



# Cloud and data centre networking

COSC349—Cloud Computing Architecture  
David Eysers

# Learning objectives

- Illustrate how network traffic on **users' devices reaches servers** within cloud data-centres
- Explain what a **content delivery network** (CDN) is and the key features it can provide
- Describe some key design considerations for **networks in large data-centres**
- Contrast the **typical traffic patterns** in large organisational networks (e.g., the University of Otago) with those of data-centre networks

# Internet access to the cloud

- Consider IP packet travelling from **web browser to cloud**
- First data makes its way to the edge of UoO network:
  - Wi-Fi in your laptop to **layer-2 wireless access points** (APs)
    - e.g., many lecture theatres contain between 1 and 3 visible APs
  - Transition from **layer-2 (MAC) to layer-3 (IP)** occurs around here
  - **Ethernet switches** in buildings aggregate traffic
    - Likely that there are hundreds of ethernet switch devices at UoO
  - Fibre optic campus cabling brings traffic to **ITS data centres**
  - ITS data centres apply internet traffic control at **campus router**
  - Then traffic leaves the University network (now a 100Gbps link)



# Onto the NZ Internet

- Assume our traffic is 'academic':
  - Goes to our educational internet provider, REANNZ
- REANNZ 'ladder' topology:
  - 23 points of presence (PoPs)
  - **Primary backbone: 100 Gbps**
  - Secondary backbone: 10 / 20 Gbps
  - 'rungs' provide redundancy
  - RTT **DUD** ↔ **AKL around 23 ms**
    - RTT: round trip time





# ... and thence to the international Internet



REANNZ is part of a global network of NRENs, connecting researchers across the globe to share information and ideas.

These lines are indicative only and do not show precise routes.

COSC349 Lecture 17, 2020

# Using AWS or Azure via REANNZ

- Typical commercial use of cloud **traverses Internet**
  - Border Gateway Protocol used for **global-scale IP routing**
    - BGP numbers 'autonomous systems' for routing IP prefixes
    - 90,000+ autonomous system numbers registered by 2019
  - By design, **not known which autonomous systems** will be used
- Large cloud providers facilitate more direct IP routing
  - **AWS DirectConnect** maps clients' networks into Amazon cloud
    - uses 802.1q VLANs: virtual Ethernets supported by most switches
  - **Azure ExpressRoute** maps clients' networks into Microsoft cloud
    - uses Multiprotocol Label Switching (MPLS): label-based routing



# Caching within New Zealand

- Lots of content is **cached within New Zealand**
  - Far better to do so than redundantly use International links
  - ... but needs appropriate software support
- Commercial ISPs and REANNZ host caches, e.g.,
  - Google, Facebook, Akamai and Netflix all hosted on REANNZ
  - Caches may be **installed at expense of origin organisation**
    - Expense will be worth it, to improve their customers' experience
- Often this sort of **caching relates to static content**
  - ... but this includes large objects such as high-definition movies

# Content Delivery Networks (CDNs)

- A CDN is a **globally distributed network** of edge servers
  - CDNs can be thought of as distributed **caching as a service**
  - & go beyond caches: compression, edge computing, routing
- **Amazon CloudFront** is AWS's CDN offering (190 PoPs)
  - Caching of static content regionally + some edge computing
  - Automatic failover between multiple AWS origin servers
  - Easy integration with AWS services such as EC2 and S3
  - Client device detection; client country detection
  - DDoS protection; certificate management; threat scanning



# Data-centre networking

- Assume our traffic needs to reach the origin server VM
- Having traversed Internet, **reach a data-centre** (DC)
  - Data-centres long pre-date the public cloud—different types
- DCs are often constrained by their network links
  - ... although this does depend on the workloads being run
  - Typical trade off **provisioning peak and average load** (+\$\$\$)
- Big cloud providers' DCs are now termed '**hyperscale**'
  - Have **very high efficiencies**: power; cooling; networking, ...

# DC physical and network layout

- Data-centres have **structured physical layout**
  - Physical servers are grouped into racks, which usually provide:
    - intra-rack network switches; power distribution; wiring management
  - Racks grouped into **aisles or zones**—for maintenance / access
- DC networking is **hierarchically organised**:
  - **Virtual networking** within physical servers themselves
  - **Top-of-rack switches** installed within racks
  - Inter-rack and Internet connectivity—but with **what topology?**

# DC networking requires a topology

- Wiring up complete graph isn't practical ( $\mathcal{O}(n^2)$  cost)
- So need topology that's **cheap but also effective**
  - Must consider degrees of freedom regarding contention
- High Performance Computing (HPC) has explored this
  - HPC (supercomputing) often has structured data processing
  - Topologies used include: **mesh hypercube; fat tree** (Clos)
- HPC often uses high numbers of concurrent flows
  - In contrast, cloud workloads are likely to be **more bursty**
  - **Clusters of activity** will depend on application; time-of-day; ...



# DC network designs: three-tier (old)

- A historically common, hierarchical DC network design:
  - **Access layer** switches connect to servers
    - Typically commodity devices
  - **Aggregate layer** connect access layer
  - **Core layer** connect aggregate layer to Internet
  - Separation of concerns between the different layers
- Upper-level routers: **highly specialised and expensive**
  - Designed to sustain high bandwidth in all directions
  - Core routers' prices have been in the order of \$100,000 a piece

# Common cloud DC network design: fat tree

- **Fat tree** structures have **thicker branches near root**
  - In the network case, this means higher bandwidth links
  - ... however such “links” need not be a single “cable”
- Modern trend is toward **using commodity switching kit**
  - Create an aggregation switch from a set of cheaper switches
  - Employs a particular addressing scheme and routing algorithm
- Often related to a topology known as a Clos network
  - Have **ingress, middle and egress stages** built from switches
  - ... and the structure can be used recursively (expand middle)

# Types of traffic and effect on workload

- Domestic / campus traffic? Typically **to/from Internet**
  - (Also could be to/from an internal DC like CS or ITS)
  - This means that the **node-to-node internal network use is small**
- Consider scale-out applications in datacenter instead
  - Interacting servers do so with **distributed effect & high volume**
    - (Although many servers won't interact—e.g., different tenants' VMs)
- Some 'interesting' traffic patterns: (*i.e.*, non-unicast)
  - **Anycast**—traffic reaches any 'nearby' applicable host
  - **Multicast** or 1-N—traffic is being sent to multiple hosts



# Open Compute Project

- Initiated by **Facebook** (Prineville Oregon DC's designs)
  - DCs just an overhead for them: expenses down → profit up
  - Aiming to **incorporate commodity parts** in custom, open designs
    - Google had previously pioneered using commodity kit for servers
  - Covers: DC facility; racks; power; networking; servers; storage; ...
- OCP networking scope includes:
  - **Disaggregated and open** network hardware and software
  - Automated **configuration management and provisioning**
  - Switch motherboard hardware and form-factor mapping
  - Also, **Software Defined Networking** (SDN) ... (see next lecture...)