Data Center Networks / TCP

Preview Session (SIGCOMM 2016)

Tobias Flach
flach@google.com

These slides:
http://goo.gl/PUExfz

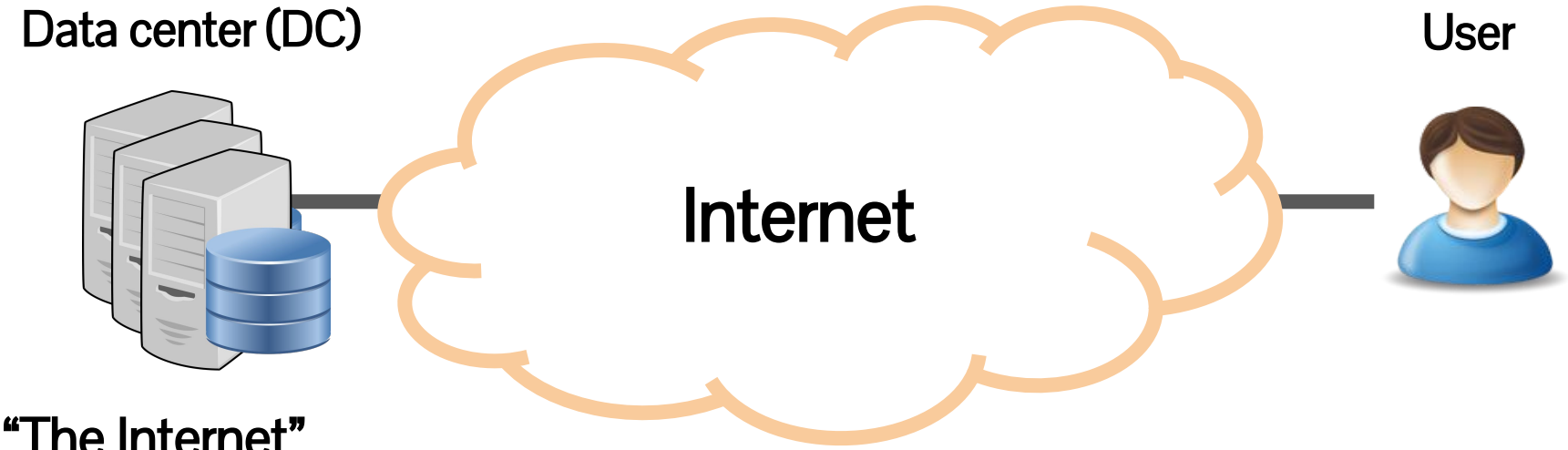
These slides:
http://goo.gl/PUExfz

TCP (Transmission Control Protocol)

"Provides **reliable**, **ordered**, and error-checked **delivery of a stream** of octets between applications running on hosts communicating over an IP network"

[Wikipedia]

Should be fast!

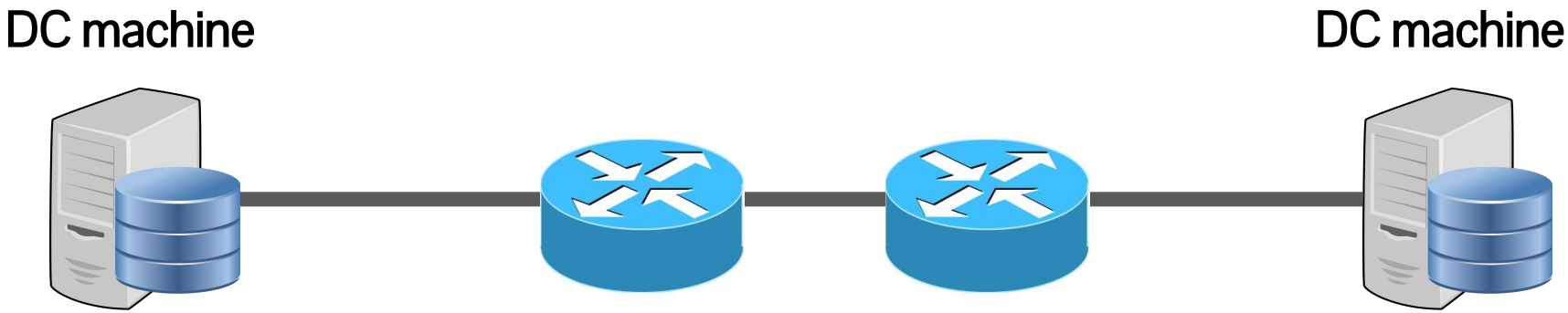


“The Internet”

- Multiple control domains
- Unknown network infrastructure
- High delays (RTTs > 200 milliseconds are common)

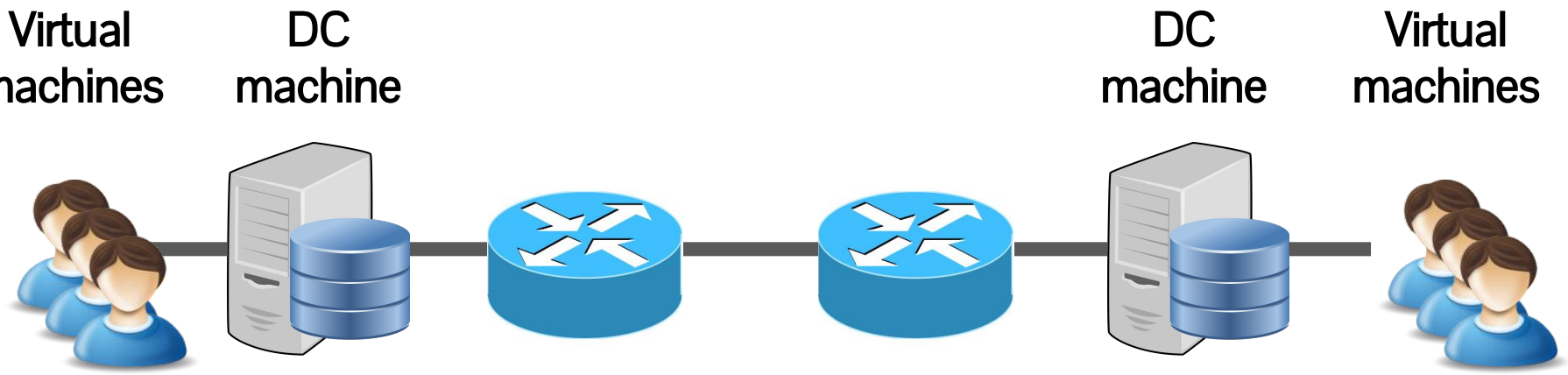
Ping issued yesterday:

```
64 bytes from 8.8.8.8: icmp_seq=25 ttl=57 time=41469.635 ms
```



Single-tenant data center

- Single control domain
- Known (and modifiable) network infrastructure
- Very low delays (RTT ≈ 100 microseconds)
- Variable traffic patterns (e.g. incast from a large-scale map/reduce job)



Multi-tenant data center

- Multiple control domains (autonomous VMs + DC infrastructure)
- Network infrastructure
  - Known to DC operators
  - Unknown to VMs
- Very low delays (RTT ≈ 100 microseconds)

From the perspective of DC operators

	Internet	Single-tenant DC	Multi-tenant DC
Endpoint control	Servers only	Yes	No
Network control	No	Yes	Yes
Round-trip times	High	Very low	Very low

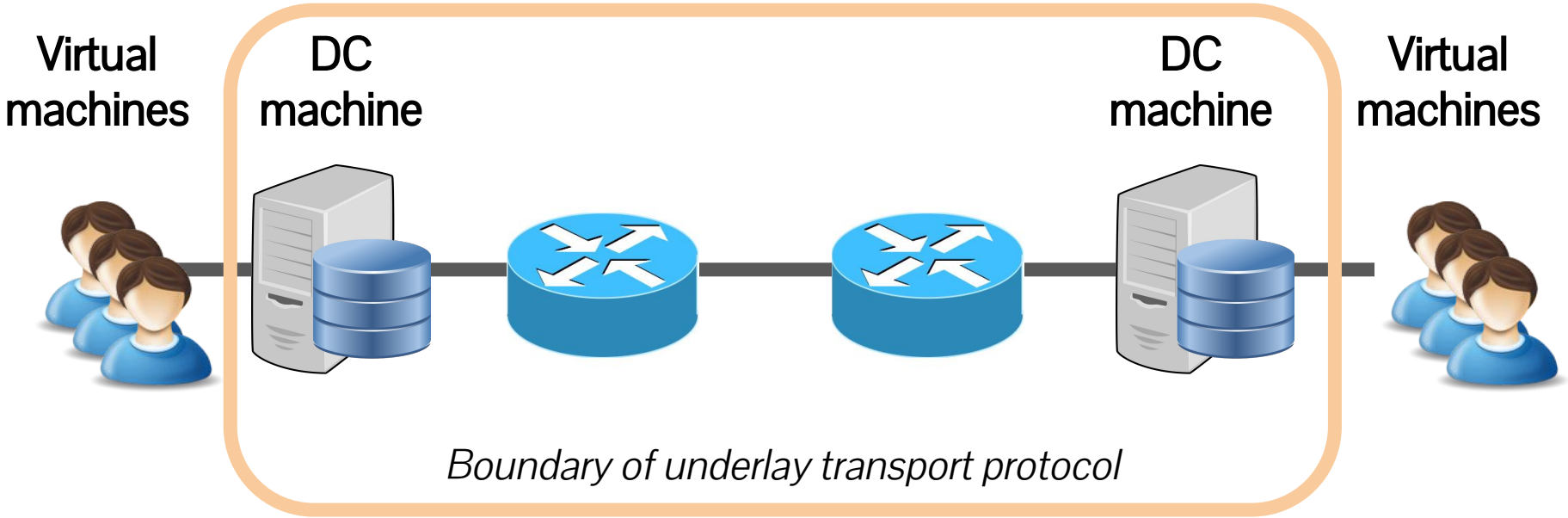
- Hard to evolve network protocols
- Need to be compatible with other protocols and legacy network infrastructure

- Easy to evolve network protocols
- Adjustable to specific environments
  - Use low timeouts
  - Explicit congestion notification

From the perspective of DC operators

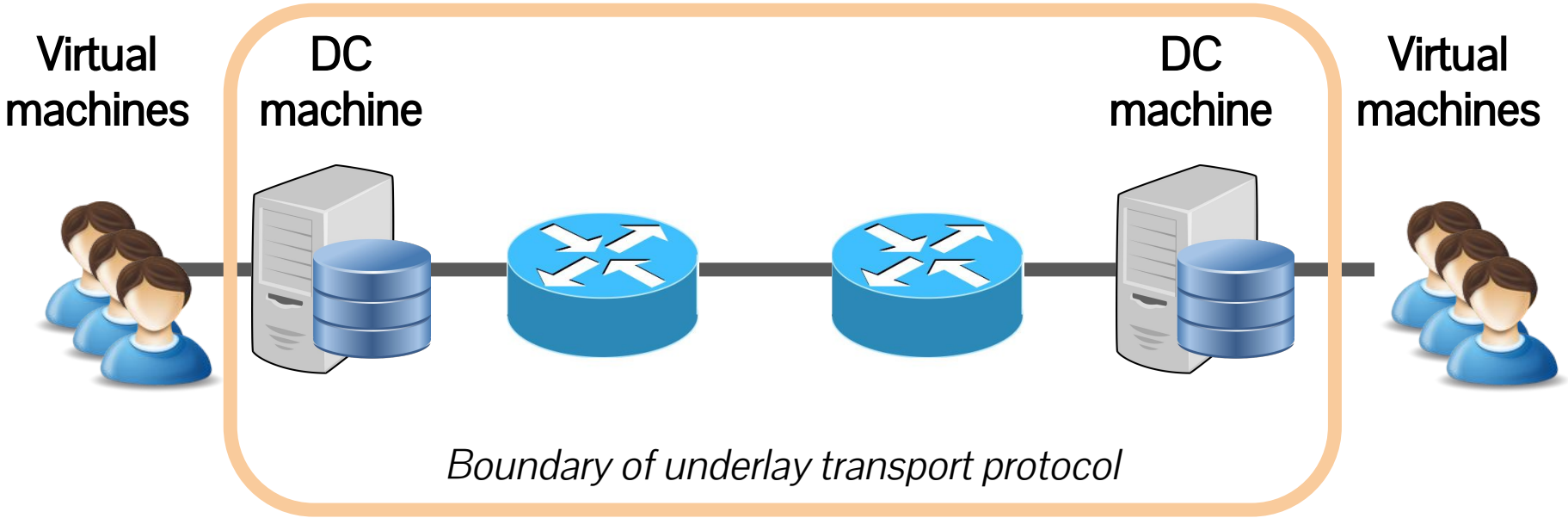
	Internet	Single-tenant DC	Multi-tenant DC
Endpoint control	Servers only	Yes	No
Network control	No	Yes	Yes
Round-trip times	High	Very low	Very low

No control of either endpoint:
I'm the boss but I can't do anything!



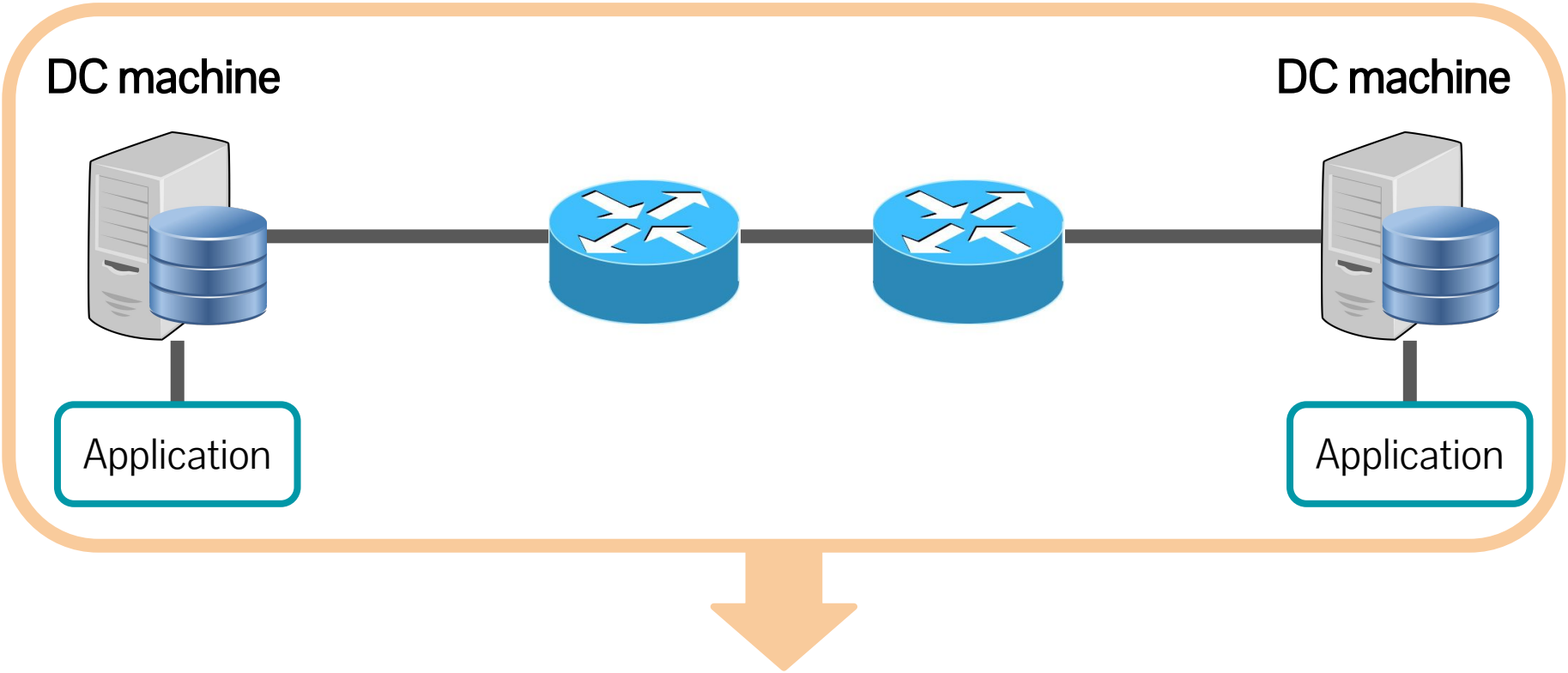
- Hypervisor* translates to underlay transport used between DC machines
- Can reap benefits of a single-tenant DC
- All virtual machines use the same protocol without knowing it!
  - Get fair network allocation even when using a legacy transport protocol

* Hypervisor: manages VMs and all VM traffic passes through it

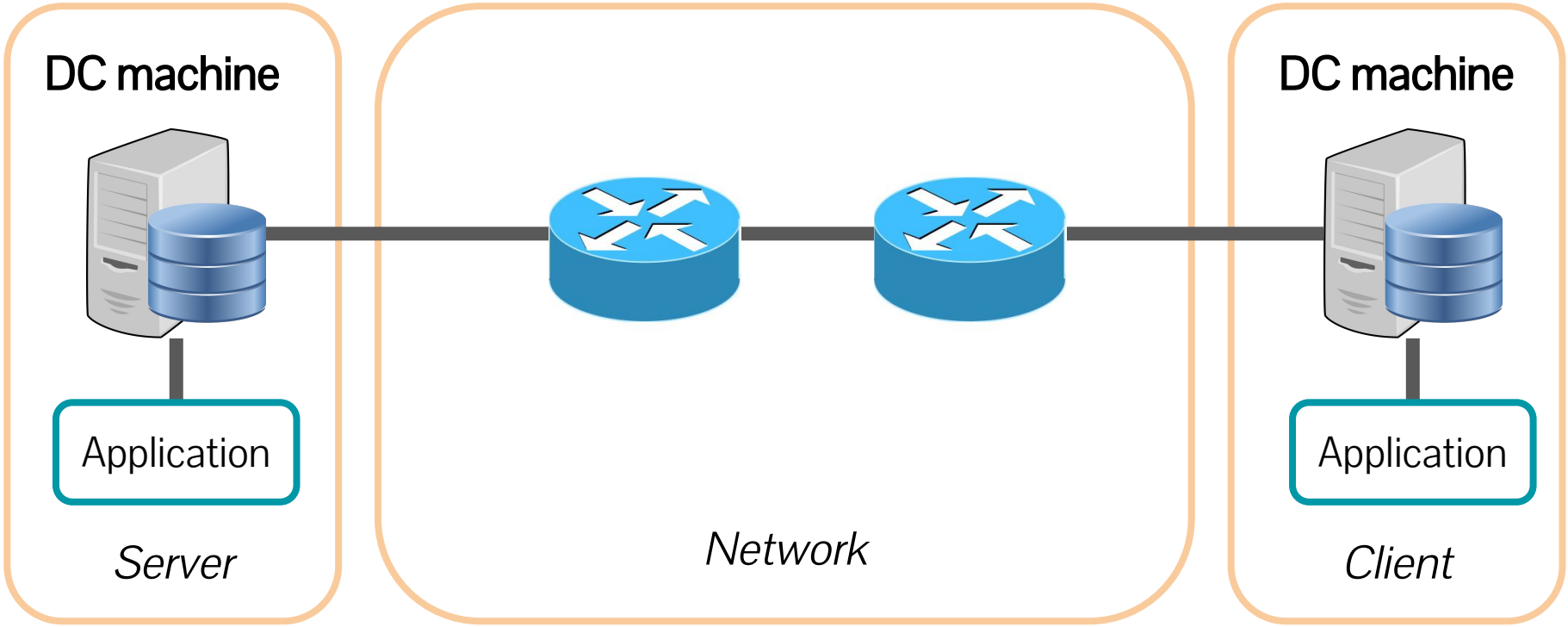


"Virtualized Congestion Control"
(Session 9, Thursday, 11:05am)

"AC/DC TCP: Virtual Congestion Control Enforcement for Datacenter Networks"
(Session 9, Thursday, 11:05am)



Application observes high latency - but why!?



- High latency could be caused by any of these entities (e.g. high I/O load on servers, packet loss on the network)
- Each entity is its own control domain (different parties without easy access to other entities)

TCP to the Rescue!

- Faults are visible to the transport protocol even if they are not network-related
- Examples:
  - Bad server: fewer client requests can be processed
    → TCP sees: less data transmitted
  - Bad client: incoming data cannot be processed quickly enough
    → TCP sees: receive window exhausted
  - Bad network: packets get dropped or reordered
    → TCP sees: packet retransmissions, reordering

"Taking the Blame Game out of Data Center Operations with NetPoirot"
(Session 9, Thursday, 11:05am)

Session 9, Thursday, 11:05am

"Virtualized Congestion Control"

"AC/DC TCP: Virtual Congestion Control Enforcement for Datacenter Networks"

"Taking the Blame Game out of Data Center Operations with NetPoirot"

These slides: **http://goo.gl/PUExfz**