Data Center Networking

(ENCS 691K – Chapter 6)

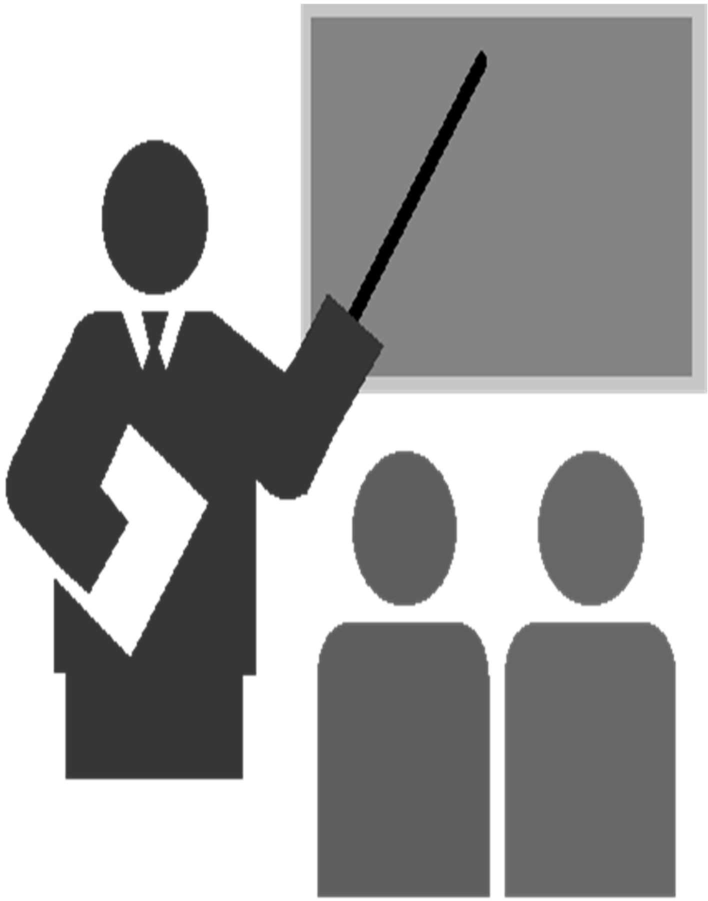
Roch Glitho, PhD
Associate Professor and Canada Research Chair
My URL - http://users.encs.concordia.ca/~glitho/

Data Center Networking

- **Basics**

- **Challenges and traditional protocols**

Data Center Networking - Basics



- Concepts
- Data Center Topology
- Data Center Virtualization

References

1. MD. Bari et al., Data Center Networking Virtualization: A Survey, IEEE Communications Surveys & Tutorials, Vol.15, No2, Second Quarter 2013

Concepts

Data Center

Consists of:

- Servers (Physical machines)
- Storage
- Network devices (switch, router, cables)
- Power distribution systems
- Cooling systems

Data Center Network

Communications infrastructure – Could be described by:

- Topology
- Routing / switching equipment
- Protocols

Data Center Network

Data Center Network vs. ISP Networks:

- Number of nodes
  - ISPs backbones (hundreds)
    - 487 for AT&T – Ref. 1
  - Data centers (thousands)
    - Google (12 000) Ref. 1

- Topology
  - Topology with specific properties are used for data center in order to allow topology specific routing optimization

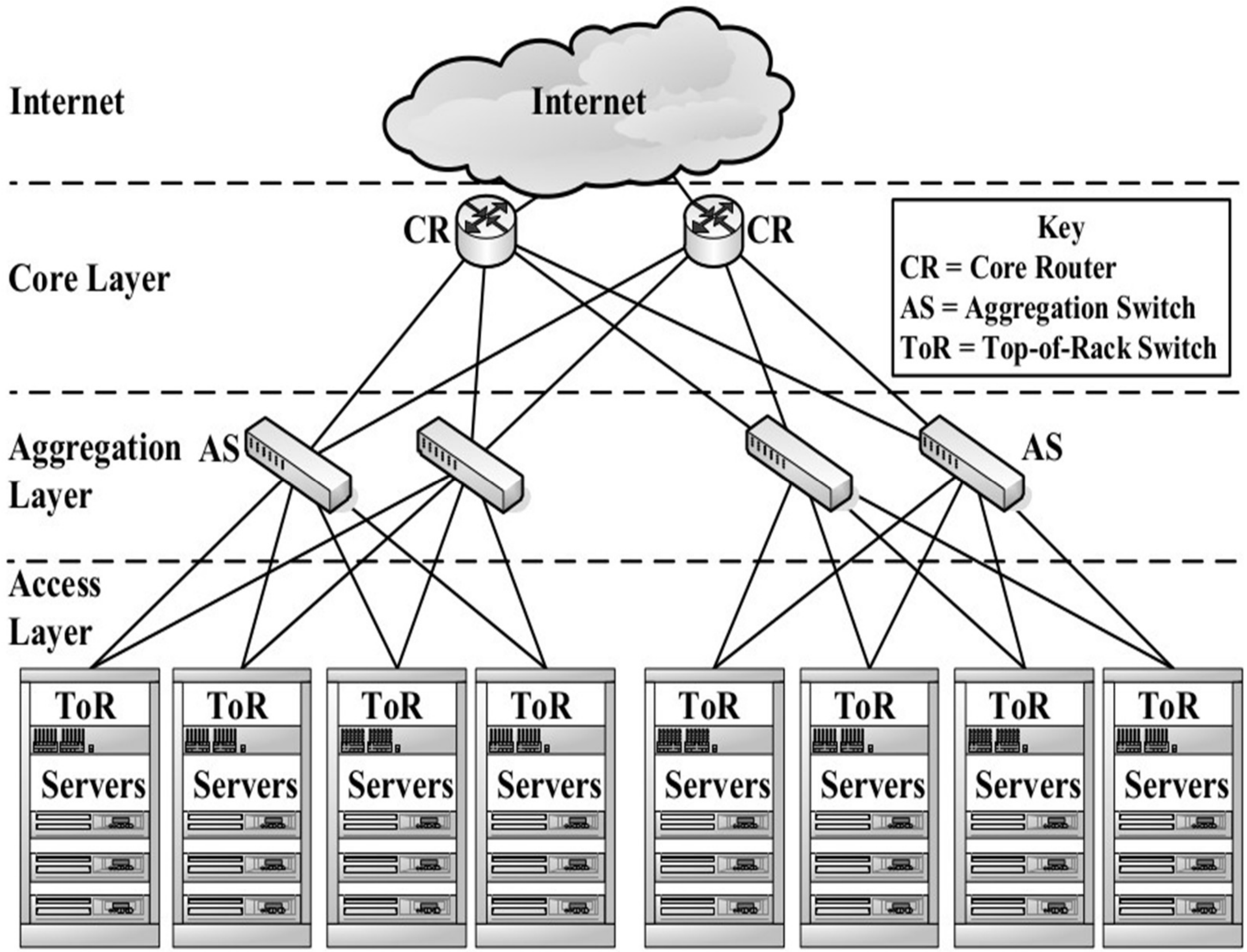
Data Center Network Topology

Conventional Topology

Three layers:

- Access layer with Top of the Rack (ToR) switches
- Aggregation layer
- Core layer

Conventional Topology (REf1)



Conventional Topology

Specific case:

- Flat layer 2 topology: Only layer 2 switches

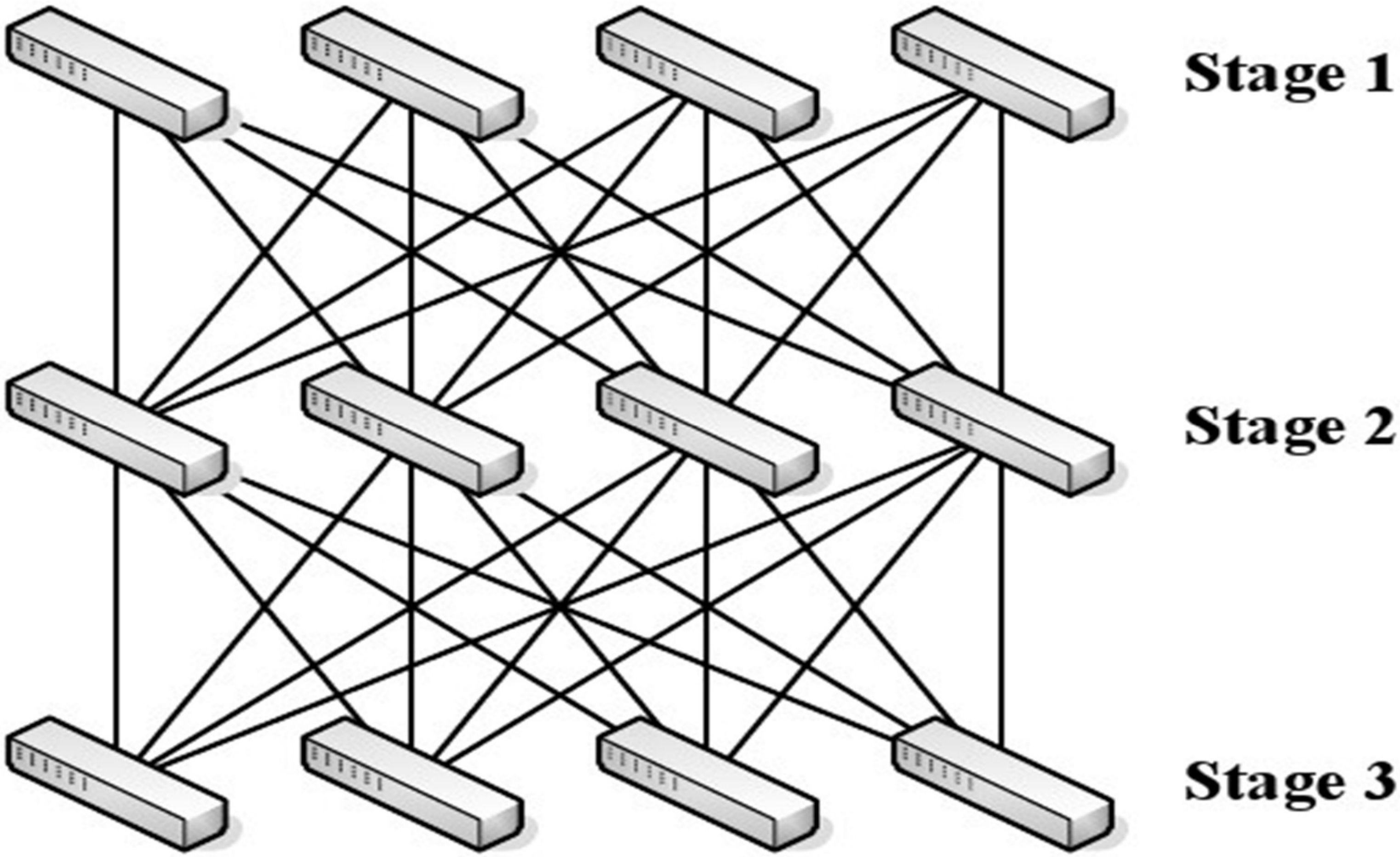
Clos Topology

Hierarchical / staged/layered:

- Each switch in a stage is connected to all the switches in the next stage

- Key benefit:
  - Extensive path diversity

Clos Topology (Ref. 1)

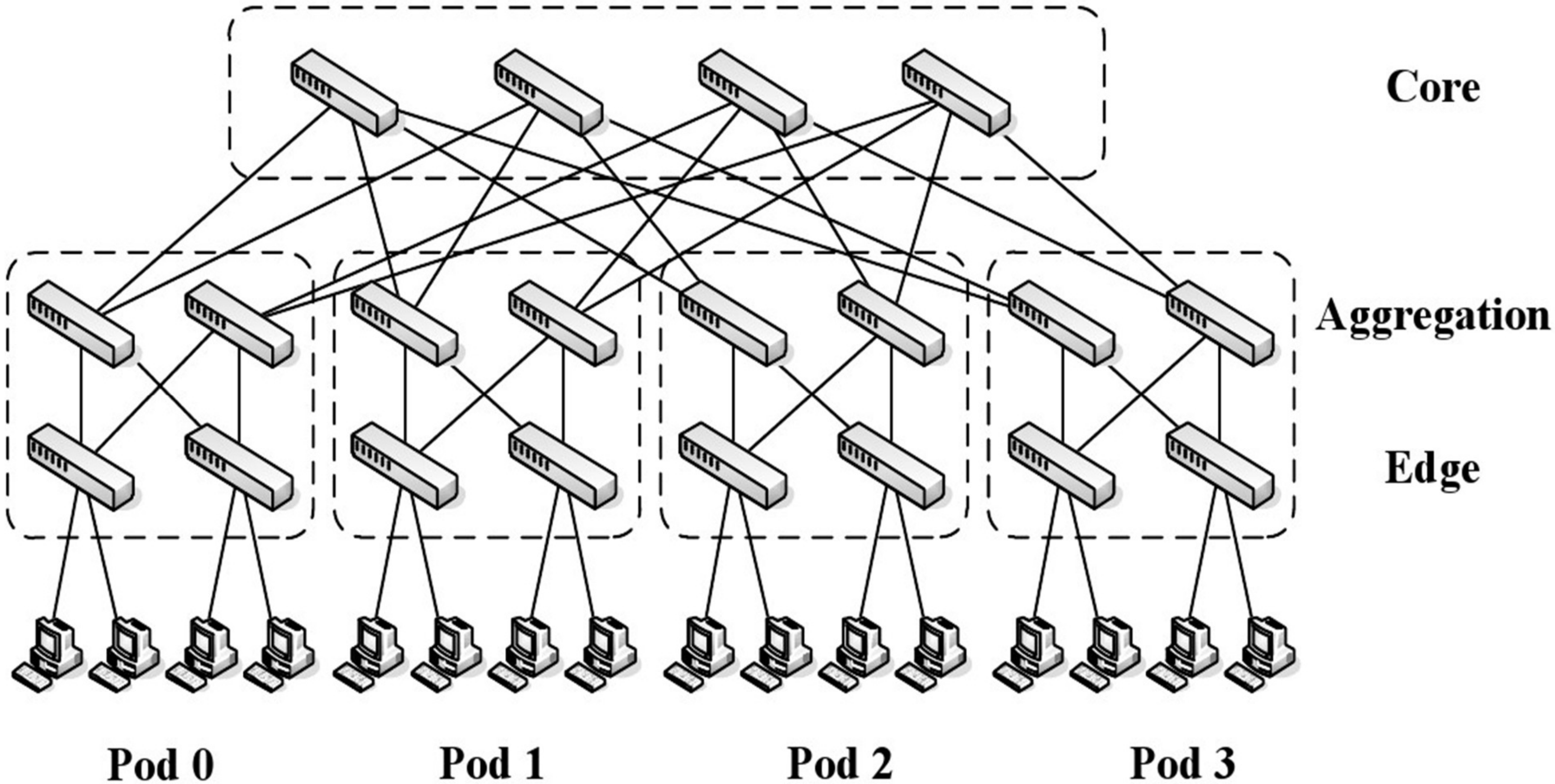


Clos Topology

Specific case: Fat tree

- Built in a tree like structure

Clos Topology – Fat tree (Ref. 1)



Data Center Virtualization

Virtualized Data Center

Data center with some or all the hardware virtualized

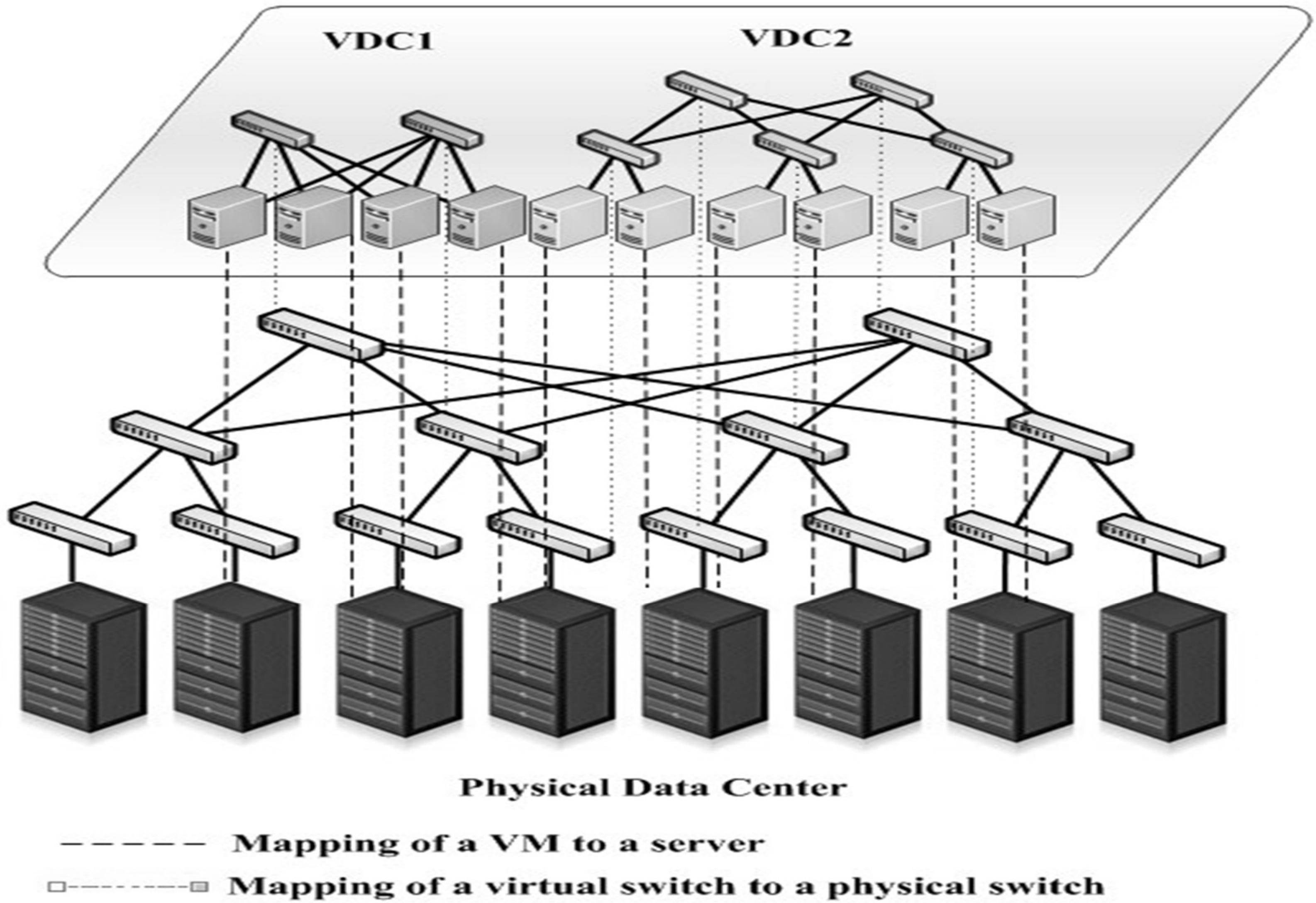
- Servers (Physical machines)
- Storage
- Network devices (switch, router)
- Power distribution systems
- Cooling systems

Virtual Data Center

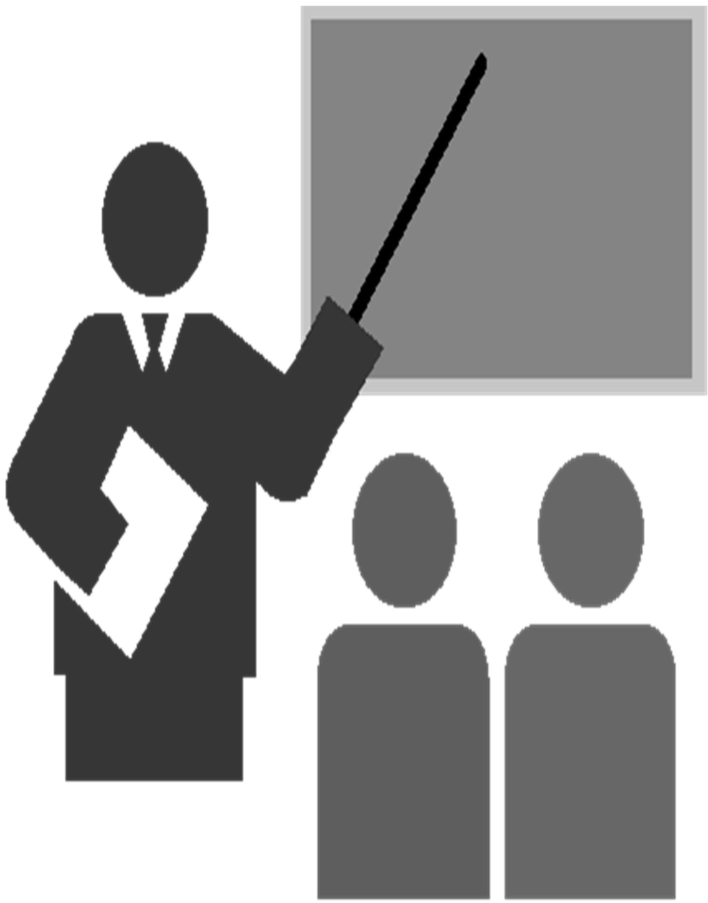
Collection of virtual resources, e.g.

- Virtual machine
- Virtual switches
- Virtual links

Virtual Data Center (Ref. 1)



Data Center Networking: Challenges and Traditional Protocols



- Data Center Networking Challenges
- Traditional Transport Protocols (Beyond TCP / UDP)
- Traditional Transport Protocols vs. Challenges

Data Center Networking Challenges

References

1. K. Kant, Towards a Virtualized Data Center Transport Protocol, Infocom Workshop, 2008
2. M Alizadeh, Data Center TCP, ACM Sigcom 2011

Data Center Networking Challenges

Why is it necessary to re-think networking in cloud data center settings?

- Very high data rates (e.g. 100 Gb/sec Ethernet)
    - TCP can hardly cope with 10 GB/sec
        - New techniques are needed to make TCP cope, e.g.
            - Hardware acceleration
    - Need for QoS mechanisms
        - A single MAC pipe can carry data with different QoS requirements

Data Center Networking Challenges

Why is it necessary to re-think networking in cloud data center settings?

- Wide range of physical layer
  - Wired
  - Wireless
  - Optical
- Emerging PHY/MAC layers, e.g.
  - Ultra Wide Band
    - Huge amount of data over a short distance

Data Center Networking Challenges

Why is it necessary to re-think networking in cloud data center settings?

- Multiple level virtualization and cluster enabled applications
  - Real time applications / soft real time applications vs. other applications

Data Center Networking Challenges

An illustration:

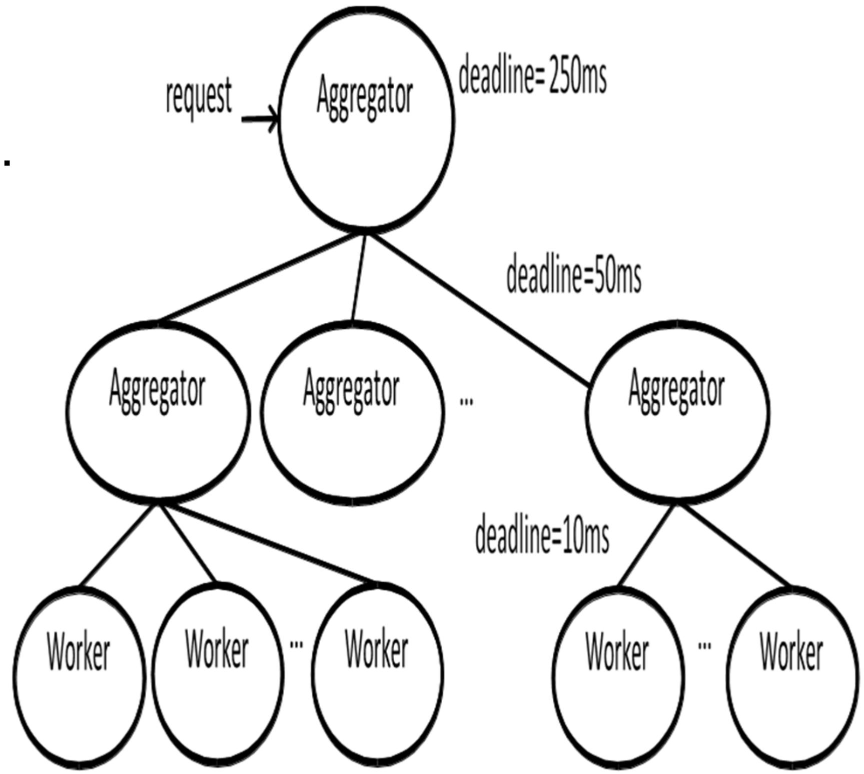
Soft real time applications, e.g.

- Web search
- Advertisement
- Retail

Data Center Networking Challenges

Partition / Aggregate pattern

.



reque
st

Data Center Networking Challenges

An illustration:

Examples of requirements:

- Low latency
- High burst tolerance

Important: Many other applications with conflicting requirements reside in the same data center

Data Center Networking Challenges

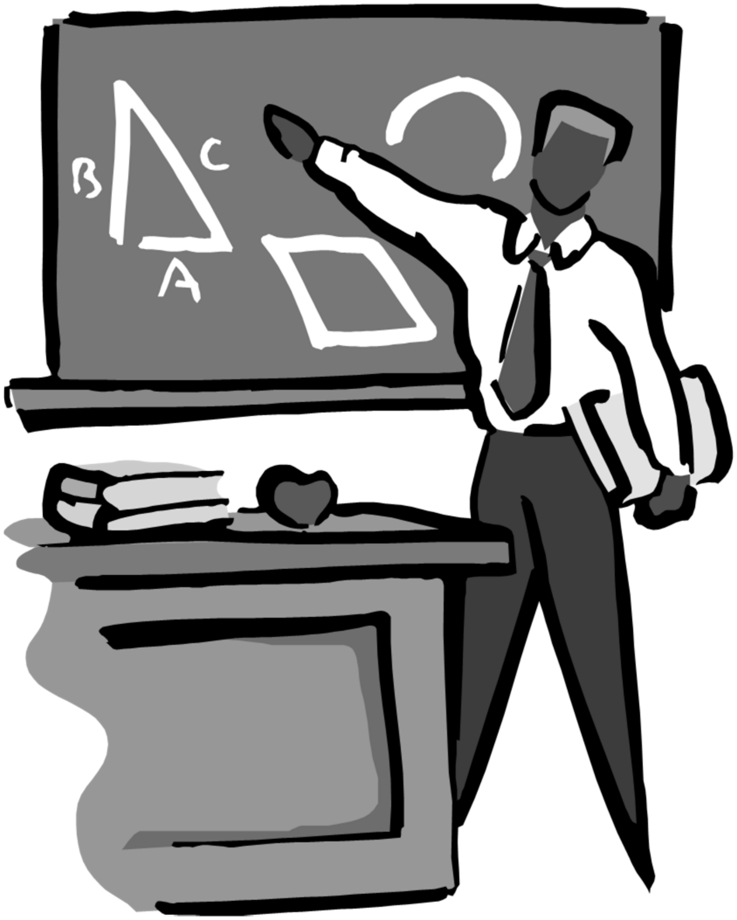
Let us focus on transport layer protocols requirements

- High data rate support (Up to 100 GB/s)
- User Level Protocol Indicator Support
- QoS friendly
- Virtual cluster support
- Data center flow / cong. Control
- High availability
- Compatibility with TCP/IP base
- Protection against DoS

On Transport Layer

Traditional Transport Layers
(Beyond TCP / UDP)

Transport Layer Basics



- 1 - Protocol layering
- 2 - Transport layer basics

Layered Architectures

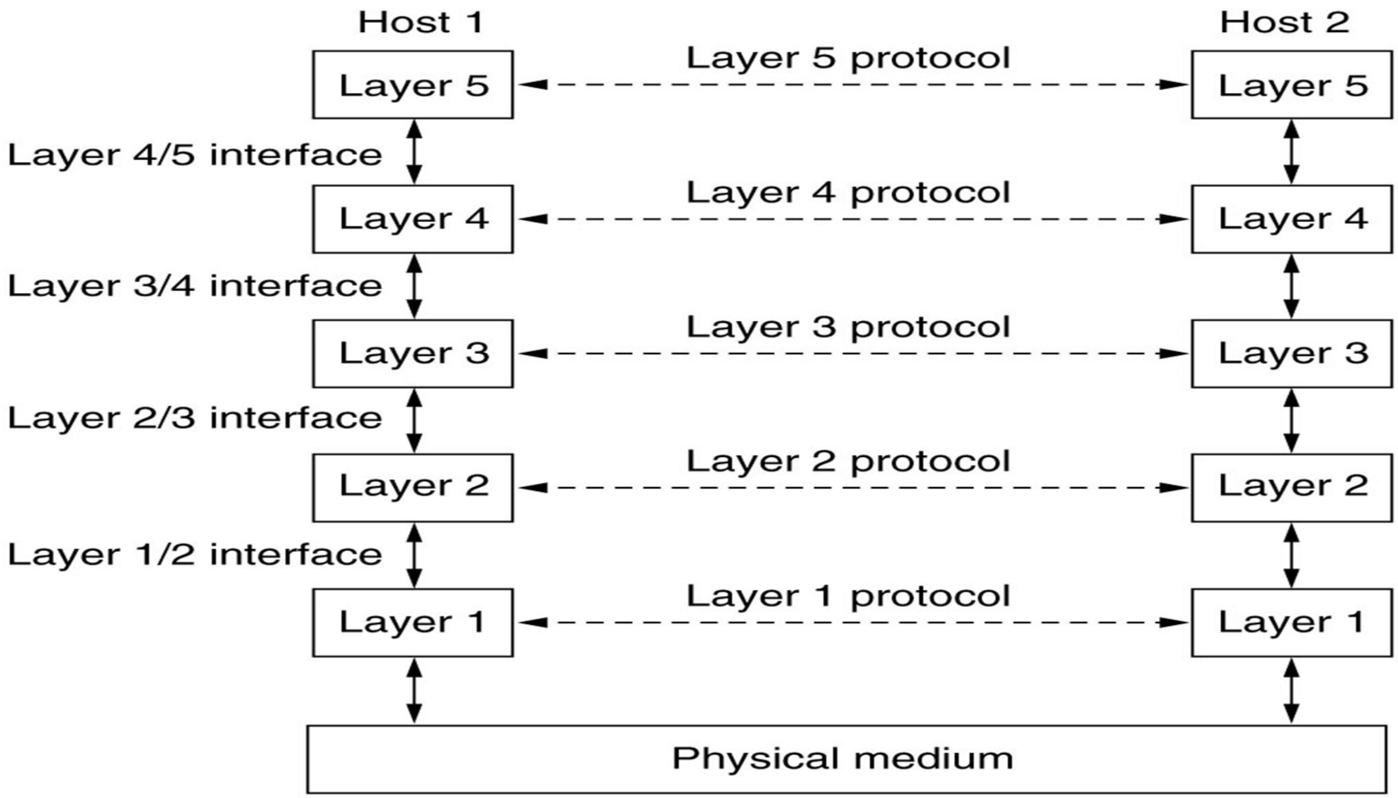


Figure 1.13 (Reference [1])

Layered Architectures

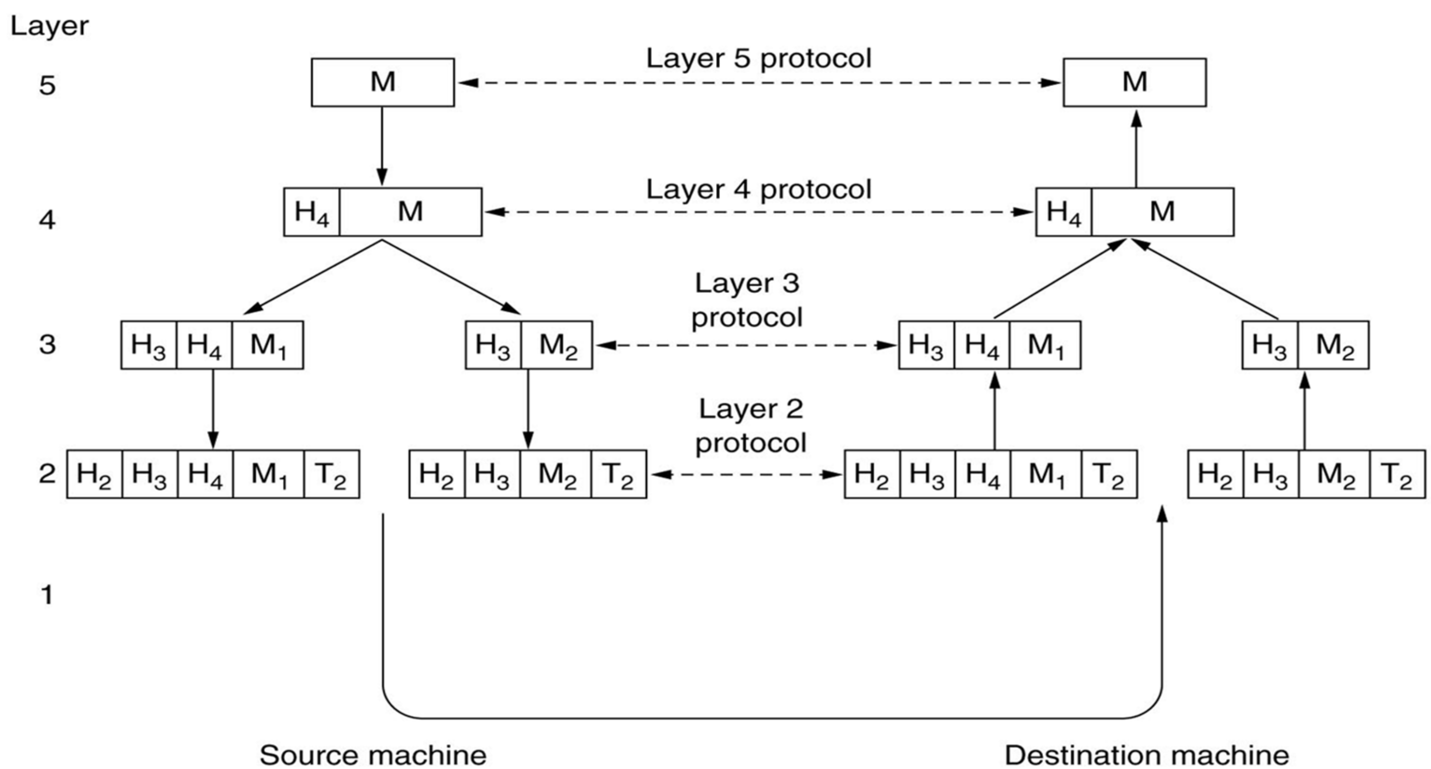
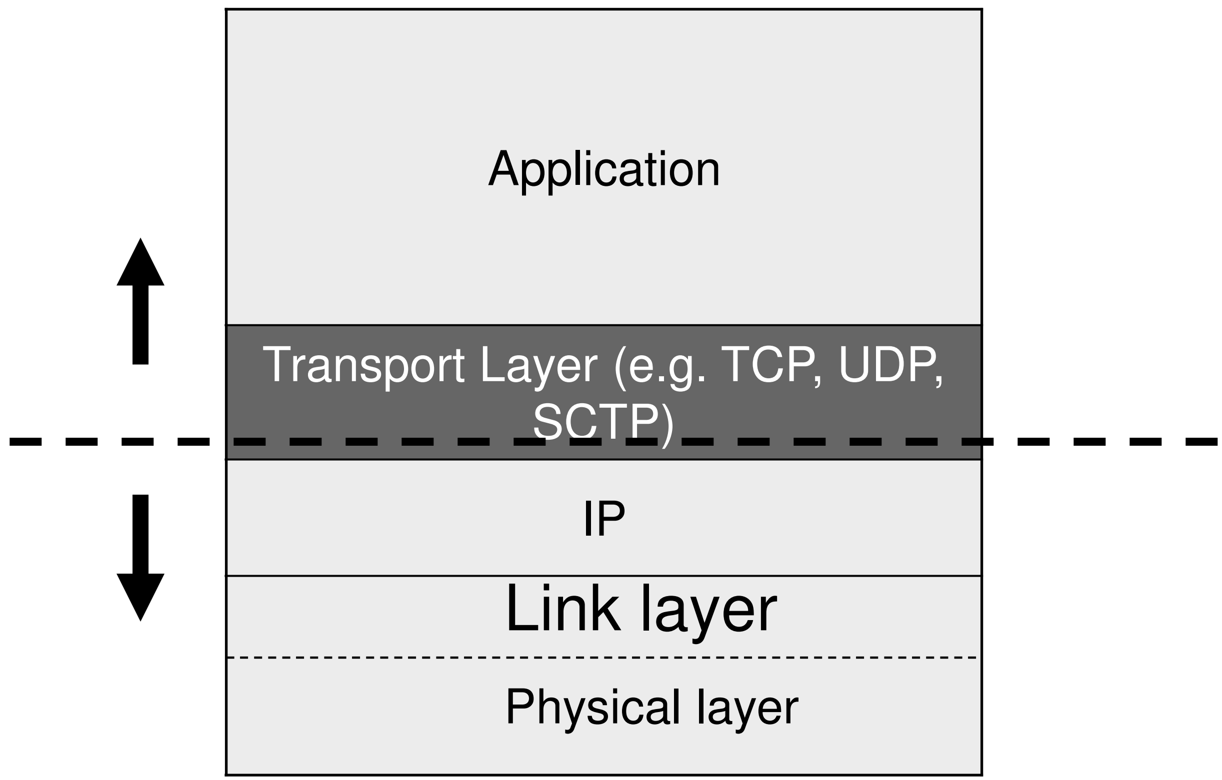


Figure 1.15 (Reference [1])

Cross Layered Architecture

- Definition of cross layer design
  - Violation of the principles of layered protocol architectures
    - Examples
      - Allowing communications between non adjacent layers
      - Sharing variables between layers
      - Designing protocols that span several layers

On The Transport Layer



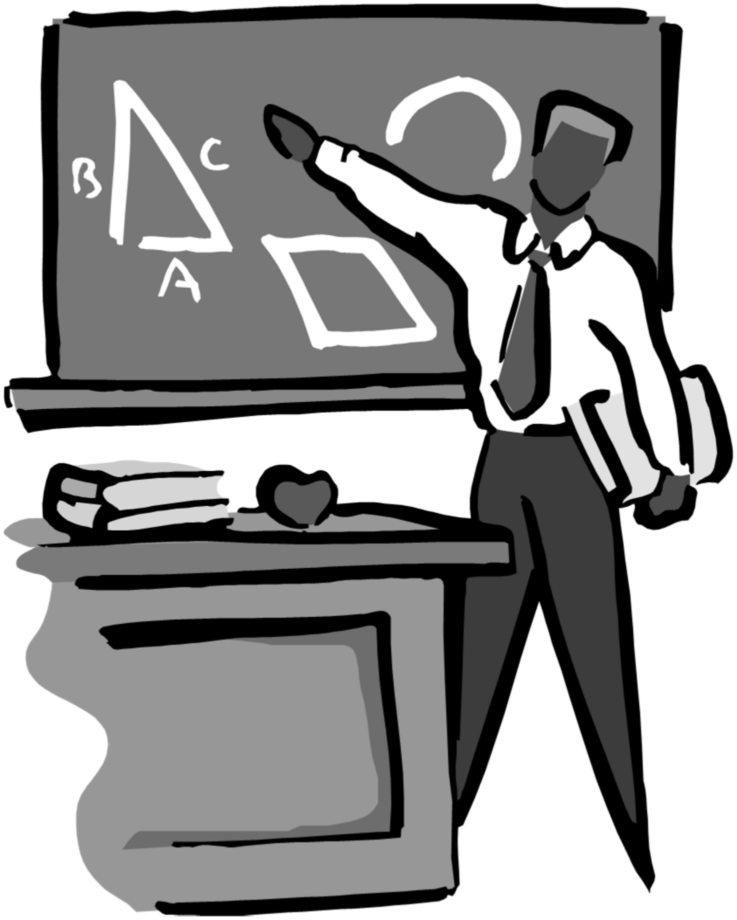
On The Transport Layer

- Provide service to application layer by using the service provided by network layer
- Hide physical network
  - Hide processing complexity
  - Hide different network technologies and architectures
- Provides host-to-host transport

On The Transport Layer

- Addressing
- Connection Establishment
- Connection Release
- Flow Control
- Error Detection and Crash Recovery

The Other Transport Protocols



- 1 - Motivations and taxonomy
- 2 - SCTP
- 3 - DCCP

References

1, IETF RFC 3550, RTP / RTCP

2. A. Caro et al., SCTP: A Proposed Standard for Robust Internet Data Transport, IEEE Computer November 2003
3. S. Fu and M. Atiquzzaman, SCTP: State of the Art in Research, Products and Technical Challenges, IEEE Communications Magazine, April 2004
4. P. Natarajan et al., SCTP: What, Why and How? IEEE Internet Computing, September / October 2009
5. Y-C Lai, DCCP: Transport Protocol with Congestion Control and Unreliability, IEEE Internet Computing, September / October 2008

Motivations and Taxonomy

Key characteristics of TCP

- Reliability
  - Three way handshake connection
  - Re-transmission
- Congestion control
  - Windows
    - Transmission rate reduction
- Uni-homing

Motivations and Taxonomy

Key characteristics of UDP

- No reliability
- No congestion control
- Uni-homing

Motivations and Taxonomy

The one size (either TCP or UDP) fits all philosophy does not always work

- What about
  - Applications requiring reliability but real time delivery (i.e. no retransmission)?
    - Interactive audio/video (e.g. conferencing)
  - Applications requiring more reliability than what is provided by TCP?
    - Multimedia session signalling
  - Applications requiring real time delivery, low reliability, but congestion control?
    - Multi party games

Motivations and Taxonomy

Two possible approaches

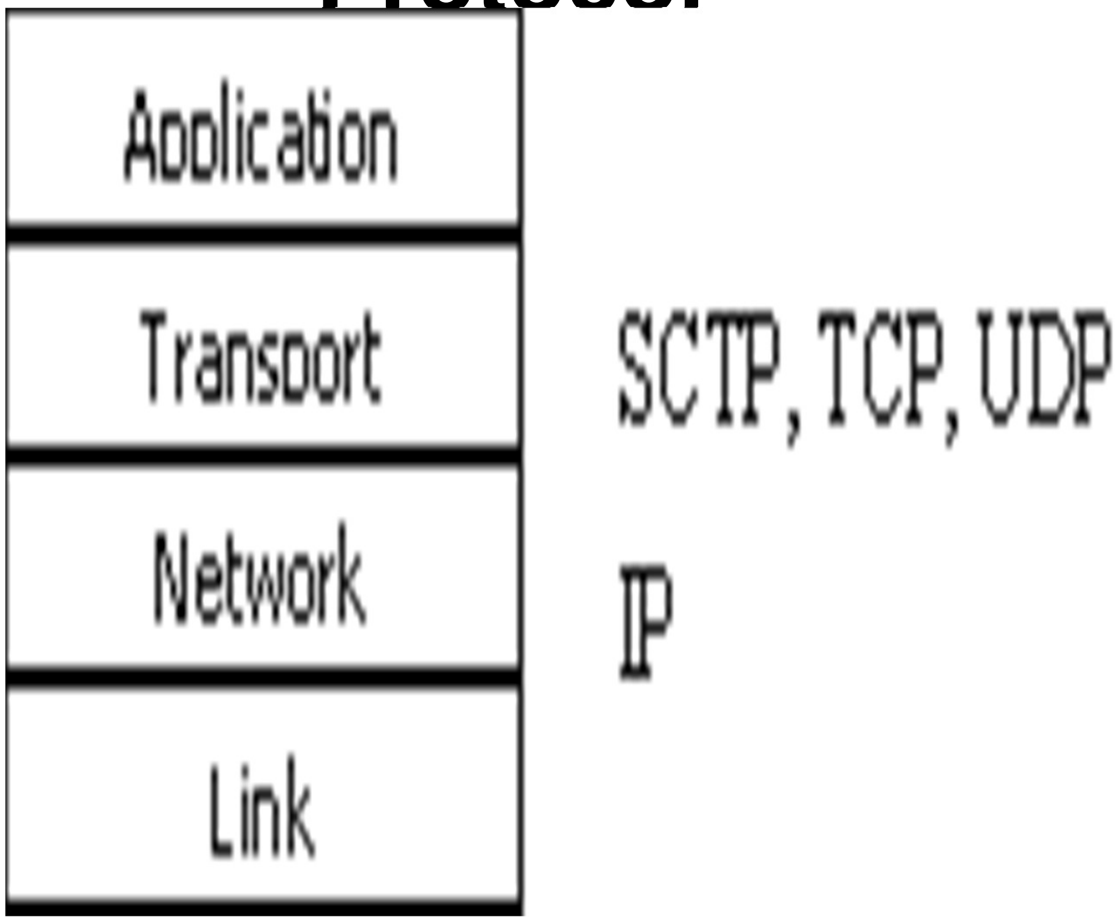
- Build a new transport protocol that complements / runs on top of existing transport protocols (e.g. UDP)
  - RTP/RTCP on top of UDP and application using RTP/RTCP
- Build a new transport protocol from scratch (i.e. runs on top of IP)
  - SCTP
  - DCCP

Stream Control Transmission Protocol (SCTP)

Designed in early 2000s to carry multimedia session signaling traffic over IP, then subsequently extended to meet the needs of a wider range of application

- Design goals much more stringent than TCP design goals (e.g. redundancy, higher reliability)
- Offer much more than TCP
- A sample of additional features
  - Four way handshake association instead of three way handshake connection
  - Multi-homing instead of uni-homing
  - Multi-streaming instead of uni-streaming

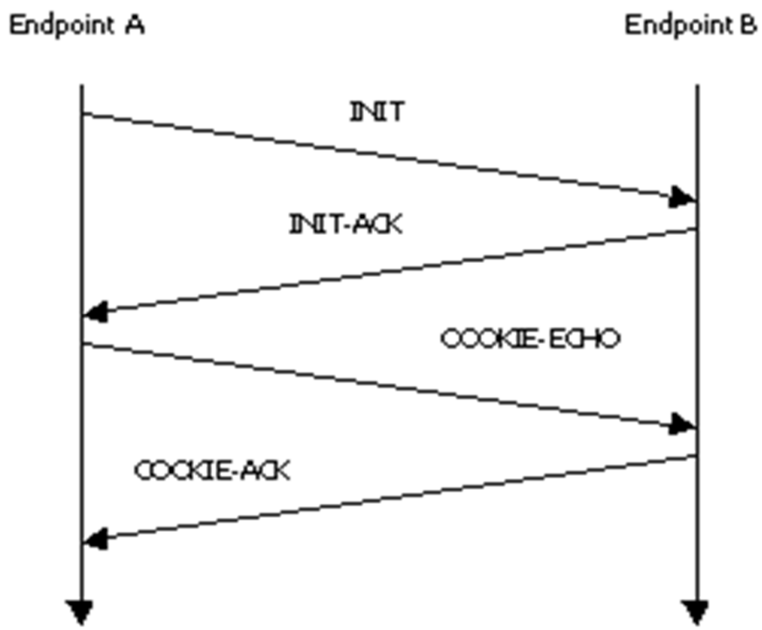
Stream Control Transmission Protocol

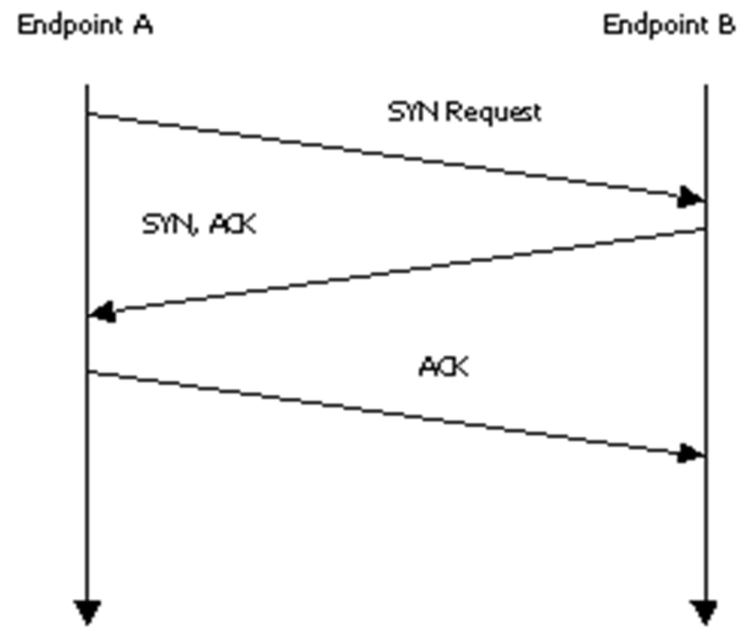


Four way handshake

Why?

- Key reason: Make SCTP resilient to denial of service (DOS) attacks, a feature missing in TCP





SCTP
TCP

Multi-homing

Why?

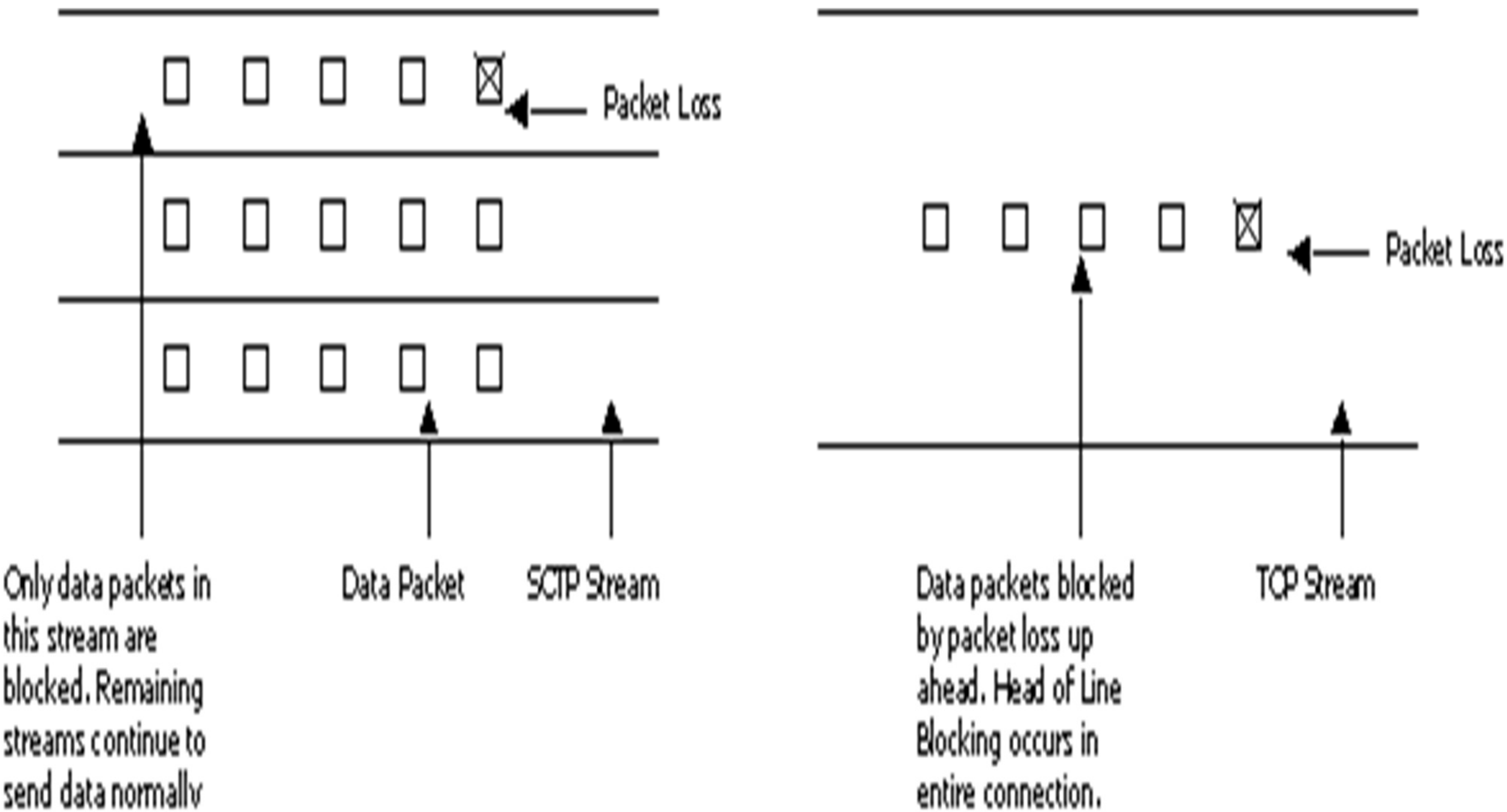
- Key reason: Make SCTP resilient in resource failures, a feature missing in TCP (High availability)
  - Multi-homed host: Host accessible via multiple IP addresses
  - Use cases
    - Subscription to multiple ISP to ensure service continuity when of the ISP fails
    - Mission critical systems relying on redundancy
    - Load balancing

Multi-homing

Why?

- Key reason: Make SCTP resilient in resource failures, a feature missing in TCP
  - Multi-homing with SCTP (only for redundancy)
    - Multi-homed host binds to several IP addresses during associations unlike TCP which binds to a single IP address
      - Retransmitted data is sent to an alternate IP address
      - Continued failure to reach primary address leads to the conclusion that primary address has failed and all traffic goes to alternate address

Multi-streaming



Data Congestion Control Protocol (DCCP)

One of the most recent transport protocols (Second half of the 2000s)

- Primary goal:
  - Delivery of real time media (somehow similar to the goal assigned to RTP / RTCP)
- Build on the experience acquired in protocol design / deployment since the design of RTP / RTCP (ie. Early 1990s)
  - Some examples of improvements:
    - Congestion control incorporated in the transport protocol (unlike RTP/RTCP)
    - Possibility to avoid DoS

Overall view

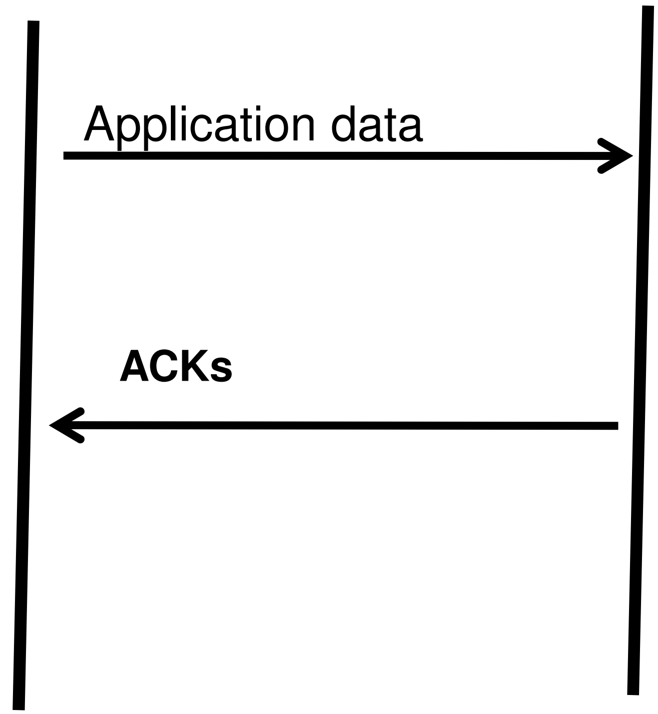
- Three way handshake connection like TCP
  - In-built possibility to use cookies during response phase to avoid DoS
  - A connection can be seen as two half-connections (i.e. uni-directional connections)
    - Possibility for a receiver to send only ACK
- Reliable connection establishment and feature negotiation
- Unreliable data transfer (no retransmission)
- Feature negotiation

The protocol states

```
Client                                        Server
  ------                                        ------
              (0) No connection
  CLOSED                                        LISTEN

              (1) Initiation
  REQUEST     DCCP-Request -->
                  <-- DCCP-Response         RESPOND
  PARTOPEN    DCCP-Ack or DCCP-DataAck -->

              (2) Data transfer
  OPEN         <-- DCCP-Data, Ack, DataAck -->    OPEN

              (3) Termination
                  <-- DCCP-CloseReq         CLOSEREQ
  CLOSING     DCCP-Close -->
                   <-- DCCP-Reset            CLOSED
  TIMEWAIT
  CLOSED
```

Half connection

Use case: Unidirectional streams (e.g. Streaming applications)



Data transfer

- Packets have sequence numbers
  - Client – server and server – client sequence numbers are independent
    - Tracking on both sides is possible
- Acknowledgements report last received packet
- Data drop option
  - Examples
    - Application not listening
    - Receiver buffer
    - Corrupt
  - May help in selecting congestion control mechanism

Data transfer

- Packets have sequence numbers
  - Client – server and server – client sequence numbers are independent
    - Tracking on both sides is possible
- Acknowledgements report last received packet
- Data drop option
  - Examples
    - Application not listening
    - Receiver buffer
    - Corrupt
  - May help in selecting congestion control mechanism

Feature negotiation

- Enable dynamic selection of congestion mechanism
  - Data drop option may help
    - Tracking on both sides is possible
  - TCP congestion control may be used
  - Other mechanisms may also be used

Traditional Transport Protocols vs. Challenges

References

1. K. Kant, Towards a Virtualized Data Center Transport Protocol, Infocom Workshop, 2008

Traditional Transport Protocols vs. Challenges (Ref. 1.)

Feature	TCP	SCTP	IBA
Scalability to 100 Gb/s	difficult	difficult	Easy?
Msg. based & ULP support	No	Yes	Yes
QoS friendly transport?	No	No	Yes
Virtual cluster support	No	No	limited
DC centric flow/cong. control	No	No	limited
Power aware transmission	Limited	limited	No
High availability features	Poor	Fair	Fair
Compatible w/ TCP/IP base	Yes	Yes	No
Protection against DoS attacks	Poor	Good	No

The End