Workshop on Computational Science Infrastructure and Applications for Academic Development

28 Sep – 09 Oct 2015

"Trieste, ICTP"

Introduction: Networking & Routing for HPC

MAKWEBA, Damas

Instructor/HPC Section (Head)

India-Tanzania Centre of Excellence in ICT (ITCoEICT)

Dar es Salaam Institute of Technology (DIT)

dmakweba@dit.ac.tz / dmakweba@ictp.it

Networking and Routing for HPC

- **Main objective:**
  - Provide you with the basic concepts encountered in HPC network and how they applied in practice
- **Appreciates:**
  - DIT, TERNET, NSRC and ICTP

Topics

- Briefly about my place
- Basic networking fundamentals
- Network for Clusters
- Discussion

About my place

Africa


Africa
International Boundary
Road
River
National Capital
City or Town
0 400 1000 KM
0 400 1000 Miles
© 2007 Geology.com
Atlantic Ocean
Indian Ocean
Mediterranean Sea
PORTUGAL
SPAIN
ITALY
GREECE
TURKEY
CYPRUS
SYRIA
LEBANON
ISRAEL
JORDAN
IRAQ
IRAN
KUWAIT
QATAR
U.A.E.
SAUDI ARABIA
OMAN
YEMEN
Red Sea
Gulf of Aden
Algiers
Annaba
Tunis
Rabat
Casablanca
Fes
Oran
Safi
MOROCCO
Marrakech
TUNISIA
Tripoli
Misratah
Benghazi
Alexandria
Cairo
EGYPT
Asyut
Sohag
Aswan
ALGERIA
LIBYA
WESTERN SAHARA
MAURITANIA
Nouakchott
MALI
NIGER
CHAD
Port Sudan
Atbara
Khartoum
Wad Medani
El Obeid
Kosti
SUDAN
ERITREA
Asmara
DJIBOUTI
Djibouti
Addis Ababa
ETHIOPIA
SOMALIA
Mogadishu
THE GAMBIA
Dakar
SENEGAL
Banjul
Bissau
GUINEA
Conakry
Freetown
SIERRA LEONE
Monrovia
LIBERIA
Bamako
Bobo Dioulasso
BURKINA FASO
Ouagadougou
Niamey
Korhogo
COTE D'IVOIRE
Yamoussoukro
Abidjan
GHANA
Accra
Lome
TOGO
BENIN
Porto-Novo
Sokode
Parakou
NIGERIA
Abuja
Moundou
Lake Chad
Abeche
N'Djamena
Sarh
CENTRAL AFRICAN REPUBLIC
Bangui
CAMEROON
Yaounde
Malabo
EQUATORIAL GUINEA
Libreville
GABON
REP. OF THE CONGO
Tchibanga
Brazzaville
Pointe-Noire
Kinshasa
Kisangani
DEMOCRATIC REPUBLIC OF THE CONGO
Mbuji-Mayi
Nimule
UGANDA
Kampala
KENYA
Nairobi
Kigali
Bujumbura
Arusha
RWANDA
BURUNDI
Tanga
Zanzibar
Dar es Salaam
Dodoma
TANZANIA
Luanda
ANGOLA
Lobito
Huambo
Lubumbashi
Ndola
ZAMBIA
Lusaka
MALAWI
Lilongwe
Salima
Blantyre
COMOROS
Mahajanga
Harare
ZIMBABWE
Bulawayo
Gweru
MOZAMBIQUE
Beira
Antananarivo
Toamasina
MADAGASCAR
Reunion (France)
NAMIBIA
Windhoek
BOTSWANA
Gaborone
Pretoria
Pietersburg
Mmabatho
Maputo
SWAZILAND
Mbabane
Maseru
LESOTHO
Kimberley
Bloemfontein
Durban
SOUTH AFRICA
Umtata
East London
Cape Town
Port Elizabeth

Tanzania

Capital: Dodoma

Largest: Dar es Salaam

Lang: Swahili, English

Area: 947,303 sq km

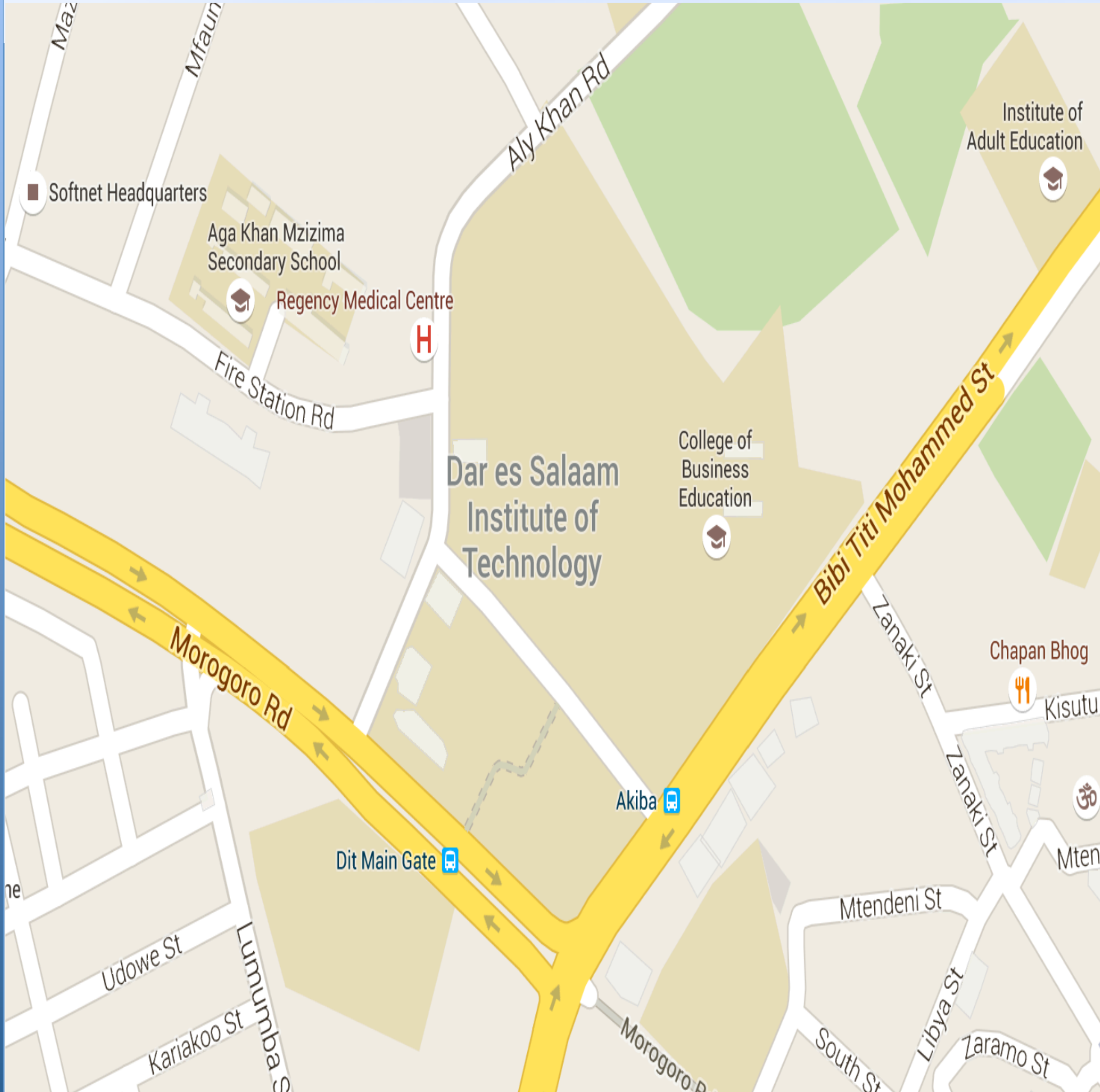
Pop: ~50.76 million
Dar: 4+ million

Currency: TZS

Time zone: (UTC +3)

Source: wikipedia

DIT



Established: 1957

Formally:

- **Vocation Training (1957)**
- **Technical College (1962)**
- **Institute of Technology**

Departments: ~ 11

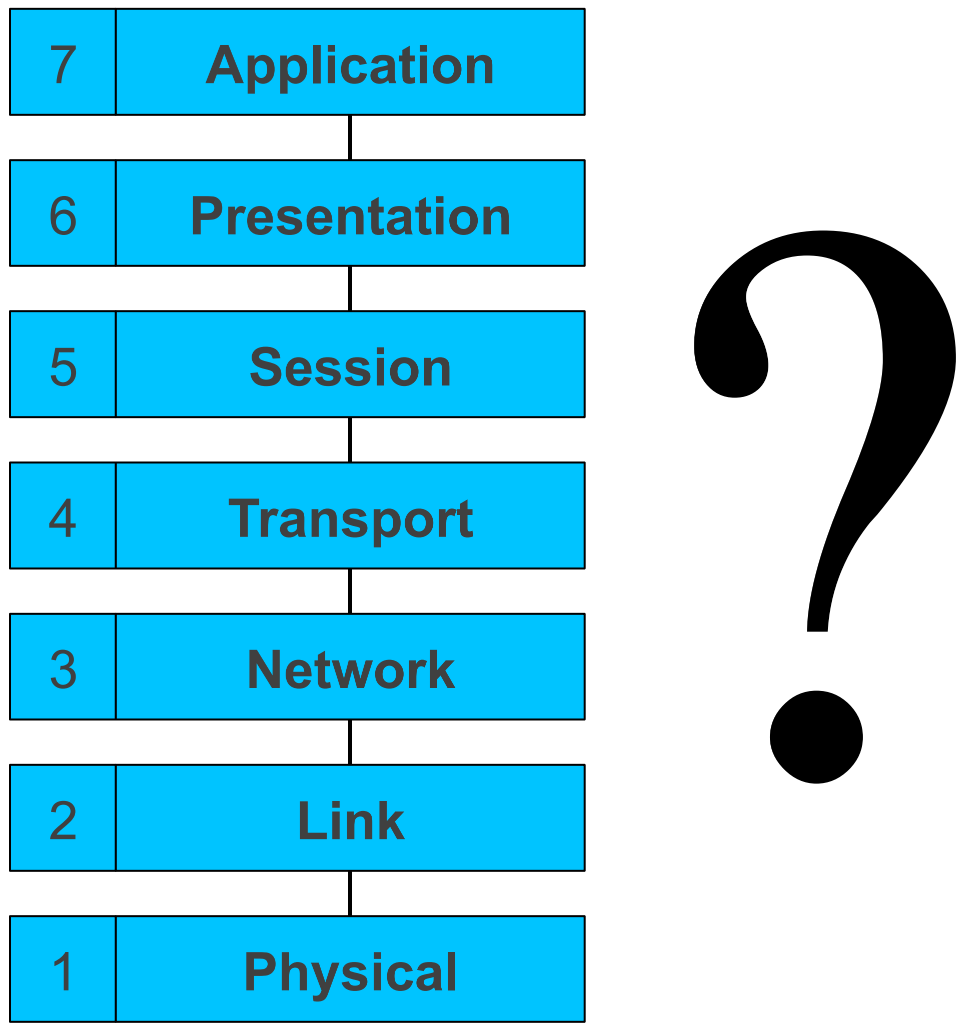
Courses:

- **MSc (2)**
- **BSc (6)**
- **Ordinary Diploma (12)**
- **Prof (10+)**

Source: wikipedia/DIT

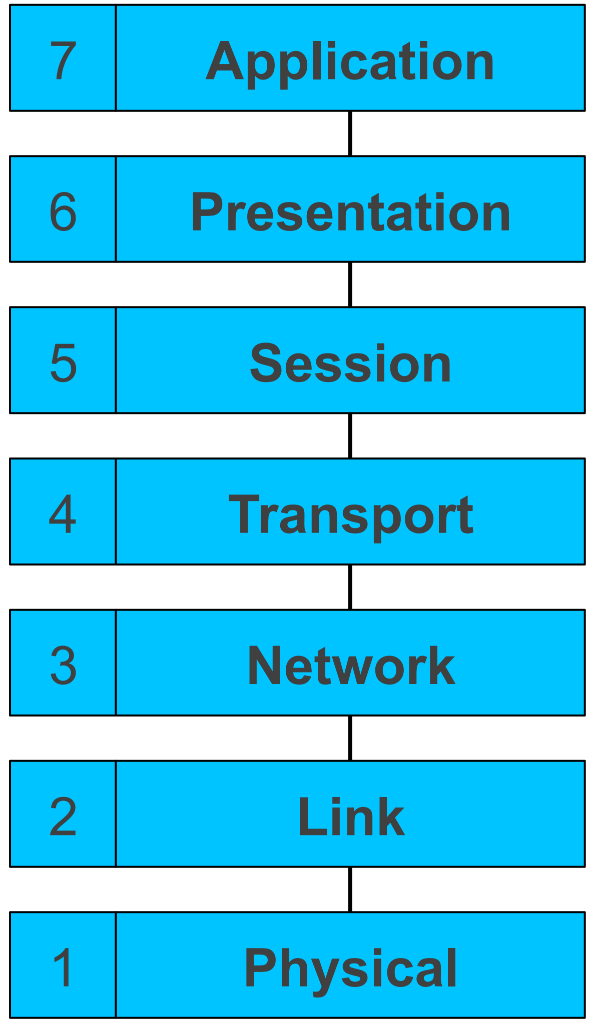
Networking Fundamentals

What is this


7 Application
6 Presentation
5 Session
4 Transport
3 Network
2 Link
1 Physical
?

Network Stack

Network Stack Model



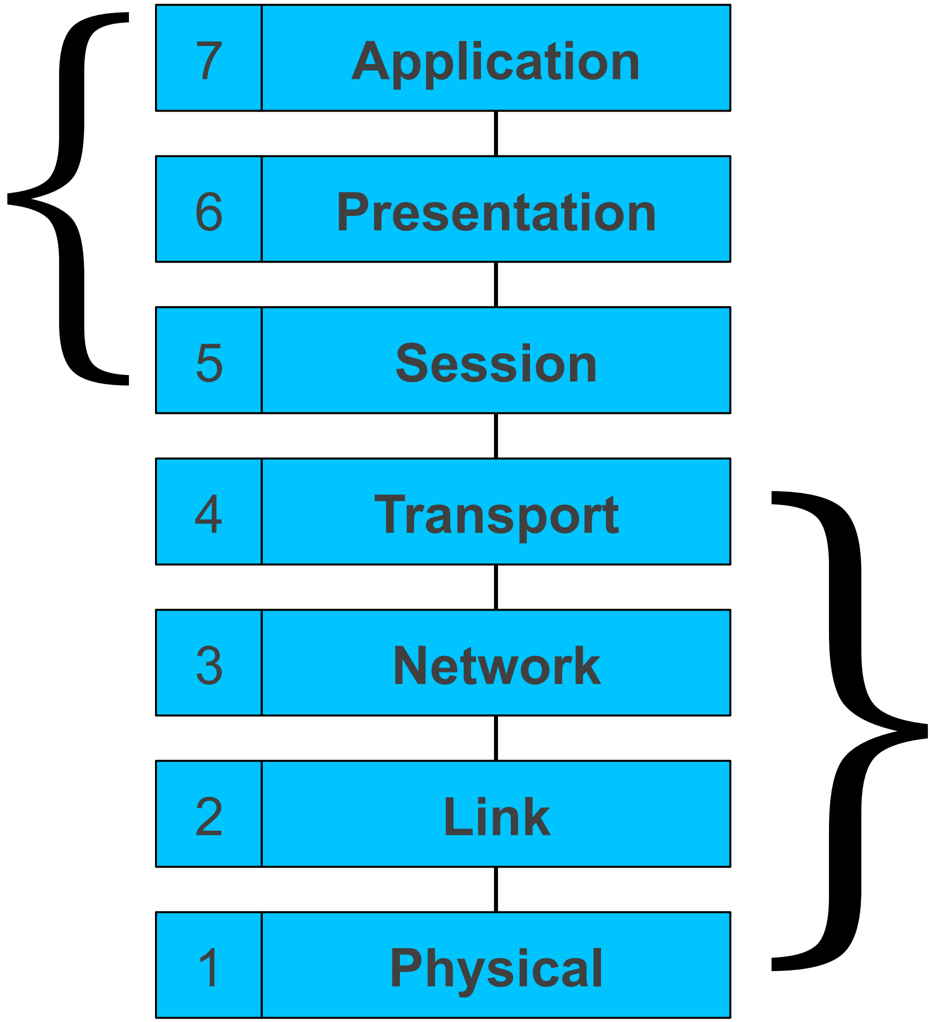
OSI Model levels

- Higher (apps)
- Lower (data flow)

Network Stack

Upper Layers

- users interaction
- Implement apps
- Relay on lower lever to deliver data



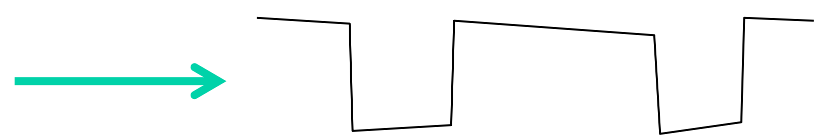
Lower Layers

- formatting, encoding and transmit data
- Don't care about data, just moving around

Layer 1: Physical Layer

- Transfers a stream of *bits*
- Defines physical characteristics
  - Connectors, pinouts
  - Cable types, voltages, modulation
  - Fibre types, lambdas
  - Transmission rate (bps)
- No knowledge of bytes or frames

101101



Qns: - What are the equipment operate over the layer?
- What challenge on this layer?

Layer 2: (Data)Link Layer

- Organises data into *frames*
- <u>May</u> detect transmission errors (corrupt frames)
- <u>May</u> support shared media
  - Addressing (unicast, multicast) – who should receive this frame
  - Access control, collision detection
- Usually identifies the layer 3 protocol being carried

Layer 3: (Inter)Network Layer

- Connects Layer 2 networks together
  - Forwarding data from one network to another
- Universal frame format (datagram)
- Unified addressing scheme
  - Independent of the underlying L2 network(s)
  - Addresses organised so that it can scale globally (aggregation)
- Identifies the layer 4 protocol being carried
- Fragmentation and reassembly

Layer 4: Transport Layer

- Identifies the *endpoint process*
  - Another level of addressing (port number)
- <u>May</u> provide reliable delivery
  - Streams of unlimited size
  - Error correction and retransmission
  - In-sequence delivery
  - Flow control
- Or might just be unreliable datagram transport

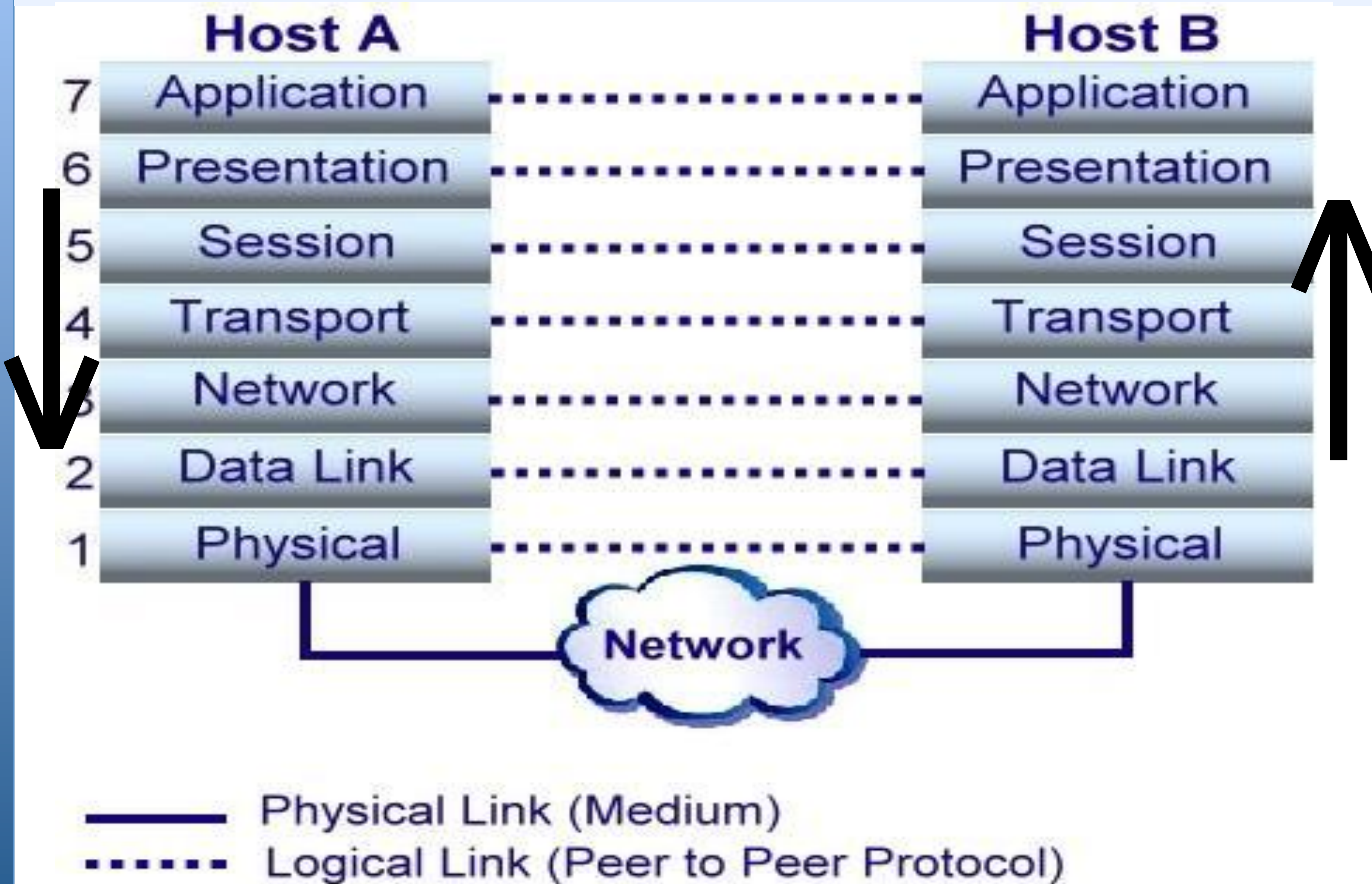
Layers 5 and 6

- Session Layer: long-lived sessions
  - Re-establish transport connection if it fails
  - Multiplex data across multiple transport connections
- Presentation Layer: data reformatting
  - Character set translation
- Neither exist in the TCP/IP suite: the application is responsible for these functions

Layer 7: Application layer

- The actual work you want to do
- Protocols specific to each application
- *Examples?*

Encapsulation & De-encapsulation



OSI in summary

- Layer 7 – Application (servers & clients, web browsers, httpd)
- Layer 6 – Presentation (file formats, e.g. PDF, ASCII, JPEG)
- Layer 5 – Session (conversation initialization, termination)
- Layer 4 – Transport (inter host comm – error correction)
- Layer 3 – Network (routing – path determination, IP addresses)
- Layer 2 – Data link (switching – media access, MAC addresses)
- Layer 1 – Physical (signaling – representation of binary digits)

Routing Concepts

- Hints:
  - IP Addressing
    - IPv4
    - IPv6
  - Routing & Forwarding

IP Addressing

- What do the addresses look like?
- How do they get allocated, to avoid conflicts?
- Examples to consider:
  - L2: Ethernet MAC addresses
  - L3: IPv4, IPv6 addresses
  - L4: TCP and UDP port numbers

IP Addressing

- What do the addresses look like?
- How do they get allocated, to avoid conflicts?
- Examples to consider:
  - L2: Ethernet MAC addresses
  - L3: IPv4, IPv6 addresses
  - L4: TCP and UDP port numbers

IPv4 Addresses

- 32-bit binary number
  - How many unique addresses in total?
- Conventionally represented as four dotted decimal octets

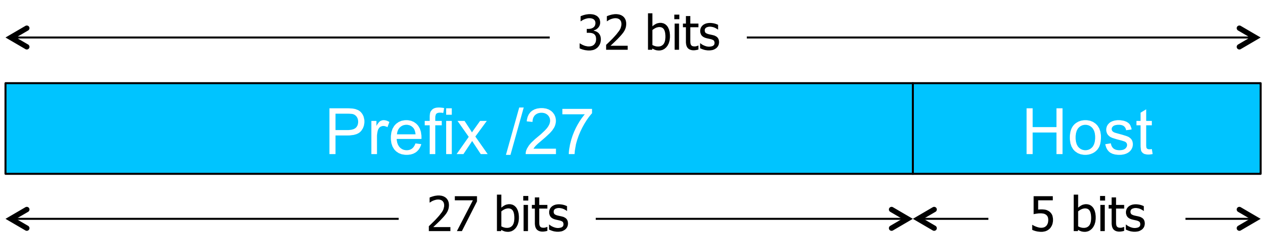
10000000110111111001110100010011

128 . 223 . 157 . 19

Hierarchical Division IP Addresses

- Network Part (Prefix): Describes which network
- Host Part (Host Address): Describes which host
- Boundary can be anywhere!
  - Used to be a multiple of 8

Prefixes



- A range of IP addresses is given as a *prefix*, e.g. 192.0.2.128/27
- In this example:
  - How many addresses are available?
  - What are the lowest and highest addresses?

Prefix calculation

192 . 0 . 2 . 128

11000000000000000000001010000000

Prefix length /27 ➔ First 27 bits are fixed

Lowest address:

11000000000000000000001010000000

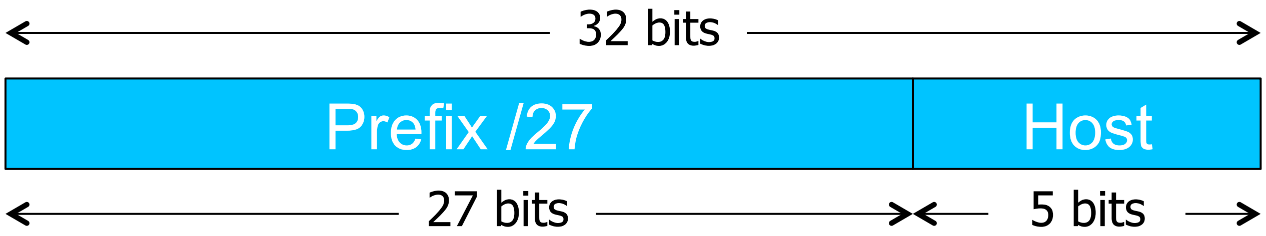
192 . 0 . 2 . 128

Highest address:

11000000000000000000001010011111

192 . 0 . 2 . 159

IPv4 “Golden Rules”



1. All hosts on the same L2 network must share the *same* prefix
2. All hosts on the same subnet have *different* host part
3. Host part of all-zeros and all-ones are reserved

Golden Rules for 192.0.2.128/27

Lowest 192.0.2.128 = network address

Highest 192.0.2.159 = broadcast address

Usable: 192.0.2.129 to 192.0.2.158

Number of usable addresses: 32 - 2 = 30

Subnetting Example

- You have been given 192.0.2.128/27
- However you want to build two Layer 2 networks and route between them
- The Golden Rules demand a different prefix for each network
- Split this address space into two equal-sized pieces
  - What are they?

IPv6 Addresses

- 128-bit binary number
- Conventionally represented in hexadecimal
  – 8 words of 16 bits, separated by colons

2001:0468:0d01:0103:0000:0000:80df:9d13

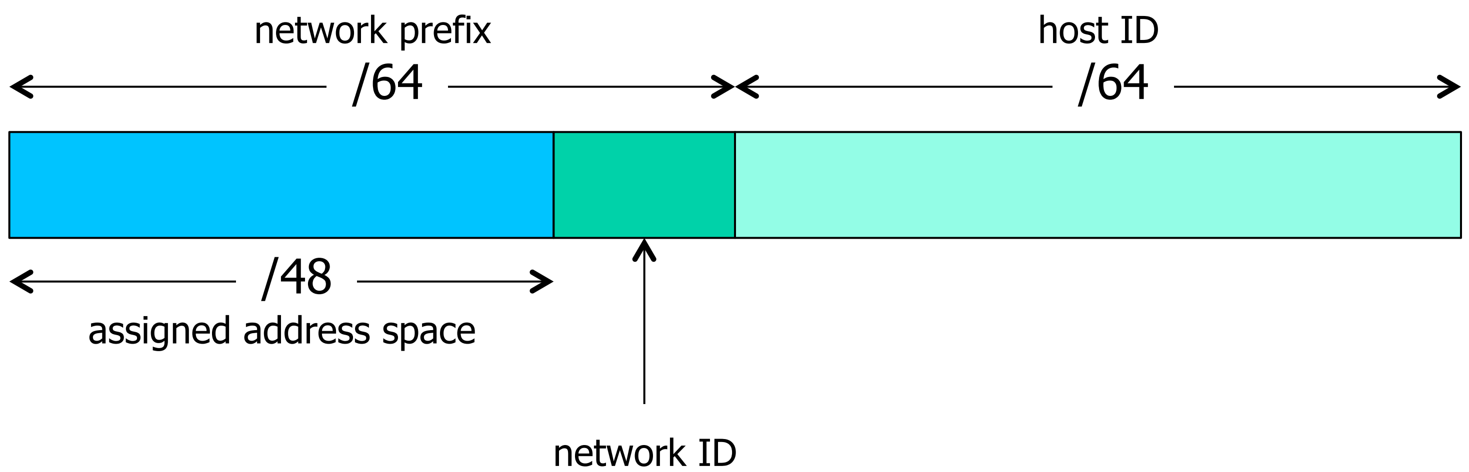
- Leading zeros can be dropped
- One contiguous run of zeros can be replaced by ::

2001:468:d01:103::80df:9d13

IPv6 Rules

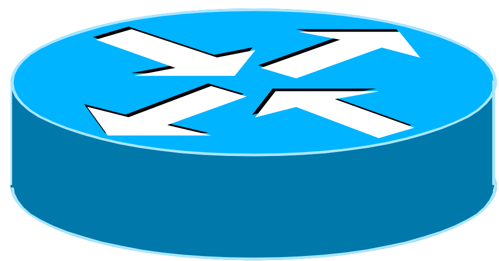
- With IPv6, every network prefix is /64
  - (OK, some people use /127 for P2P links)
- The remaining 64 bits can be assigned by hand, or picked automatically
  - e.g. derived from NIC MAC address
- There are special prefixes
  - e.g. link-local addresses start fe80::
- Total available IPv6 space is ≈ 2^{61} subnets
- Typical end-user allocation is /48 (or /56)

IPv6 addressing



How many /64 networks can you build given a /48 allocation?

What does a router do?



A day in a life of a router

find path

forward packet, forward packet, forward packet, forward packet...

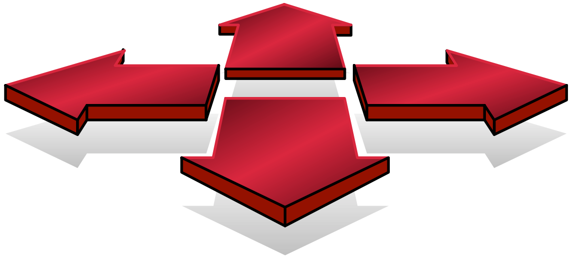
find alternate path

forward packet, forward packet, forward packet, forward packet…

repeat until powered off

Routing vs Forwarding

Routing = building maps and giving directions



Forwarding = moving packets between interfaces according to the "directions"

IP Routing – finding the path

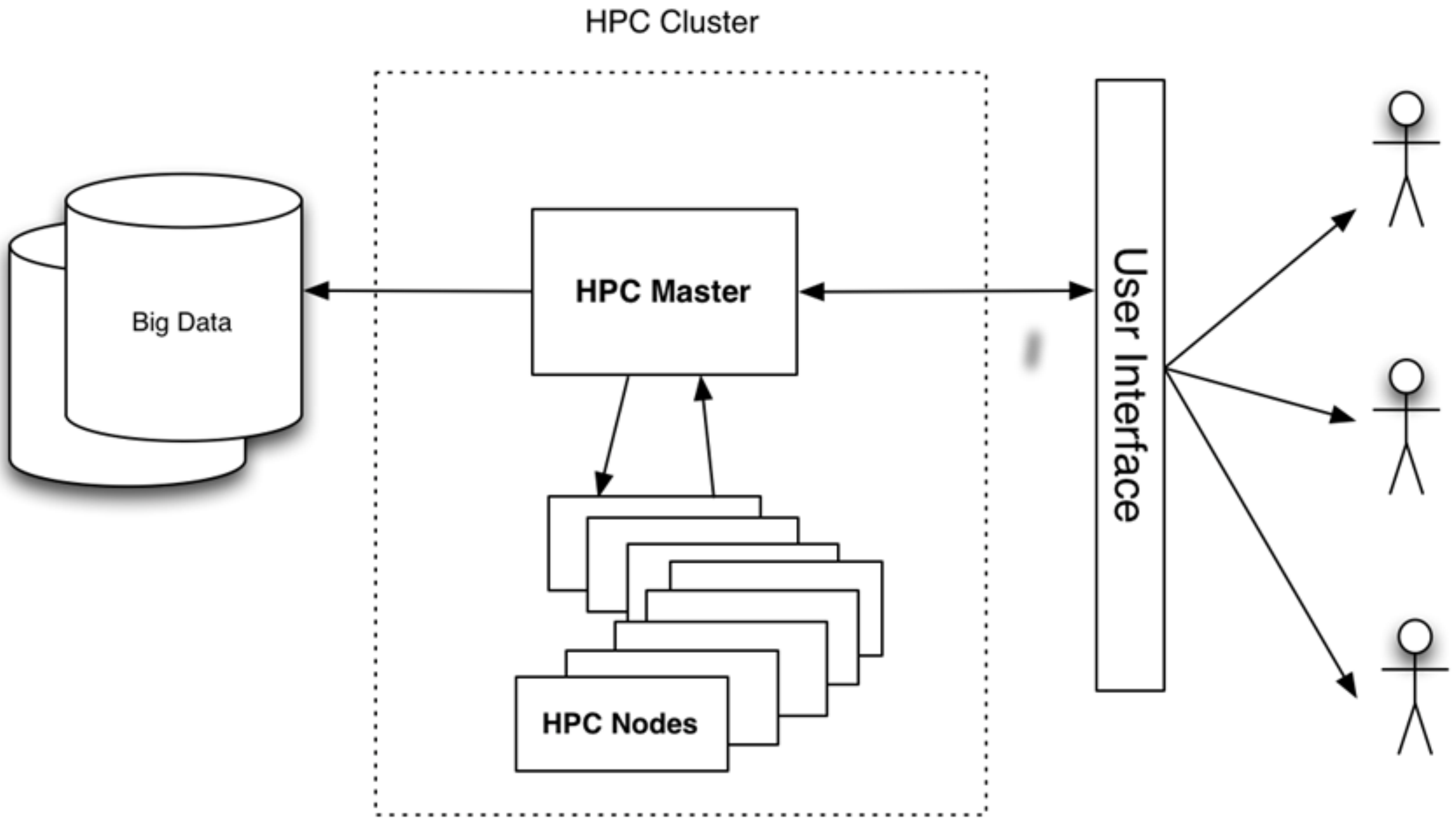
- Path derived from information received from a routing protocol
- Several alternative paths may exist
  - Best path stored in **forwarding** table
- Decisions are updated periodically or as topology changes (event driven)
- Decisions are based on:
  - Topology, policies and metrics (hop count, filtering, delay, bandwidth, etc.)

IP Forwarding

- Router decides which interface a packet is sent to
- Forwarding table populated by routing process
- Forwarding decisions:
  - destination address
  - class of service (fair queuing, precedence, others)
  - local requirements (packet filtering)
- Forwarding is usually aided by special hardware

Network for clusters

Designing Hints



Designing concepts

Topology

– Rules determining how compute nodes and network nodes are connected

–Unlike LAN or data center networks, HPC topologies are highly regular

– All connections are full duplex

Designing concepts cont….

Routing

– Rules determining how to get from node A to node B

– Because topologies are regular & known, routing algorithms can be designed a priori

– Source- vs. table-based; direct vs. indirect; static vs. dynamic; oblivious vs. adaptive

Designing concepts cont….

Flow control

– Rules governing link traversal

– Deadlock avoidance

Interconnect classification

HIGH SPEED NETWORK

- parallel computation
- low latency /high bandwidth
- Usual choices; infiniband..

I/O NETWORK

- I/O requests (NFS and/or parallel FS)
    - Latency not fundamental / good bandwidth
    - Gigabit could be ok

Interconnect classification cont..

MANAGEMENT NETWORK

– management traffic

any standard network (fast ethernet Ok)

Interconnect classification cont..

Standard vs. proprietary

– Standard ("open"): Network technology is compliant with a specification ratified by standards body (e.g., IEEE)

– Proprietary ("closed"): Network technology is owned & manufactured by one specific vendor

Characteristics of a network

Topology

- Diameter
- Nodal Degree
- Bisection bandwidth

Performance

- Latency
- Link bandwidth

Topology

- How the components are connected.
- Important properties
  - Diameter: maximum distance between any two nodes in the network (hop count, or # of links).
  - Nodal degree: how many links connect to each node.
  - Bisection bandwidth: The smallest bandwidth between half of the nodes to another half of the nodes.

A good topology: small (diameter + nodal), large bisection bandwidth

Bisection bandwidth

- Split N nodes into two groups of N/2 nodes such that the bandwidth between these two groups is minimum: that is the bisection bandwidth

Why is Bisection Bandwidth?

- if traffic is completely random, the probability of a message going across the two halves is ½
- if all nodes send a message, the bisection bandwidth will have to be N/2
- The concept of bisection bandwidth confirms that some network topology network is not suited for random traffic patterns
- your worst case scenario of HPC workload is to have random traffic patterns..

Latency in Networking

- Latency is the delay between the time a frame begins to leave the source device and when the first part of the frame reaches its destination. A variety of conditions can cause delays:
  - Media delays may be caused by the finite speed that signals can travel through the physical media.
  - Circuit delays may be caused by the electronics that process the signal along the path.

Latency in Networking cont..

- Software delays may be caused by the decisions that software must make to implement switching and protocols.

Latency in HPC

- The one-way latency may be also meant as the period of time that a 0-sized message spends traveling from its source to its destination,
- It involves the time needed to:
  - Encode,
  - send the packet,
  - receive the packet,
  - decode

Bandwidth & Speed

- Bandwidth;
  - the measure of the amount of information that can move through the network in a given period of time.
- How wide is my channel ?

Warning:

Speed

- is often used interchangeably with bandwidth, but a large-bandwidth device will carry data at roughly the same speed of a small-bandwidth device if only a small amount of their data-carrying capacity is being used.

What we need from High Speed Networks ?

- Intelligent Network Interface Cards
  - Support entire protocol processing completely in hardware (hardware protocol offload engines)
- Provide a rich communication interface to applications
  - User level communication capability
  - Gets rid of intermediate data buffering requirements

What we need from High Speed Networks ?

- No software signaling between communication layers
    - All layers are implemented on a dedicated hardware unit, and not on a shared host CPU

What is InfiniBand?

- Inductry Standard defined by the InfiniBand Trade Association – Originated in 1999
- InfiniBand Specification defines and input/ output architecture used to interconnect servers, communications infrastructure equipments, storage and embedded systems.
- InfiniBand is a pervasive, low-latency, high-bandwidth interconnect which requires low processing overhead and is ideal to carry multiple traffic type (clustering etc)

What is InfiniBand? Cont..

- InfiniBand is now used in thousands of High Performance Compute clusters.

InfiniBand Architecture

- Defines Systems Area Network architecture
  - Comprehensive specs: from physical to applications processor
- Architecture supports
  - Host Channel Adapters
  - Switchers /Routers
- Facilitated HW design for
  - Low latency / high bandwidth
  - Transport offload

Which network for your clusters?

- Some questions to help
  - Which kind of cluster (HTC or HPC)?
  - Which kind of application?
    - Serial / parallel
    - Latency or bandwidth dominated?
- Input/Output considerations
  - Only MPI or storage as well?
- Budget consideration

Thanks

Ahsante

$\Rightarrow$ **For discussion**

For discussion

- Can you give examples of equipment which operates at layer 4? At layer 7?
- At what layer does a wireless access point work?
- What is a “Layer 3 switch”?
- How does traceroute find out the routers which a packet traverses?

For discussion cont…..

- Network 10.10.10.0/25

  How many addresses in total?

  How many usable addresses?

  What are the lowest and highest usable addresses?

- Network 10.10.20.0/22

  How many addresses in total?

  How many usable addresses?

  What the the lowest and highest usable addresses?

For discussion cont.....

- What is the purpose of TCP/UDP port numbers?

  a. indicate the beginning of a three-way handshake

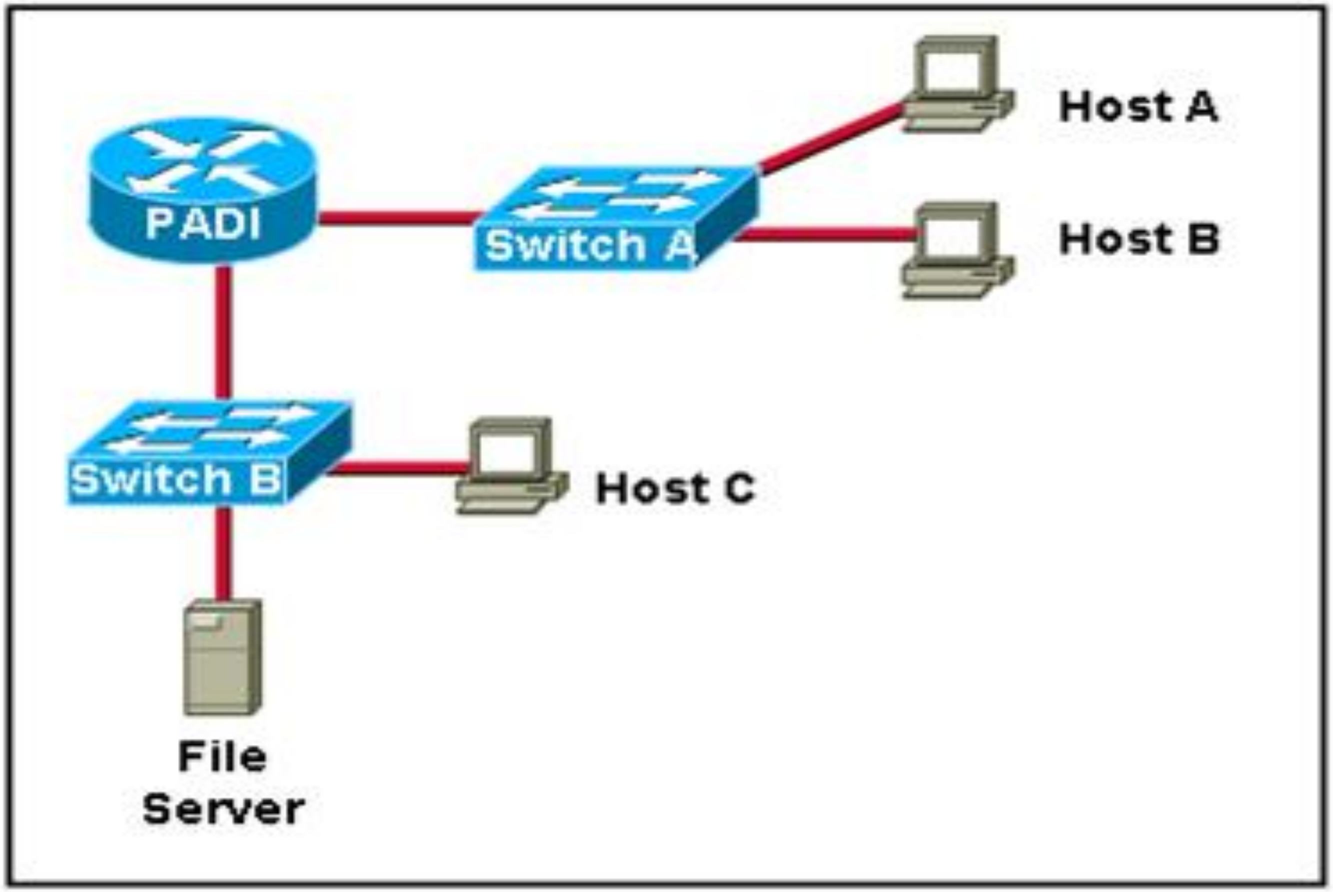
  b. reassemble the segments into the correct order

  c. identify the number of data packets that may be sent without acknowledgment

  d. track different conversations crossing the network at the same time

For discussion cont…..

Exhibit



For discussion cont.....

- Refer to the exhibit (previous slide). What must be configured on Host B to allow it to communicate with the file server? (Choose three.)
  - a. the MAC address of the file server
  - b. the MAC address of the PADI router interface connected to Switch A
  - c. the IP address of Switch A
  - **d**. a unique host IP address
  - **e**. the subnet mask for the LAN
  - **f**. the default gateway address

And so many as can be requested !

dmakweba@dit.ac.tz