ALBERT-LUDWIGS-
UNIVERSITÄT FREIBURG

Communication Systems

Ethernet

University of Freiburg
Computer Science
Computer Networks and Telematics
Prof. Christian Schindelhauer

CoNe Freiburg

IIF
INSTITUT FÜR
INFORMATIK
FREIBURG

Copyright Warning

- **This lecture is already stolen**
- **If you copy it please ask the author**
  - Prof. Dr. Gerhard Schneider
- **like I did**

Communication Systems Organization and Q&A

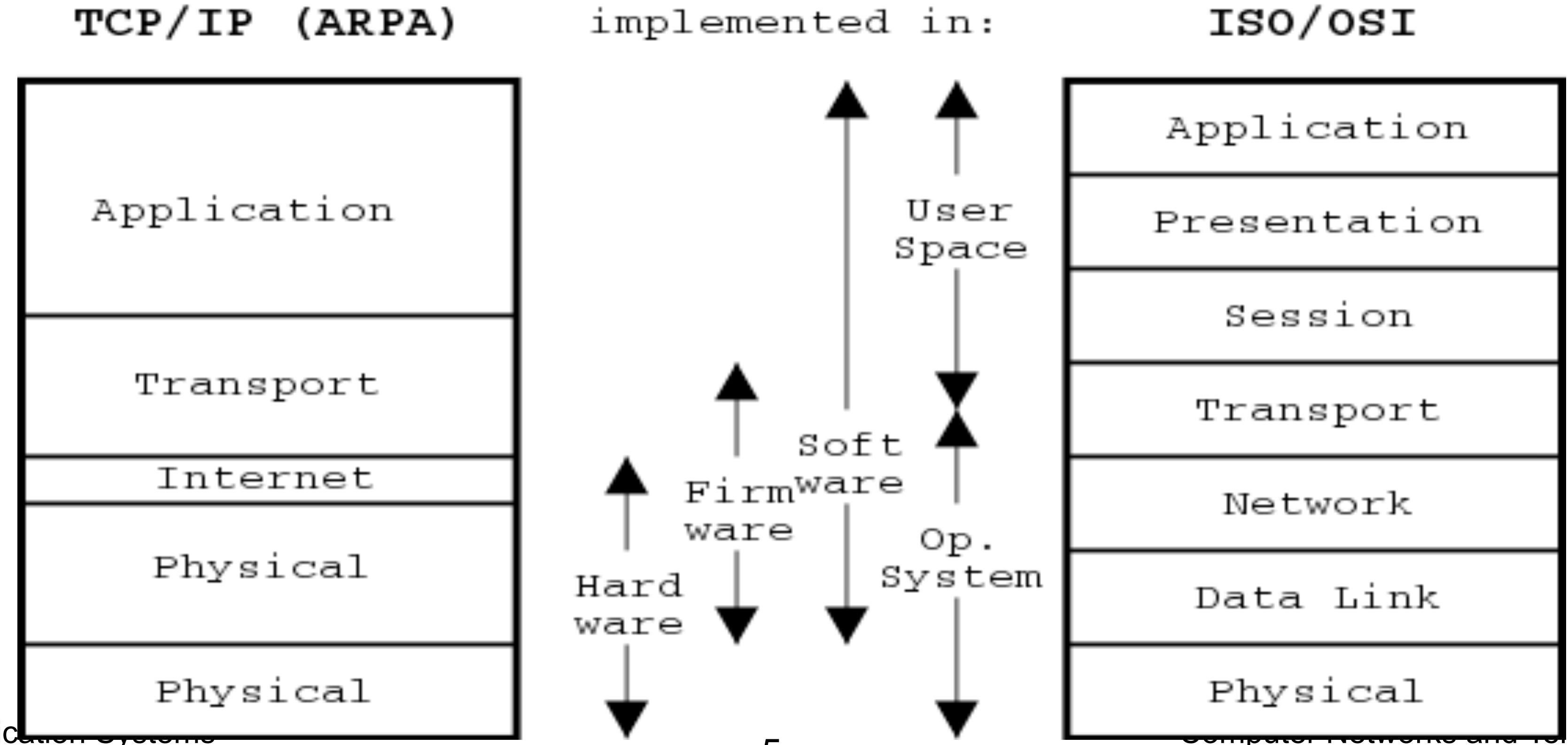
- **Everybody should have gotten two emails of the comsysWS08 mailing list**
  - Communicating organizational issues
  - Giving literature hints
  - Handing out the (non-mandatory) theoretical questions for upcoming lecture
- **Any questions from last lecture?**
- **Questions on lecture organization?**

Communication Systems
Q&A of exercises / homework

- What is a protocol, what do we have standards for?
- Why stacks of protocols?
- Draw ISO/OSI protocol stack, compare it to TCP/IP stack of Tanenbaum!

Communication Systems
Q&A of exercises / homework

- **Comparison of both stacks, typical implementation methods (ARPA is four layers; Tanenbaum stack lists the Data Link Layer comparable to ISO/OSI)**

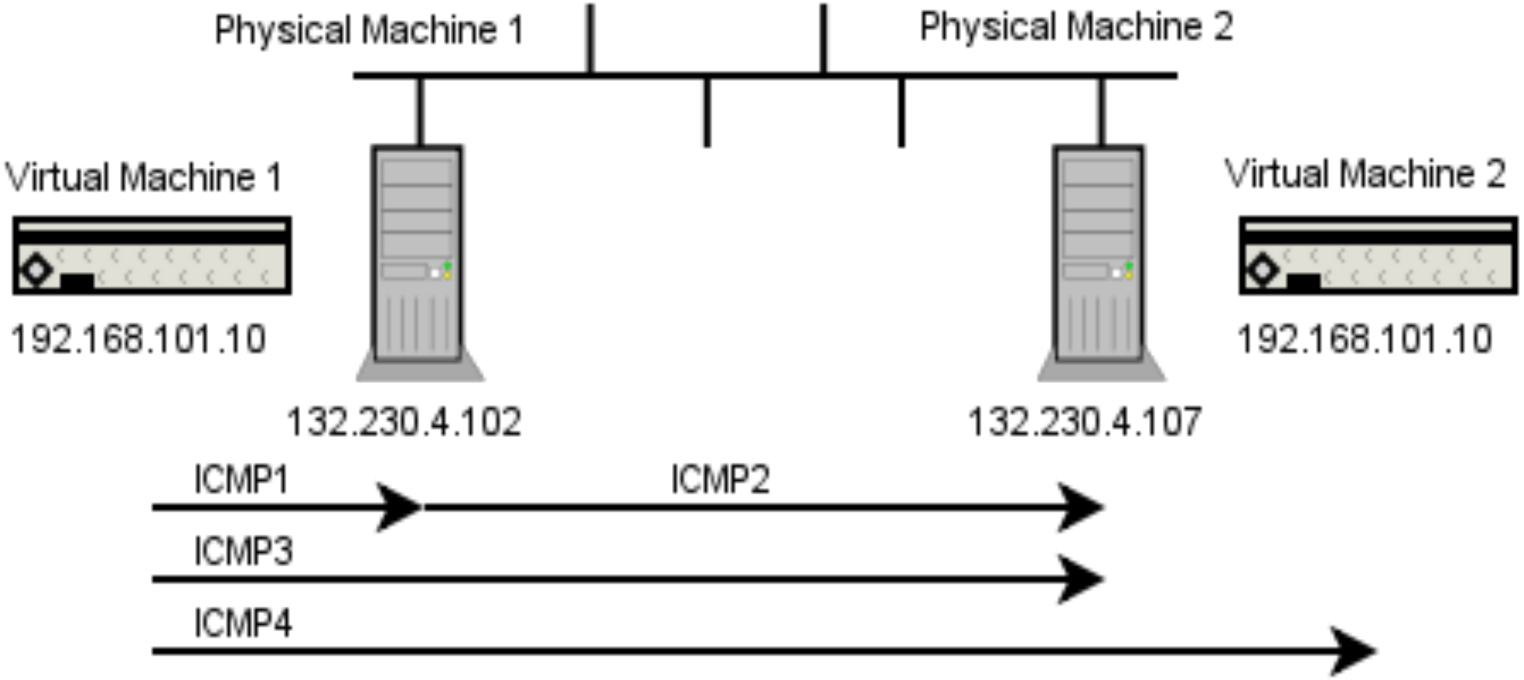


Communication Systems
Q&A of exercises / homework

- **How it is possible to deploy duplicated IP addresses?**

Communication Systems
Q&A of exercises / homework

- **More on NAT and related issues in the static IP routing lecture**
- **Which pings (IP / ICMP messages) would be properly exchanged, which would fail!?**



Communication Systems
Ethernet – the physical and data link

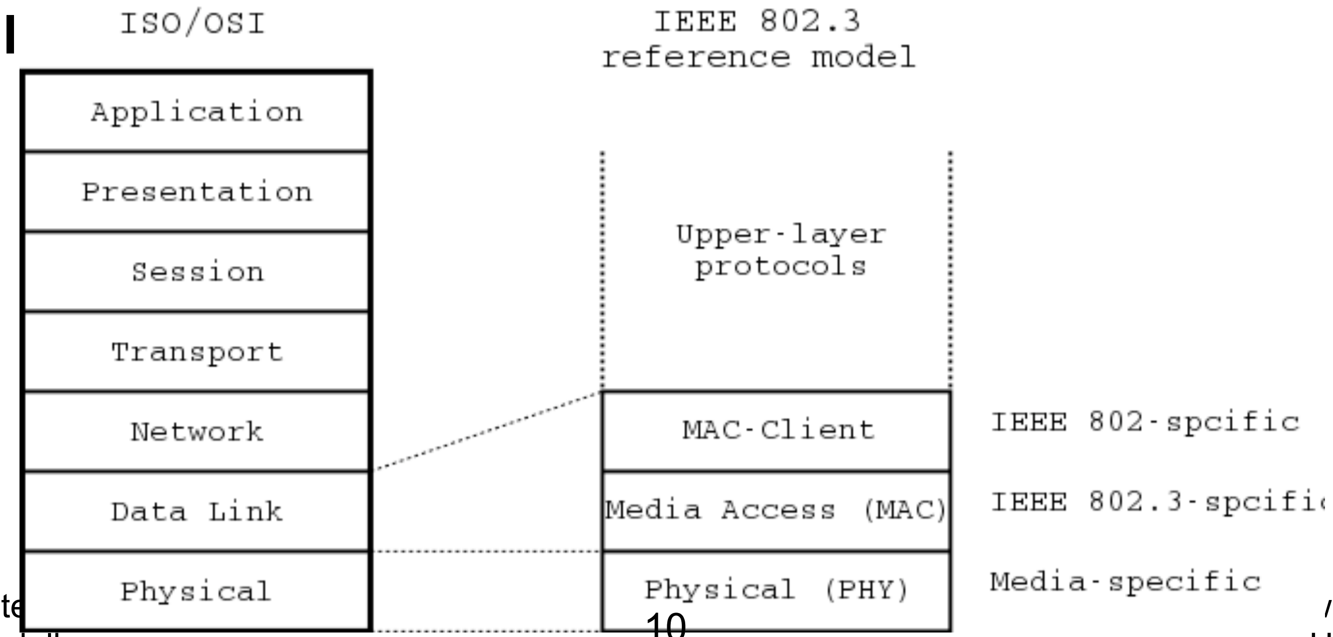
- **Low level packet networking standard evolved end of seventies**
- **A standard to cope with rising numbers of locally networked machinery was needed (point-2-point connections as in the first iterations of Internet wouldn't do in long run)**
- **Started among other technologies like TokenRing, ArcNET or AppleTalk but left behind all these technologies in the mid 1990s with the creation of the 100Mbit/s standard**
- **Disputes in the CSMA (Carrier Sense Multiple Access) groups, whether**
  - CD (Collision Detection) as in Ethernet
  - Or CA (Collision Avoidance) as in TokenRing and alike is the better strategy

Communication Systems
Ethernet – the physical and

- **CSMA/CD protocol was originally developed as a means by which two or more stations could share a common media in a switch-less environment. Its specifics:**
  - The protocol does not require central arbitration, access tokens, or assigned time slots to indicate when a station will be allowed to transmit.
  - Each Ethernet MAC determines for itself when it will be allowed to send a frame
- **Check for a more in depth explanation of CSMA/CD in background literature!**

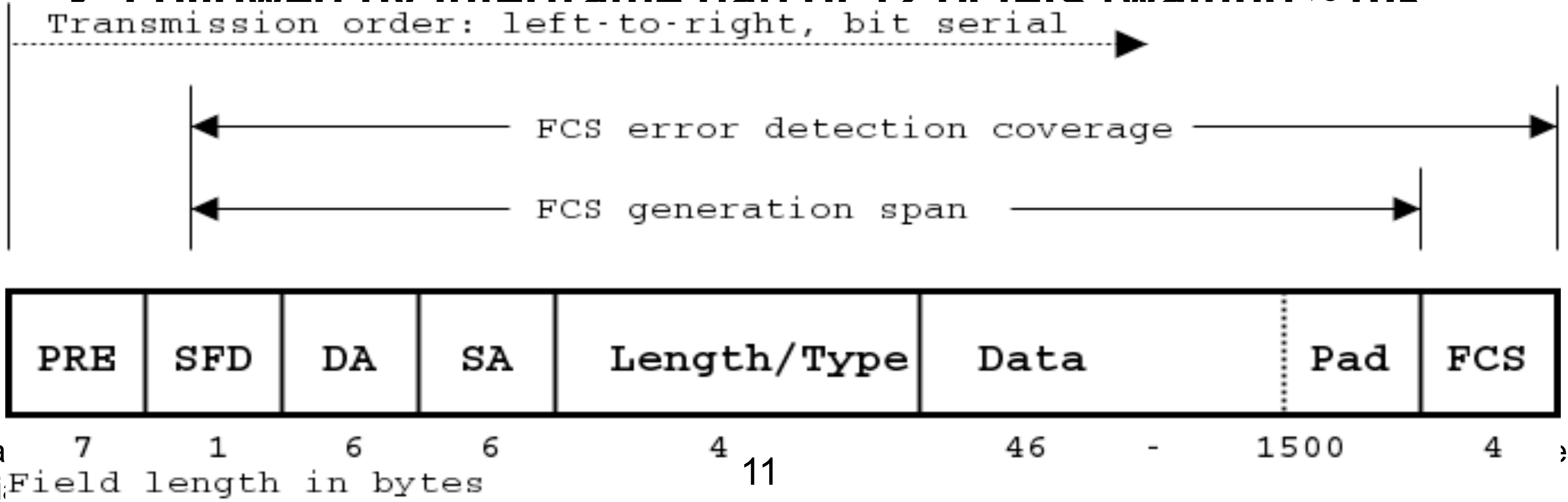
Communication Systems
Ethernet – the physical and

- LAN protocol family which refers to the IEEE 802.3 standards
- IEEE 802 protocols devide the ISO data link layer into two IEEE 802 sublayers, the Media Access Control (MAC) sublayer and the MAC-client sublayer. The IEEE 802.3 physical layer corresponds to the ISO physical l



Communication Systems
Ethernet – the physical and

- For modulation, base band encoding and the wiring standards of early versions refer to Systems II lecture
- Basic data frame format required for all MAC implementations of interest here as it is not changed up to the 10GBit/s version
- In basic version (standard extensions next lecture) seven fields
- Followed by interframe gap of 12 octets (waiting ~10k



Transmission order: left-to-right, bit serial

FCS error detection coverage

FCS generation span

PRE	SFD	DA	SA	Length/Type	Data	Pad	FCS
7	1	6	6	4	46 - 1500		4

Field length in bytes

Communication Systems Ethernet – the physical and

- There is a preamble (PRE) of seven octets of alternating zeros and ones – for what reason?

Communication Systems
Ethernet – the physical and data link

- **The preamble (PRE) tells receiving stations that a frame is coming, and that provides a means to synchronize the frame-reception portions of receiving physical layers with the incoming bit stream**
- **Start-of-frame delimiter (SOF) consists of 1 Byte. The SOF is an alternating pattern of ones and zeros, ending with two consecutive 1-bits indicating that the next bit is the left-most bit in the left-most Byte of the destination address**
- **Destination address (DA) - 6 Bytes identifies which station(s) should receive the frame**
  - left-most bit in the DA field indicates whether the address is an individual address (indicated by a 0) or a group address (indicated by a 1)

Communication Systems
Ethernet – the physical and data link

- **Destination address (DA)**
  - Second bit from the left indicates whether the DA is globally administered (indicated by a 0) or locally administered (indicated by a 1)
  - Remaining 46 bits are a uniquely assigned value that identifies a single station, a defined group of stations, or all stations on the network
  - Prefixes are world wide unique, assigned by FCC, thus possible to distinguish the chips/adapters of different vendors
  - Why locally/LAN-wide uniqueness is enough? Why nevertheless a world wide assignment? Why machine virtualization produces new challenges on MAC address assignment?
  - Addresses are programmed to the adapters, but are easy to change (experimental part of the lecture)

Communication Systems
Ethernet – the physical and data link

- **Source addresses (SA) - Consists of 6 Bytes**
  - The SA field identifies the sending station
  - Source address is always an individual address and the left-most bit in the SA field is always 0
- **Length/Type - Consists of 4 Bytes**
  - This field indicates either the number of MAC-client data Bytes that are contained in the data field of the frame, or the frame type ID if the frame is assembled using an optional format
  - If the Length/Type field value is less than or equal to 1500, the number of LLC bytes in the Data field is equal to the Length/Type field value
  - If the Length/Type field value is greater than 1536, the frame is an optional type frame, and the Length/ Type field value identifies the particular type of frame being sent or received

Communication Systems
Ethernet – the physical and data link

- **Type defines the next layer protocol to hand the packets over, e.g. 0X800hex for IPv4, see www.iana.org/assignments/ethernet-numbers for other types or check with wireshark for ARP and IPv6 (in the practical part)**
- **Data - Is a sequence of n bytes of any value, where n is less than or equal to 1500 (this is the MTU max. transfer unit - size reported to upper layers, see todays exercise on MTU)**
  - If the length of the Data field is less than 46, the Data field must be extended by adding a filler (a pad) sufficient to bring the Data field length to 46 Bytes
  - Minimum length of a packet, why?
- **Frame check sequence (FCS) of 4 Bytes length, 32-bit cyclic redundancy check (CRC), is generated over the DA, SA, Length/Type, and Data fields**

Communication Systems
Ethernet – collisions and detection

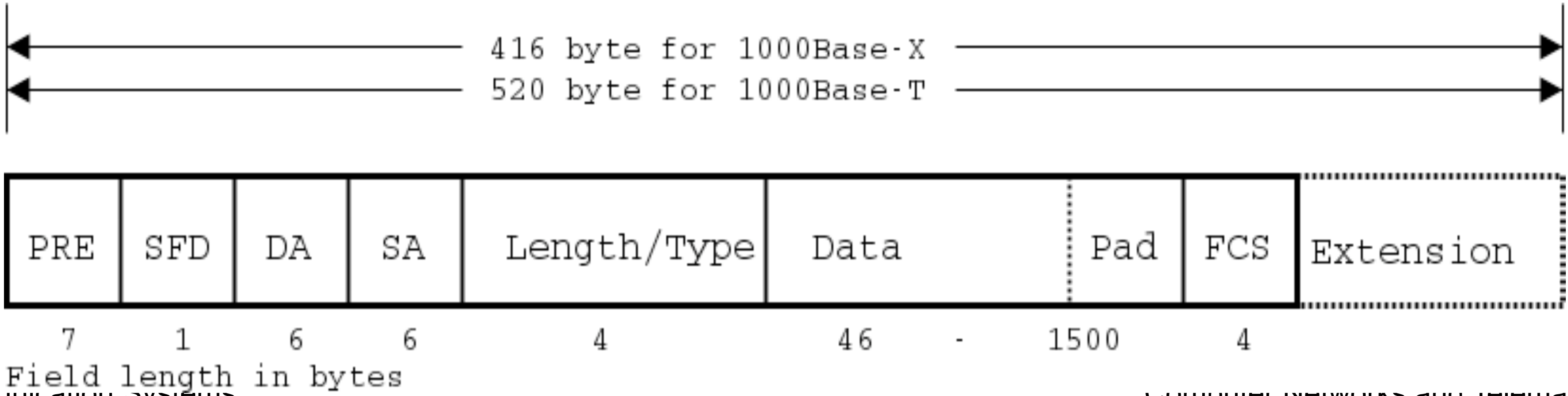
- **Packet collisions in the several Ethernet standards**
  - If collision happens, each transmitting station must be capable of detecting that a collision has occurred before it has finished sending its frame
- **Concept not without problems, especially for the newer standards**
  - Worst-case situation is given when the two most-distant stations on the network need to send a frame
  - First sends, second starts little later (cable seems to be free), collision almost immediately near the second station, but corrupted signal has to spread way back so the first can acknowledge it
  - Maximum time for detection (collision window) must be estimated (twice of signal end to end propagation time)
  - Minimum frame length and the maximum collision diameter are directly related to the slot time

Communication Systems
Ethernet – collisions and detection

- Longer minimum frame lengths translate to longer slot times and larger collision diameters
- Shorter minimum frame lengths correspond to shorter slot times and smaller collision diameters
- Defined network diameter of 2500m with 10Mbps
- Problems occur with this setup at speeds of 100 and 1000Mbit/s, because time required to transmit a frame is inversely related to the transmission rate
- 100Mbit/s therefor only defined for Twisted Pair media with reduced length of roughly one tenth (200m: e.g. computer- hub -computer)
- Decreasing network diameters by another factor of 10 (to approximately 20m) for 1000Mbit/s operation is simply not practical

Communication Systems Ethernet – collisions and

- **Collision domain diameters like for 100Mbit/s networks should have been maintained (using same cabling ...)**
- **The apparent minimum frame size is increased by adding a variable-length nondata extension field to frames that are shorter than the minimum length (the extension field is removed during frame reception)**
- **Concept is changed for 10GBit/s for obvious reasons**



PRE	SFD	DA	SA	Length/Type	Data	Pad	FCS	Extension
7	1	6	6	4	46 -	1500	4	

Communication Systems
Ethernet – the physical and

- Why only half duplex connection in coaxial infrastructure?
- How full duplex in twisted pair setups?

Communication Systems
Ethernet – full duplex operation

- **Full duplex transmission is functionally much simpler than half-duplex transmission**
  - Each direction is its own collision domain with just one defined sender and one receiver
  - Involves no media contention, no collisions
  - No need to schedule retransmissions
  - No need for extension bits on the end of short frames
- **The result is not only more time available for transmission, but also an effective doubling of the link bandwidth because each link can now support full-rate, simultaneous, two-way transmission**
- **Only restriction is the need for a minimum-length interframe gap between successive frames**

Communication Systems
Ethernet – network bridges

- **Traditional segment connection method: hubs**
  - Doing simple signal refreshing - repeater functionality, amplifying and reconstruction of signals
  - Allow star topology to avoid electrical interference when (dis)connecting LAN stations
- **To extend the maximum network diameters or eleminate occurance of collisions special network components required**
- **Simple implementation: Ethernet bridge**
  - Separation of segments (collision domains)
  - Learn where devices are by watching MAC addresses
  - Keep intrasegment traffic away from the other segment(s)
  - Typically software implementation (bridging was possible to non-ethernet type packet networks too)

Communication Systems
Ethernet – switches

- **Software bridge in use of todays experiments setup: VMware implements a software Ethernet bridge allowing packets of the virtual machine to be directly handed over onto the physical LAN**
  - Setting up of a (software) bridge in Linux – next lecture
- **Implementation of full duplex star topology is done with switches**
- **Network components for wire speed packet transfers (software bridges often slow under high load)**
- **Switches implement**
  - Store and forward for avoidance of collisions and speed adaption for different data rates, needed if machinery of different Ethernet speeds is added to a single LAN
  - Virtual point-to-point connections between hosts (connecting the ports involved through MAC address storing)

Communication Systems
Ethernet – switches

- **Store and forward**
  - Packet has to be received completely, before it is sent out
  - Delay is L/R (packet size divided by data rate)
- **Cut-through**
  - If output queue is empty, the switch sends out the packet after receiving the destination address immediately
- **Cut-through may reduce connection delays**
- **While every packet in Ethernet is seen by every LAN station, switches introduce a little bit of "privacy”**

- Packets typically only delivered to the port intended for
- Only broadcast traffic still sent to all stations at same time
- Easy to circumvent, as demonstrated in a later exercise

Communication Systems
Ethernet – higher speeds and jumbo

- **1GBit/s Ethernet**
  - Half duplex still allowed (CD implemented), but not really used, e.g. no Gigabit hubs (there were only few Fast Ethernet hubs in the market)
  - Autonegotiation requirement
  - Specific addition: frame bursting
- **Burst mode is a feature that allows a MAC to send a short sequence (a burst) of frames equal to approximately 5.4 maximum-length frames without having to relinquish control of the medium**
- **The transmitting MAC fills each interframe interval with extension bits, so that other stations on the network will see that the network is busy and will not attempt transmission until after the burst is complete**

Communication Systems
Ethernet – higher speeds and jumbo

- **Original of 1500 MTU was due to slow communication speeds and high error rates (think of collision probability in unswitched Ethernets)**
- **Header overhead in for large IP data chunks**
  - MTU dictates packet sizes of higher level protocols and fragmentation in IPv4
- **Jumbo frames special Ethernet frames with more than 1,500 bytes of payload (MTU)**
  - Supported by many (but not all) Gigabit Ethernet switches and network interface cards
  - Not supported in Fast Ethernet, so mixed operation with both does not allow jumbo frames
  - Thus not really used at the moment

Communication Systems
Ethernet – autonegotiation

- **Autonegotiation – speciality of twisted pair Ethernet wiring**
- **Mixed Ethernet version operation needs link adaptations**
  - Half duplex / full duplex LAN stations
  - 10 / 100 / 1000 Mbit/s
- **Auto-negotiation sublayer allows the NICs at each end of the link to exchange information about their individual capabilities**
- **Then to negotiate and select the most favorable operational mode that they both are capable of supporting.**
- **Auto-negotiation is optional in early Ethernet implementations and is mandatory in later versions**

Communication Systems
Ethernet and IP

- **Ethernet handles physical connection of most Internet hosts**
  - But using flat addressing scheme
  - Number of machines in an Ethernet could be not extended infinitely (typical switch can store between 8 and 16/32k of MACs)
  - Think of broadcast floods (for e.g. ARP – later lecture)
  - Restricted network diameter (even with long range Ethernet physical layer in fiber optics)
- **Ethernet addresses fixed to hardware**
  - Automatic setup of networking
- **IP addressing – higher layer, to overcome the named restrictions: routed/hierarchical**
  - Manual setup/configuration needed (todays practical)

Communication Systems
The exercise environment

- **You should have grabbed practical exercise sheet #2 passed by**
- **This exercise block will require some more setup than the last one**
  - Updated special Linux image which will run in a virtual machine (VMware/ Player) using a virtual bridge to the physical Ethernet of the host machine
  - Start the pool system machine – select "Kursraum Entwicklung" entry in the list (just the one below the highlighted entry, otherwise the command for the virtual environment would not be found)
  - Log-on to the pool system machine with your computer center ID (choose a session like "KDE3" or "Gnome" at this moment) or
  - Choose "KDE3" or "Gnome" after login (do not select the "Communication Systems" or any Windows course session)
  - Same access procedure like last time, run run-vmware.sh /var/lib/ vmware/vmconfigs/linux-comsys-ws08.xml

Communication Systems
The exercise environment

- **No DHCP is running this time – manual IP address assignment required for network access**
  - The Linux within the virtual machine should not have an IP address assigned
  - Try to avoid duplicate addresses (or try it out :))
  - Check for the proper netmask setting
  - Use the proper default gateway – use the commands of last practical course
  - In ping tests try to use IP addresses as the resolver might not be configured properly or the nameserver is unreachable
- **Work with your neighbors and coordinate your actions!**
- **Homework: Download wireshark, try it out at home on your machine or in your dormitory ...**

ALBERT-LUDWIGS-
UNIVERSITÄT FREIBURG

Communication Systems

ARP

University of Freiburg
Computer Science
Computer Networks and Telematics
Prof. Christian Schindelhauer

CoNe Freiburg

IIF
INSTITUT FÜR INFORMATIK FREIBURG

Communication Systems Organization and Q&A

- Any questions from last lecture?
- We were dealing with Ethernets last lecture:
- Why not only Ethernets are used for networking – working plug&play, stations in the net are found automatically!?
- Why it is impossible to use jumbo frames in a mixed Gigabit and Fast Ethernet LAN?
- What is the minimum length of an Ethernet packet, why? Is that really needed in switched Ethernets, why/not?
- Which restriction may apply if a Gigabit Ethernet adapter is plugged into the old-standard PCI?

Communication Systems
10GBit/s Ethernet

- **10 Gigabit Ethernet (10GbE or 10 GigE) standard first published in 2002 as IEEE Std 802.3ae-2002**
- **Several substandards followed to define different media types (mainly fiber, later several copper variants)**
  - To support different PHYs often pluggable PHY modules deployed in switches
- **Major difference to earlier standards:**
  - Obsoleting half duplex operation and CSMA/CD
  - Only full duplex links connected by switches
  - Introducing a special 10 Gigabit Ethernet WAN standard: 802.3aq-2006 (slightly slower with extra encapsulation)

Communication Systems
10GBit/s Ethernet

- **Optical PHYs for single or multi-mode fiber**
  - Single mode: 8.3µm – thus difficult to deploy but much longer distances (some km, single path of light travelling)
  - Multi-mode: 50 or 62.5µm – multi path with the problem of differential mode delay (up to 300m, cheaper cable, optics)
- **10 Gigabit Ethernet copper standards followed later because on quite demanding wire characteristics (IEEE 802.3an-2006 for twisted pair copper wire)**
  - Allows up to 55 m (180 ft) with existing Cat 6 cabling
  - Typical 100m with partitioned Category 6a cables with reduced crosstalk between UTP cables (alien crosstalk)
  - Wire-level modulation: Tomlinson-Harashima precoded (THP) version of pulse amplitude modulation with 16 discrete levels (PAM16 – PAM5 in 1GbE TP), encoded in a two-dimensional checkerboard pattern (DSQ128)

Communication Systems
100GBit/s Ethernet

- **Because of modulation overhead TX copper standard adds rather high latency**
- **Other copper standards for low latency like for "backplane" wiring as needed in blade computer units (1m distances) or 802.3ak (15m)**
- **10 GbE deployed mostly in infrastructure, network adaptors available but not widely used (1GbE much cheaper, other strategies like channel bonding)**
- **100 Gigabit Ethernet, first discussions in 2006, standardization begun in 2007 (including standard for 40GbE)**

- Same frame formats and sizes, full duplex only like in 10 GbE
- Aiming for 10-40km with optical single mode fiber (4 wave length), copper standards too, but for much shorter ranges (~10m)

Communication Systems
Ethernet link aggregation/channel

- **Ethernet by now predominant standard in LAN / MAN – alternatives searched for not depending on next generation standard**
- **Implementations for Ethernet line backup and seemless bandwith upgrades**
  - Often impossible to switch directly to the next higher standard – imagine 100% used 1GbE link means 10+% used 10GbE
  - Idea to use more than one link between two switches (proprietary extensions exist to span link aggregation over different switches)
  - Allows network's backbone speed to grow incrementally on demand, without having to replace the whole infrastructure

Communication Systems Ethernet link aggregation/

- **IEEE 802.3ad standard for link aggregation (LA – Ethernet, channel bonding – several names exist)**
  - Either static setup or implementation with Link Aggregation Control Protocol (LACP)
  - Some proprietary alternatives exists of companies like Cisco or Nortel (offering multi switch bonding / high availability, (split) multi link trunking - SMLT)
  - Common way to balance traffic by using Layer 3 (IPv4, IPv6 address) hashes (Which problems you would expect when using this approach?)

Communication Systems Ethernet link aggregation/

- **Mixed port speed aggregation possible (useful for backup but not load balancing)**
- **Not possible to aggregate links operating mixed in full and half duplex mode**
- **Some disadvantages**
  - Depending on sessions rarely 50/50 distribution (on two lane LA) reached, often distribution like 70/30
  - Imagine 1GbE capable station connected to a switch using 2*100Mbit/s uplink – what is the resulting max. Bandwidth?
  - More advanced switches: Using Layer 4 link hashes for

Communication Systems
Ethernet link aggregation/

- **Higher layer implementations for most common Unix OS (like Linux)**
  - Bonding over different vendors Ethernet adaptors possible
  - Several setups
    - High availability and outgoing traffic equalization (independent of switch)
    - HA and link aggregation for both directions (dependent on compatibility to switch link layer aggregation)
- **Talked of bandwidth issues; addressing higher layer**

Communication Systems
Ethernet extensions

- **Large collision domains allowed restricted number of stations**
  - Broadcast traffic – even a rather moderate load can reduce Ethernet throughput significantly up to rendering a network unuseable completely
  - Switched operation reduces domains to just one wire/direction (thus no collision handling, half duplex operation standardized in 10GBit/s Ethernet)
- **Diameter of Ethernets increased throught switching and optical transmission**
  - Whole university campus connected using Ethernet: 100/1000MBit/s to the stations, up to 10GBit/s in backbone
  - Number of connected stations quite large

Communication Systems
Ethernet on Freiburg University

- **Freiburg University campus: 156 switching centers offering ~29k Ethernet ports of 100 / 1000Mbit/s speed, ~16k ports acive**
- **Ethernet only infrastructure by now (obsoleting FDDI, TokenRing, ATM)**
- **Backbone is using Ethernet trunking ??**

ALBERT-LUDWIGS-
UNIVERSITÄT FREIBURG

Communication Systems

Christian Schindelhauer

University of Freiburg
Computer Science
Computer Networks and Telematics
Prof. Christian Schindelhauer

CoNe Freiburg

IIF
INSTITUT FÜR
INFORMATIK
FREIBURG