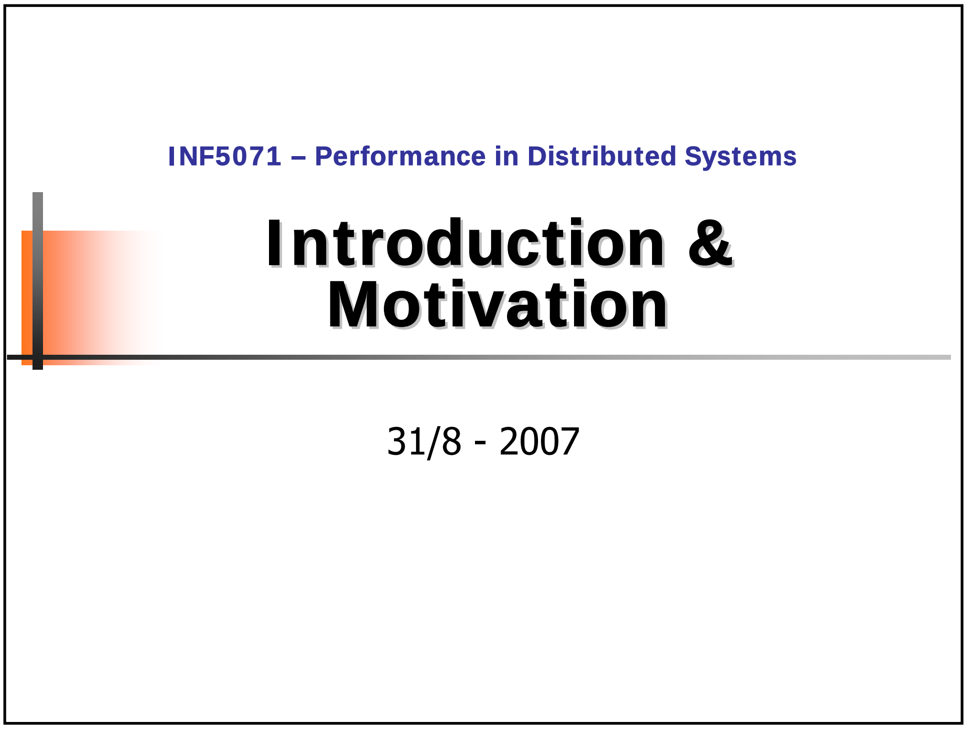
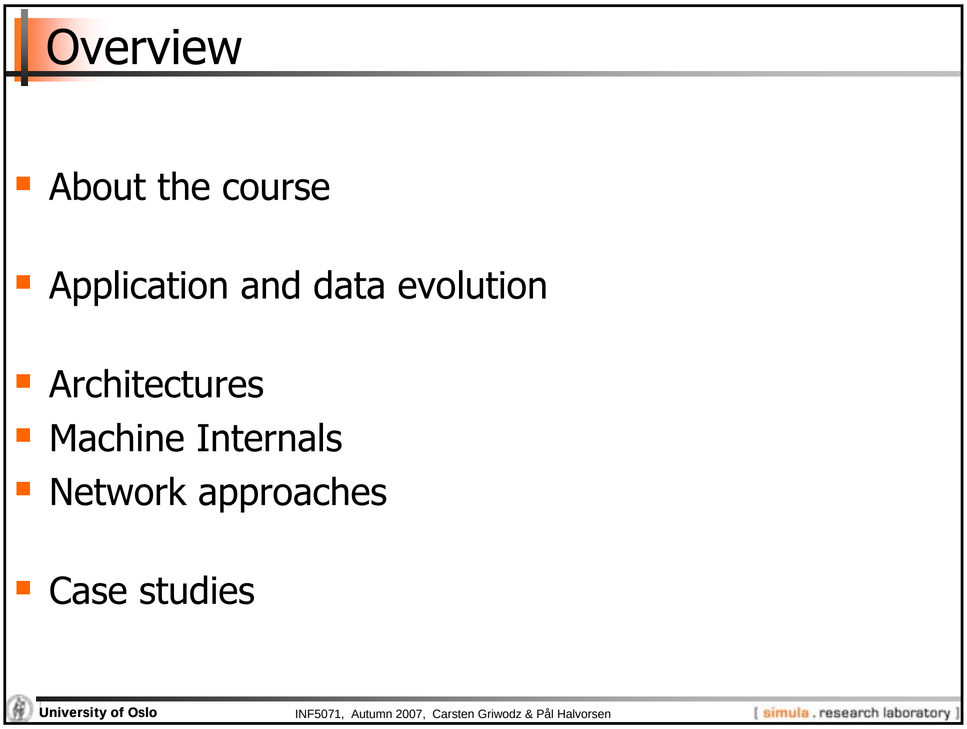


INF5071 – Performance in Distributed Systems

Introduction & Motivation

31/8 - 2007



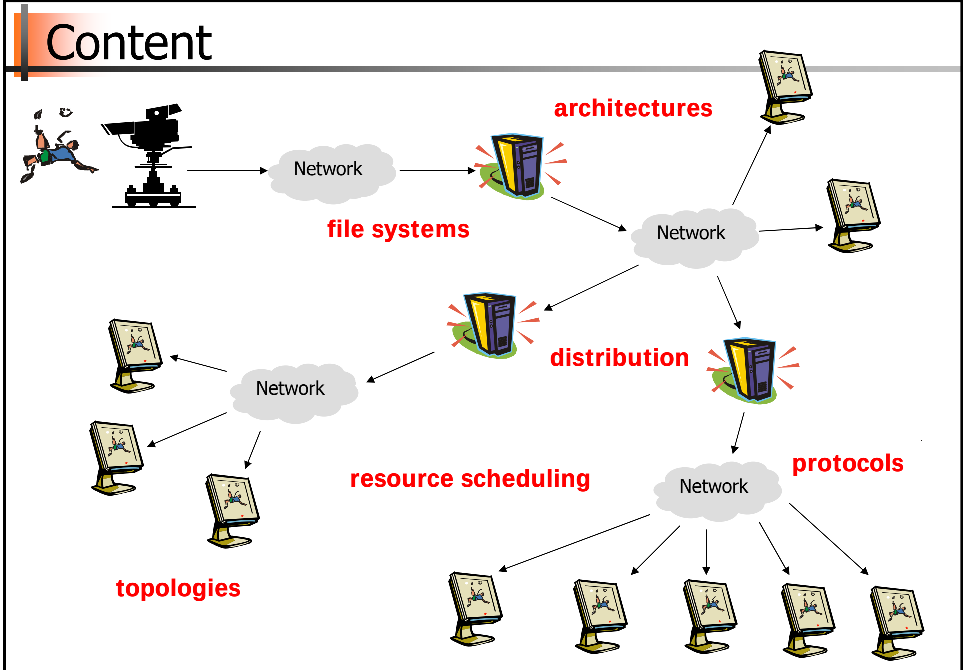
Overview

- About the course
- Application and data evolution
- Architectures
- Machine Internals
- Network approaches
- Case studies

Lecturers

- Carsten Griwodz
  - email: griff @ ifi
  - office: Simula 153
- Pål Halvorsen
  - email: paalh @ ifi
  - office: Simula 253

Content



Network

architectures

file systems

Network

distribution

Network

protocols

resource scheduling

Network

topologies

Content

- Applications and characteristics
  (components, requirements, …)

- Server examples and resource management
  (CPU and memory management)

- Storage systems
  (management of files, retrieval, …)

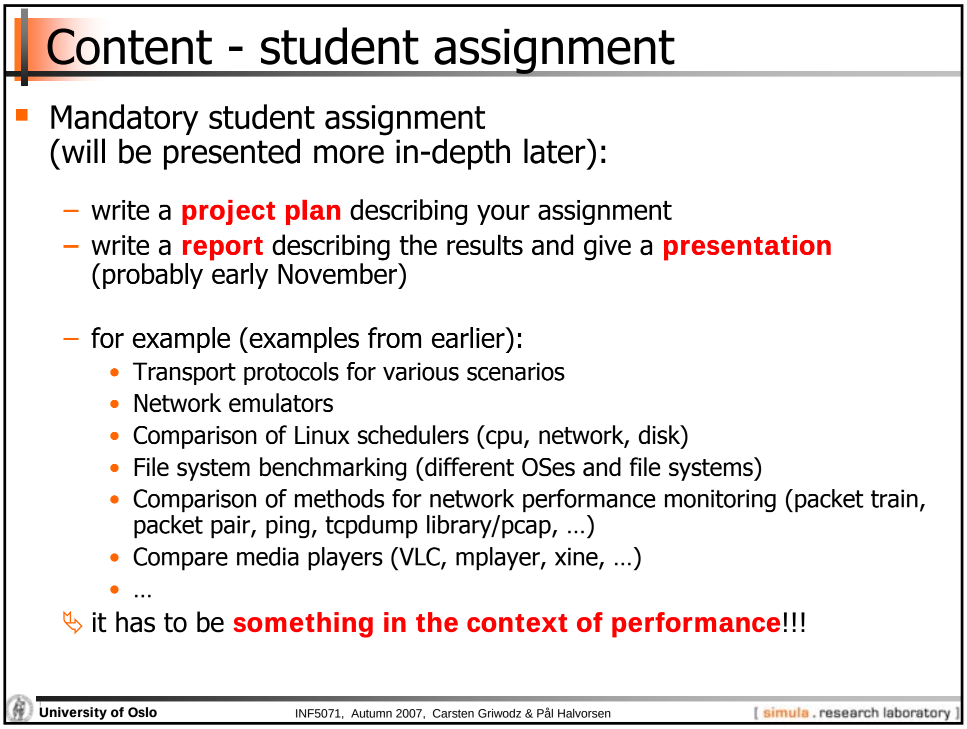
Content

- Protocols with and without Quality of Service (QoS)
  (specific and generic QoS approaches)

- Distribution
  (use of caches and proxy servers)

- Peer-to-Peer
  (various clients, different amount of resources)

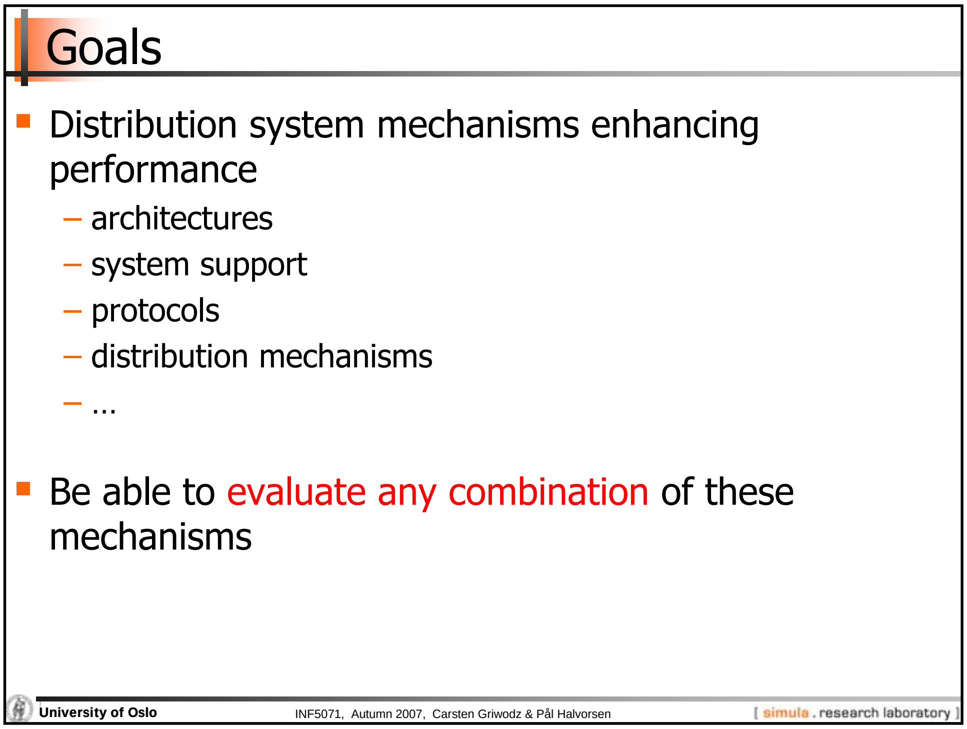
- Guest lecture: The fast searching system
  (architecture, resource utilization and performance, storage and distribution of data, parallelism, etc.)



Content - student assignment

- Mandatory student assignment
  (will be presented more in-depth later):
  - write a **project plan** describing your assignment
  - write a **report** describing the results and give a **presentation** (probably early November)
  - for example (examples from earlier):
    - Transport protocols for various scenarios
    - Network emulators
    - Comparison of Linux schedulers (cpu, network, disk)
    - File system benchmarking (different OSes and file systems)
    - Comparison of methods for network performance monitoring (packet train, packet pair, ping, tcpdump library/pcap, …)
    - Compare media players (VLC, mplayer, xine, …)
    - …

  ↳ it has to be **something in the context of performance**!!!



Goals

- Distribution system mechanisms enhancing performance
  - architectures
  - system support
  - protocols
  - distribution mechanisms
  - …
- Be able to evaluate any combination of these mechanisms

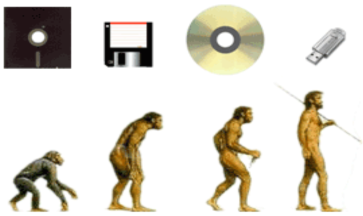
Exam

- Prerequisite:
  approved presentation of student assignment

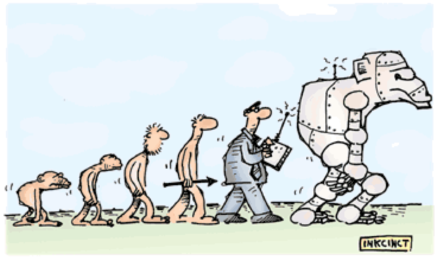
- Oral exam (**early December**):
  - all *transparencies* from lectures

    Note: we do NOT have a book, and you probably do not want to read all the articles the slides are made from!

  - content of your *own student assignment*



Evolution

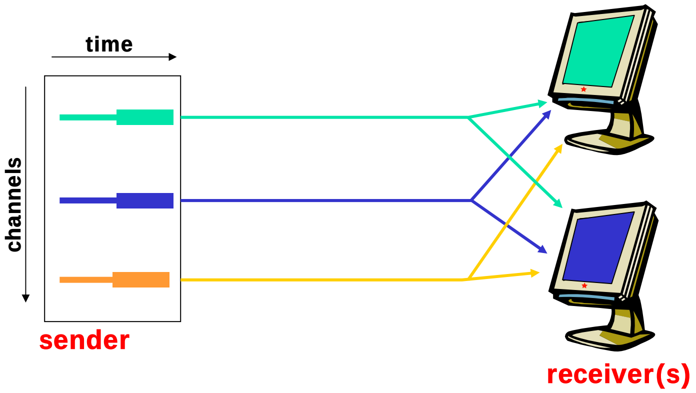


Discrete Data to Continuous Media Data

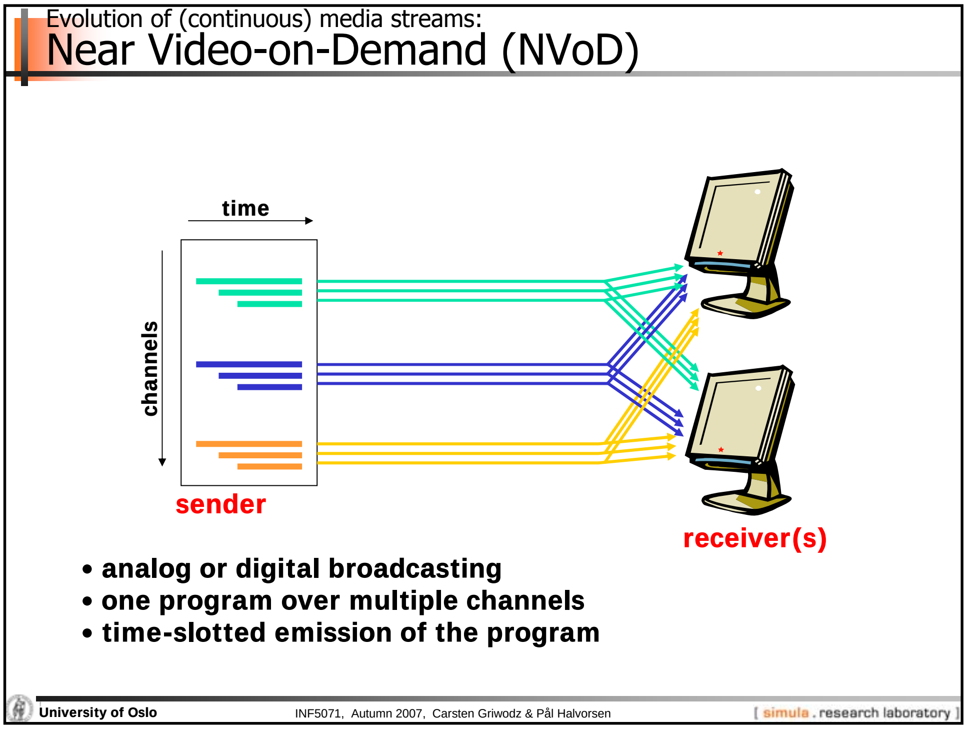
3D streaming is coming ...

Evolution of (continuous) media streams:

Television (Broadcast)



- **analog or digital**
- **traditionally, one program per channel**
  - ❑ analog use frequency division multiplexing only
  - ❑ digital may additionally use time division multiplexing inside one frequency (several programs per channel)



Evolution of (continuous) media streams:

Near Video-on-Demand (NVoD)

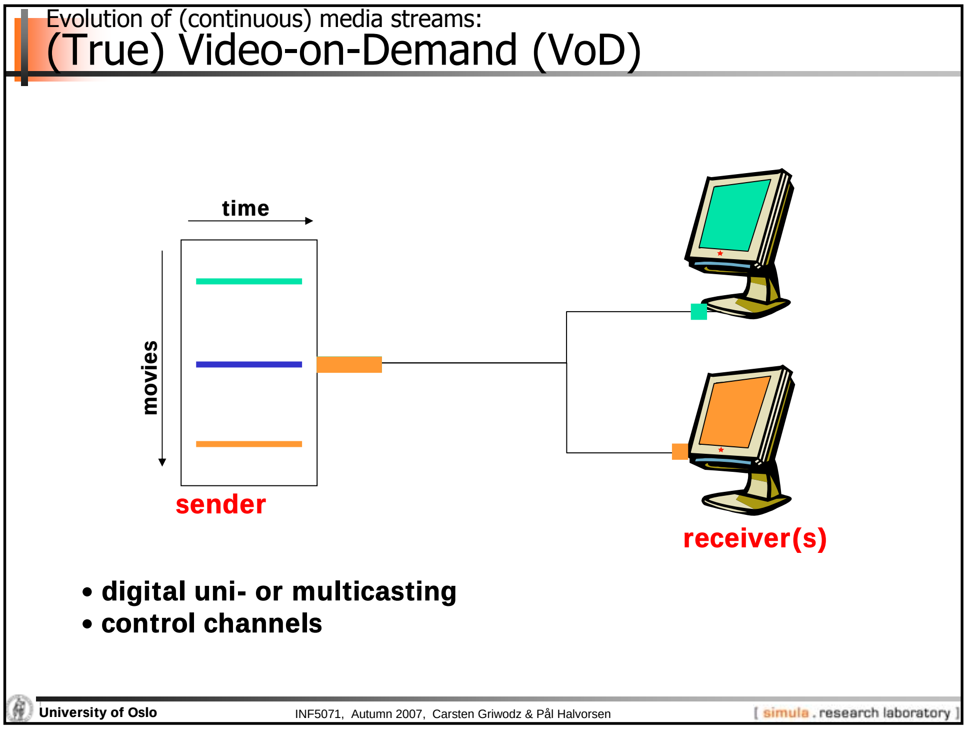
time

channels

sender

receiver(s)

- analog or digital broadcasting
- one program over multiple channels
- time-slotted emission of the program



Evolution of (continuous) media streams:

(True) Video-on-Demand (VoD)

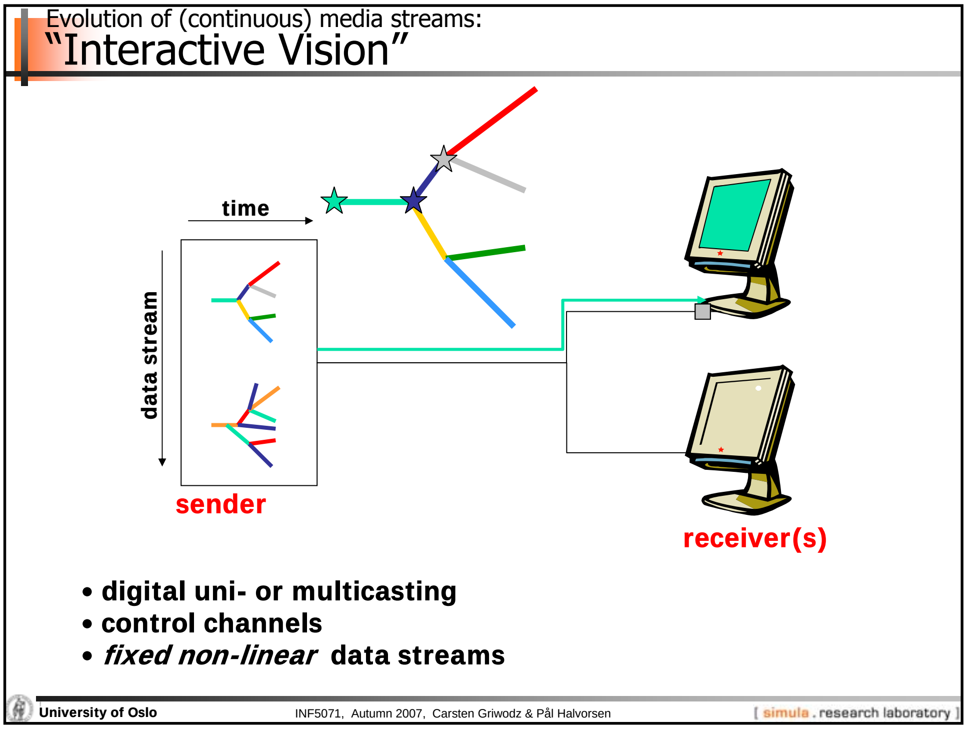
time

movies

sender

receiver(s)

- digital uni- or multicasting
- control channels



Evolution of (continuous) media streams:

“Interactive Vision”

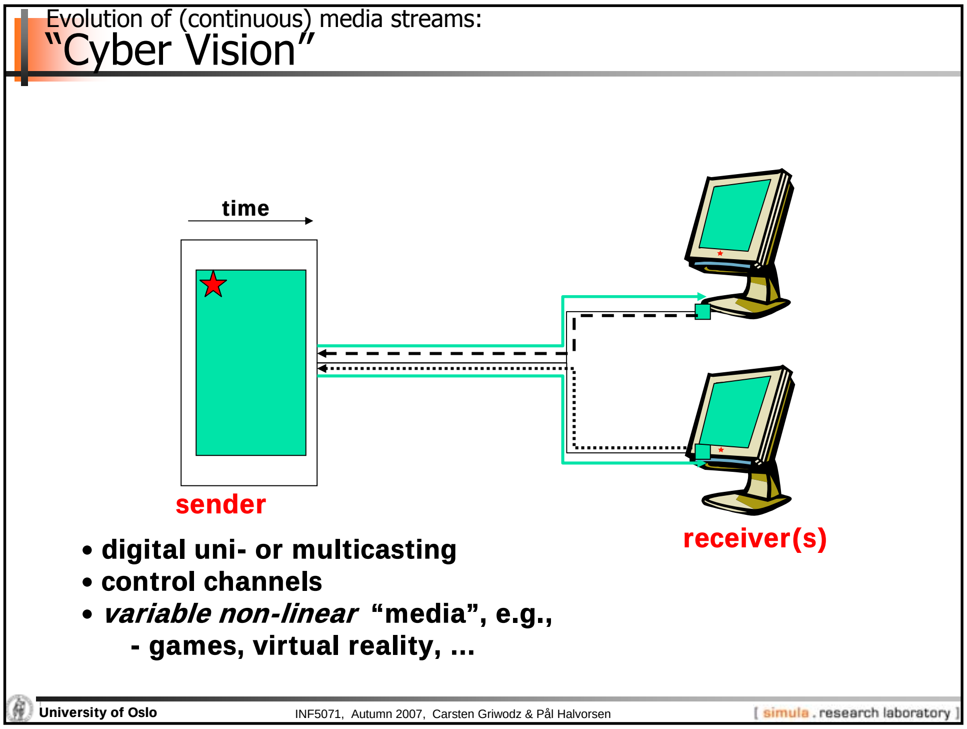
time

data stream

sender

receiver(s)

- digital uni- or multicasting
- control channels
- *fixed non-linear* data streams



Evolution of (continuous) media streams:

“Cyber Vision”

time

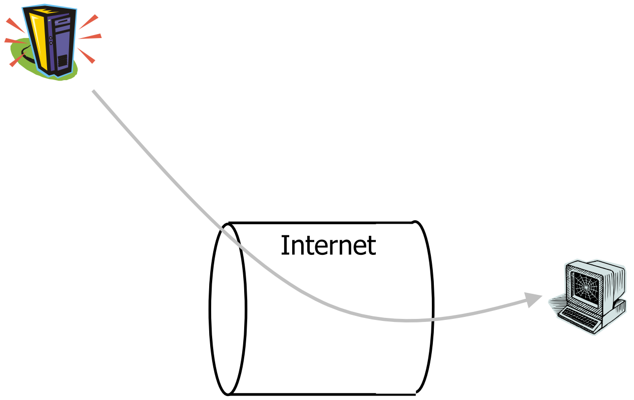
sender

receiver(s)

- digital uni- or multicasting
- control channels
- *variable non-linear* “media”, e.g.,
  - games, virtual reality, ...

Evolution & Requirements:

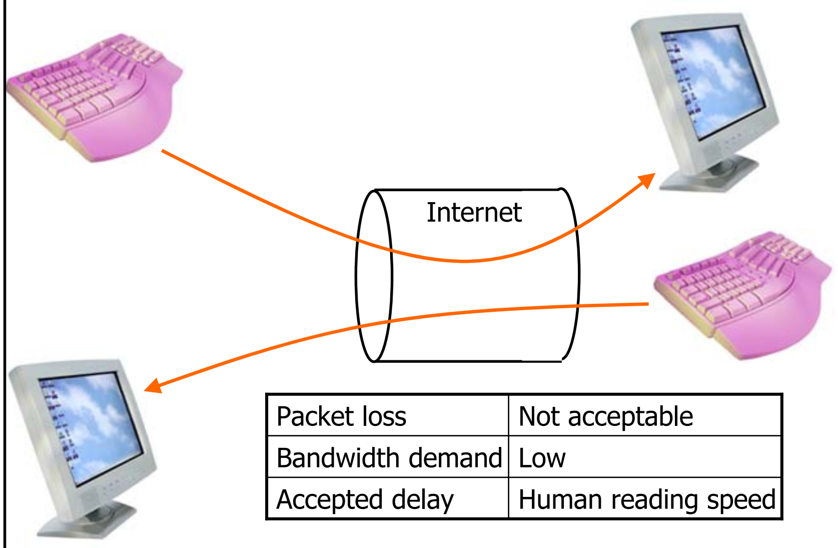
File download and Web browsing



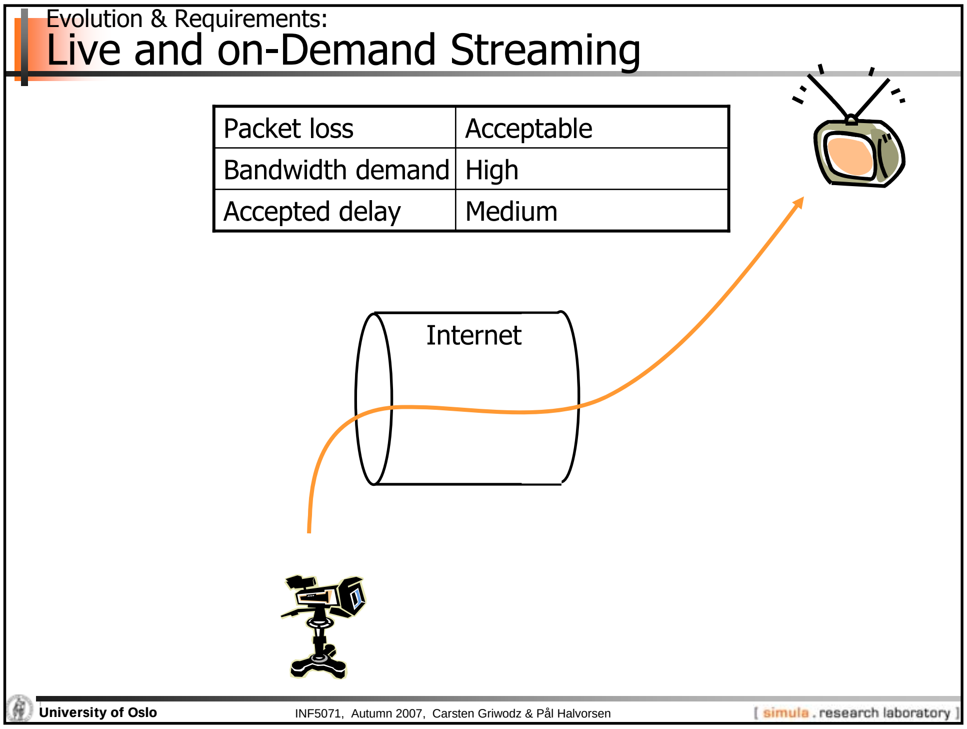
Packet loss	Not acceptable
Bandwidth demand	Low (?)
Accepted delay	Medium – High (?)

Evolution & Requirements:

Textual commands and textual chat



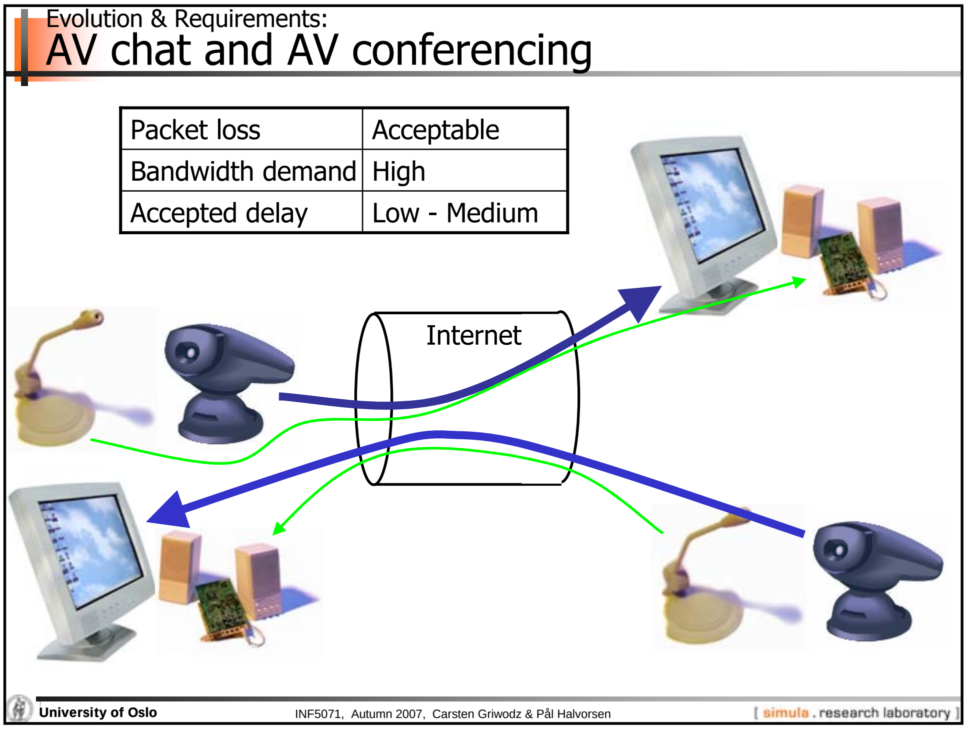
Packet loss	Not acceptable
Bandwidth demand	Low
Accepted delay	Human reading speed



Evolution & Requirements:

Live and on-Demand Streaming

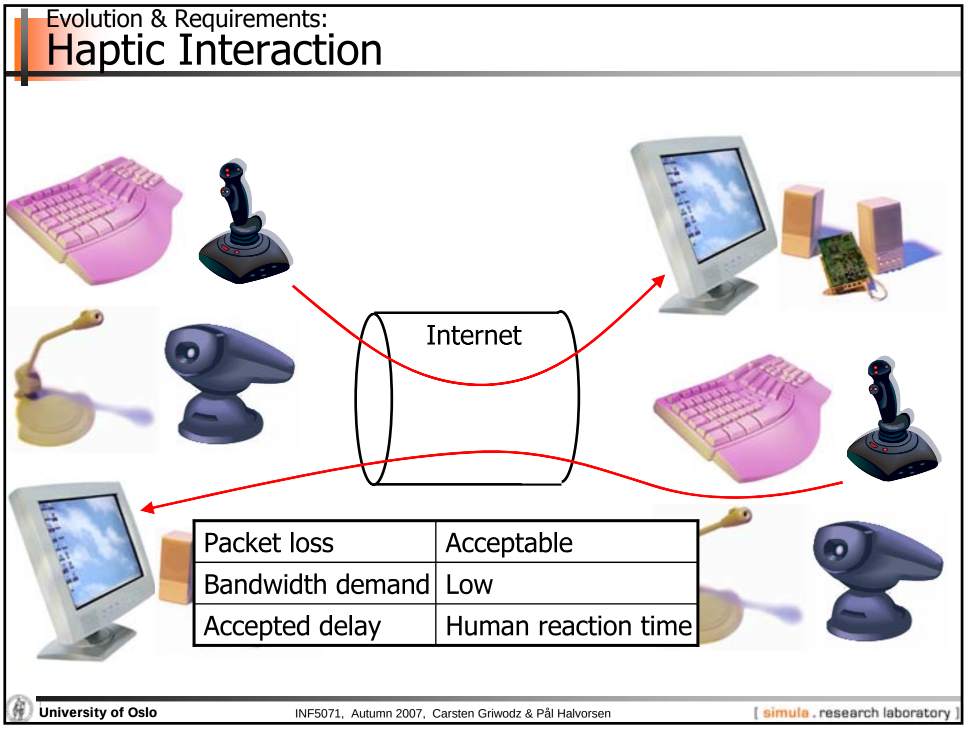
Packet loss	Acceptable
Bandwidth demand	High
Accepted delay	Medium



Evolution & Requirements:

AV chat and AV conferencing

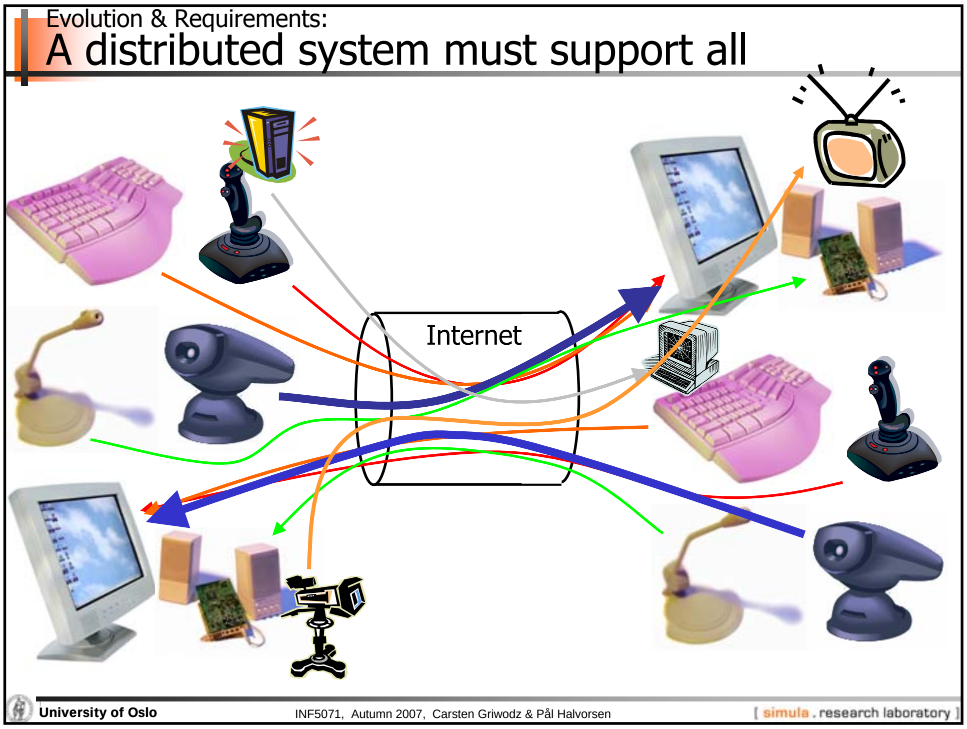
Packet loss	Acceptable
Bandwidth demand	High
Accepted delay	Low - Medium



Evolution & Requirements:

Haptic Interaction

Packet loss	Acceptable
Bandwidth demand	Low
Accepted delay	Human reaction time



Evolution & Requirements:

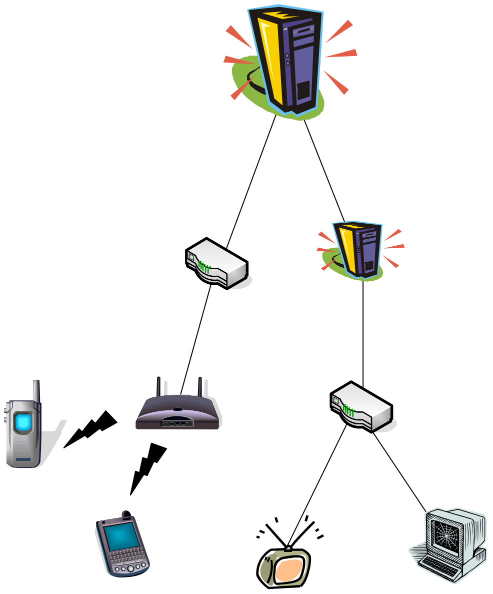
A distributed system must support all

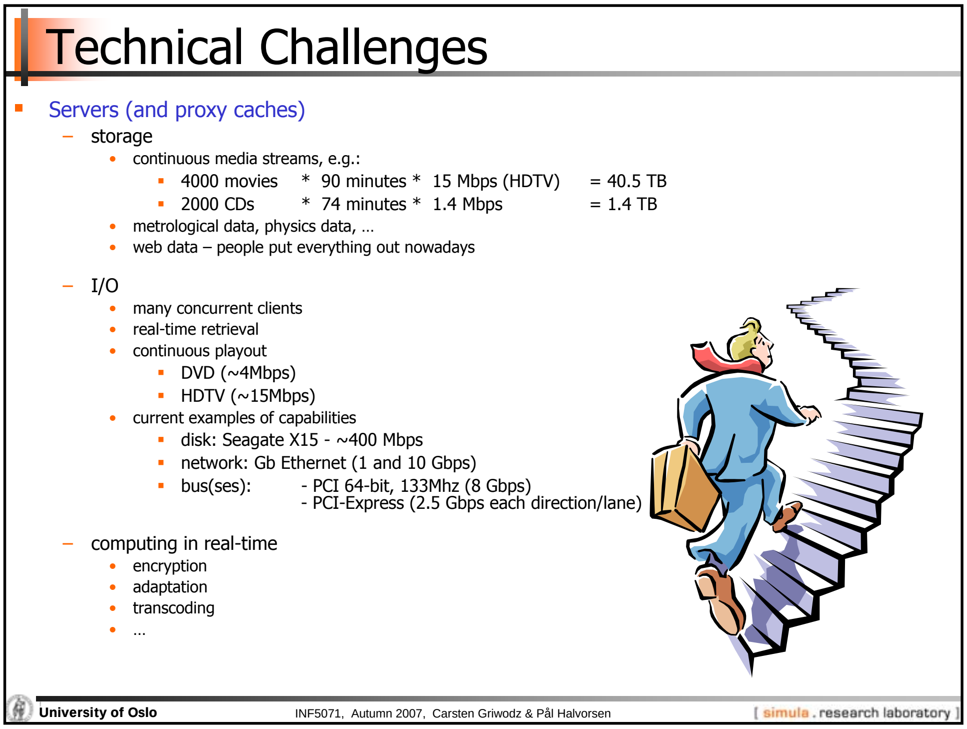
Different Views on Requirements

- Application / user
  - QoS – time sensitivity?
  - resource capabilities – bandwidth, latency, loss, reliability, …
  - best possible perception
- Business
  - scalability
  - reliability
- Architectural
  - topology
  - cost vs. performance

Components

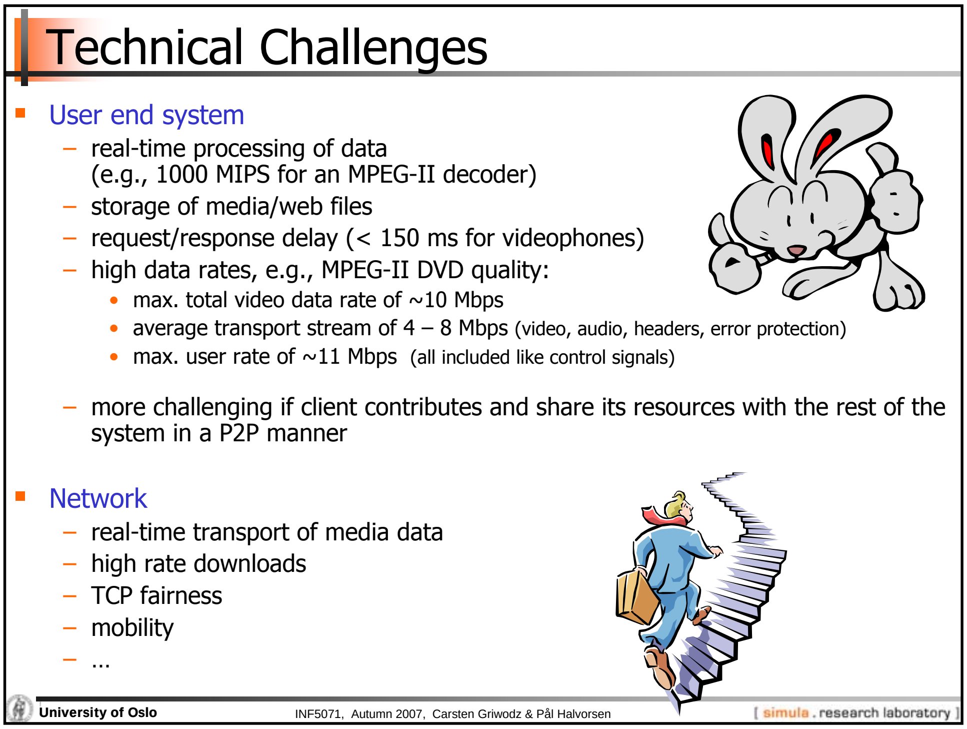
- Servers
- End-systems
  - PCs
  - TV sets with set-top boxes
  - PDAs
  - Phones
  - …
- Intermediate nodes
  - routers
  - proxy cache servers
- Networks
  - backbone
  - local networks





Technical Challenges

- Servers (and proxy caches)
  - storage
    - continuous media streams, e.g.:
      - 4000 movies * 90 minutes * 15 Mbps (HDTV) = 40.5 TB
      - 2000 CDs * 74 minutes * 1.4 Mbps = 1.4 TB
    - metrological data, physics data, …
    - web data – people put everything out nowadays
  - I/O
    - many concurrent clients
    - real-time retrieval
    - continuous playout
      - DVD (~4Mbps)
      - HDTV (~15Mbps)
    - current examples of capabilities
      - disk: Seagate X15 - ~400 Mbps
      - network: Gb Ethernet (1 and 10 Gbps)
      - bus(ses): - PCI 64-bit, 133Mhz (8 Gbps)
        - PCI-Express (2.5 Gbps each direction/lane)
  - computing in real-time
    - encryption
    - adaptation
    - transcoding
    - …



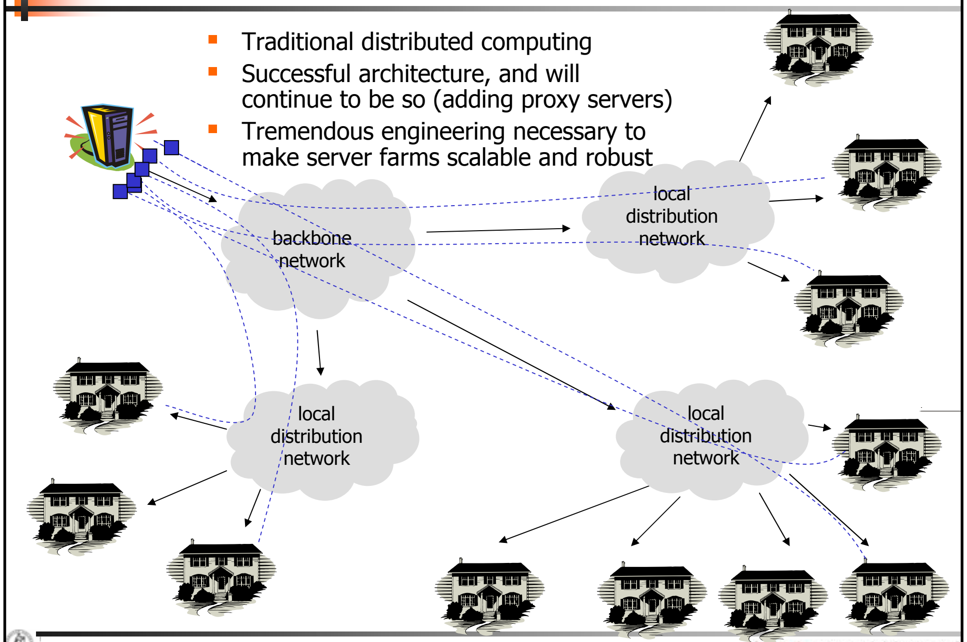
Technical Challenges

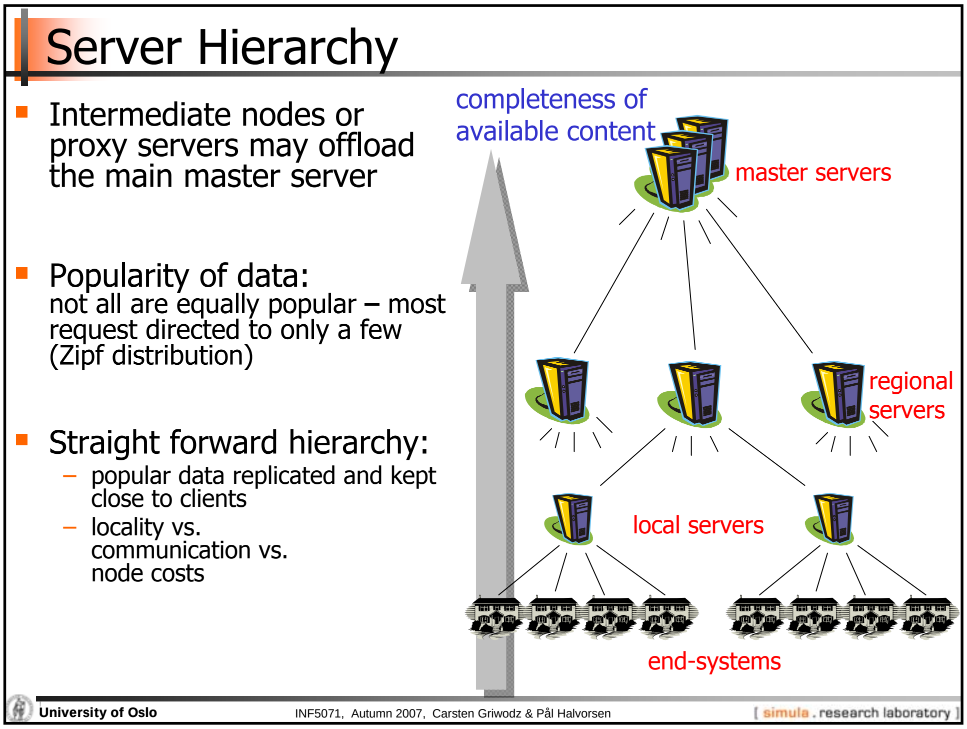
- User end system
  - real-time processing of data
    (e.g., 1000 MIPS for an MPEG-II decoder)
  - storage of media/web files
  - request/response delay (< 150 ms for videophones)
  - high data rates, e.g., MPEG-II DVD quality:
    - max. total video data rate of ~10 Mbps
    - average transport stream of 4 – 8 Mbps (video, audio, headers, error protection)
    - max. user rate of ~11 Mbps (all included like control signals)
  - more challenging if client contributes and share its resources with the rest of the system in a P2P manner
- Network
  - real-time transport of media data
  - high rate downloads
  - TCP fairness
  - mobility
  - …

Traditional Distributed Architectures

Client-Server

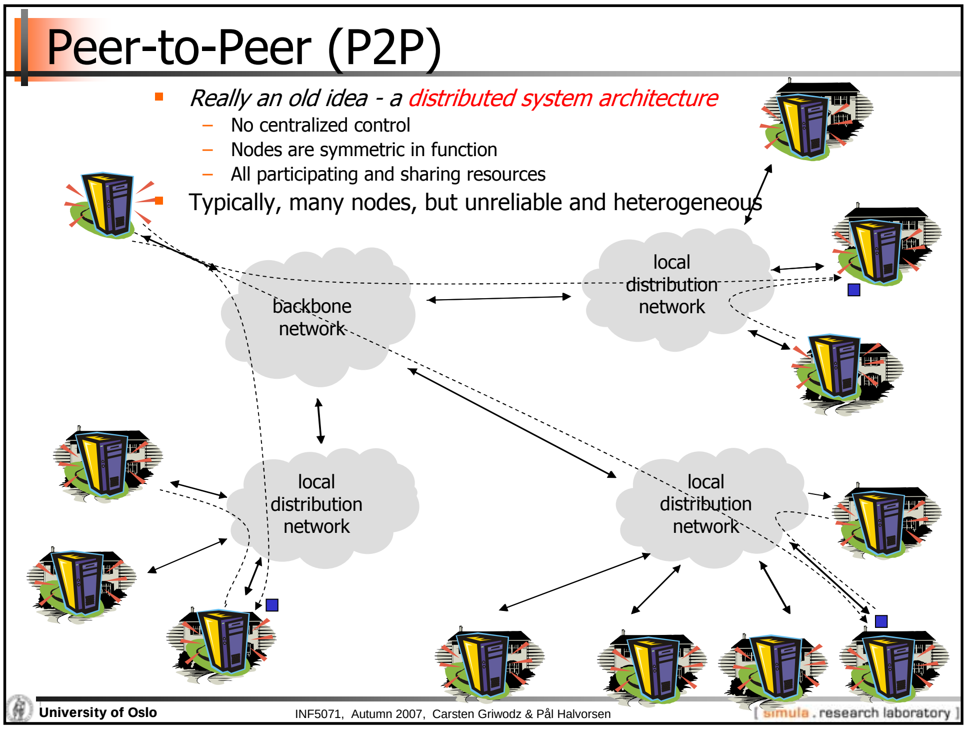
- Traditional distributed computing
- Successful architecture, and will continue to be so (adding proxy servers)
- Tremendous engineering necessary to make server farms scalable and robust


backbone network
local distribution network
local distribution network
local distribution network



Server Hierarchy

- Intermediate nodes or proxy servers may offload the main master server
- Popularity of data:
  not all are equally popular – most request directed to only a few (Zipf distribution)
- Straight forward hierarchy:
  - popular data replicated and kept close to clients
  - locality vs. communication vs. node costs



Peer-to-Peer (P2P)

- *Really an old idea - a distributed system architecture*
  - No centralized control
  - Nodes are symmetric in function
  - All participating and sharing resources
- Typically, many nodes, but unreliable and heterogeneous

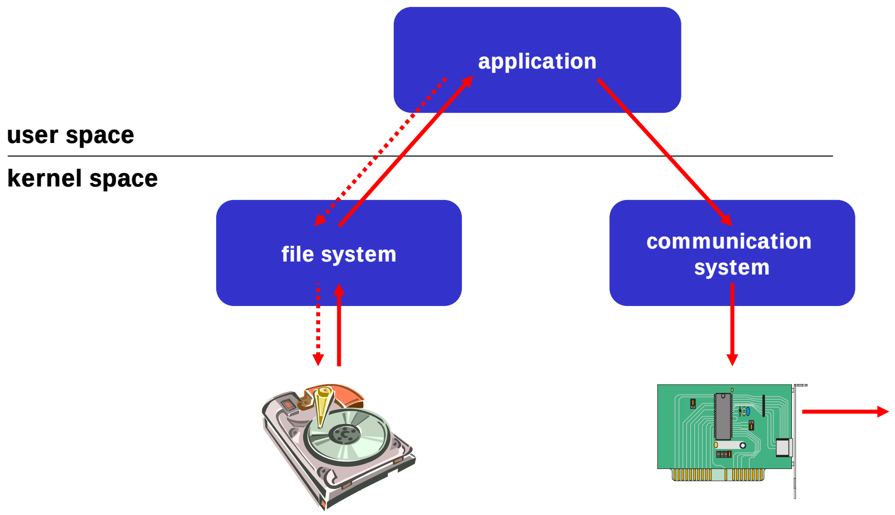
backbone network
local distribution network
local distribution network
local distribution network

Topologies

- Client / server
  - easy to build and maintain
  - severe scalability problems
- Hierarchical
  - complex
  - potential good performance and scalability
  - consistency challenge
  - cost vs. performance tradeoff
- P2P
  - complex
  - low-cost (for content provider!!)
  - heterogeneous and unreliable nodes
- We will in later lectures look at different issues for all these

Traditional Server Machine Internals

General OS Structure and Retrieval Data Path


application

user space

kernel space

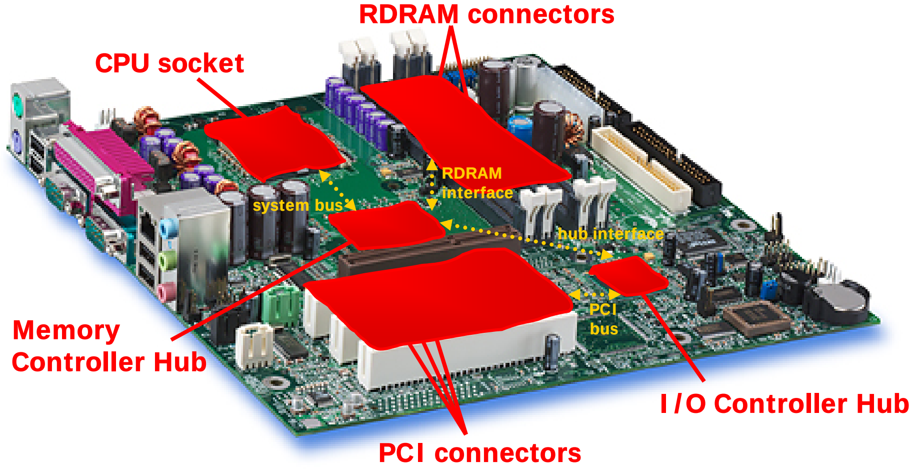
file system

communication system

Example:

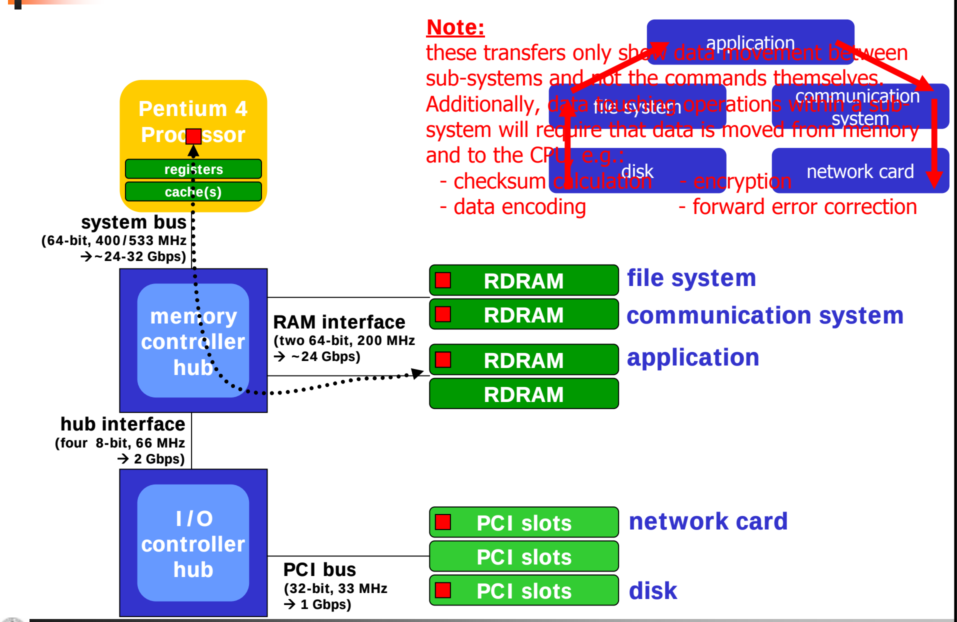
Intel Hub Architecture (850 Chipset) – I

Intel D850MD Motherboard:



Example:

Intel Hub Architecture (850 Chipset) – II

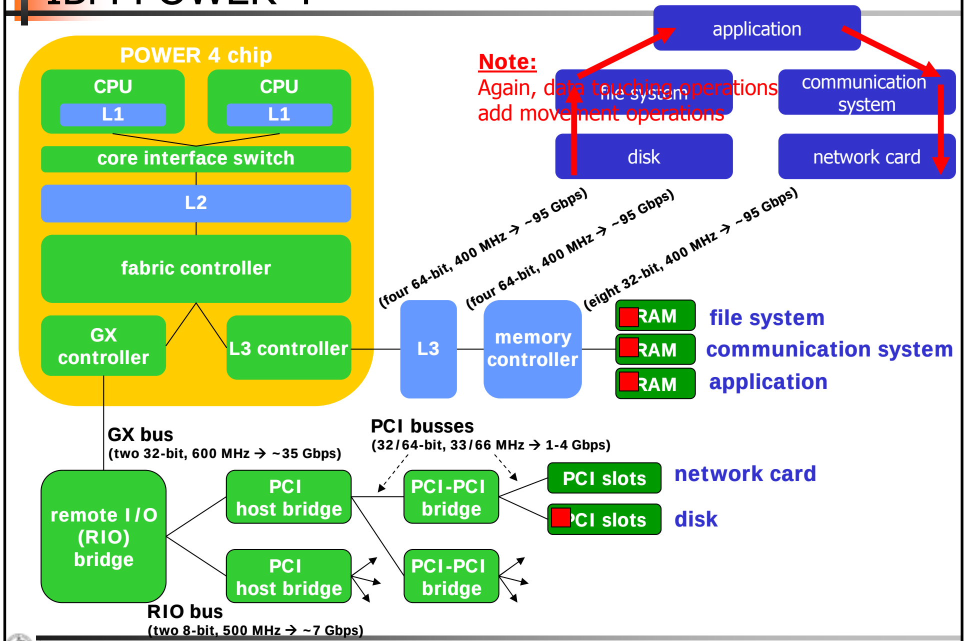


Note:
these transfers only show data movement between sub-systems and not the commands themselves.
Additionally, data touching operations within a sub-system will require that data is moved from memory and to the CPU, e.g.:

- checksum calculation
- data encoding
- encryption
- forward error correction

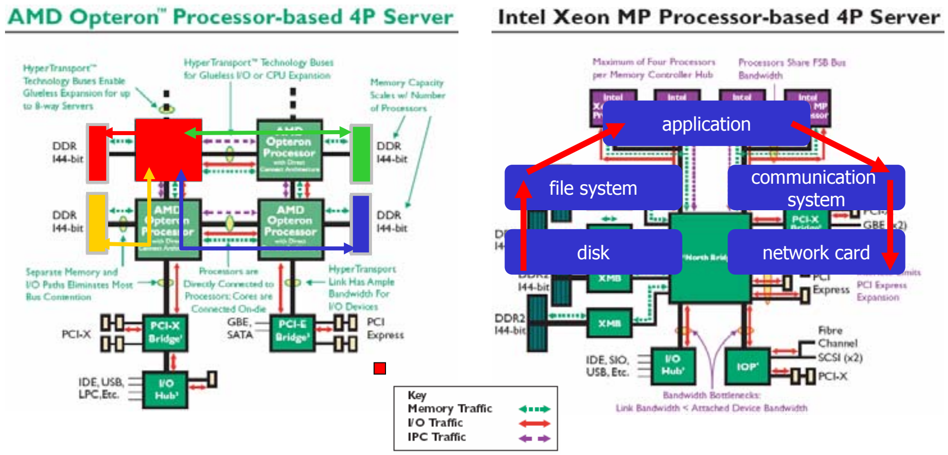
Example:

IBM POWER 4



Note:
Again, data touching operations add movement operations

Example:
AMD Opteron & Intel Xeon MP 4P servers

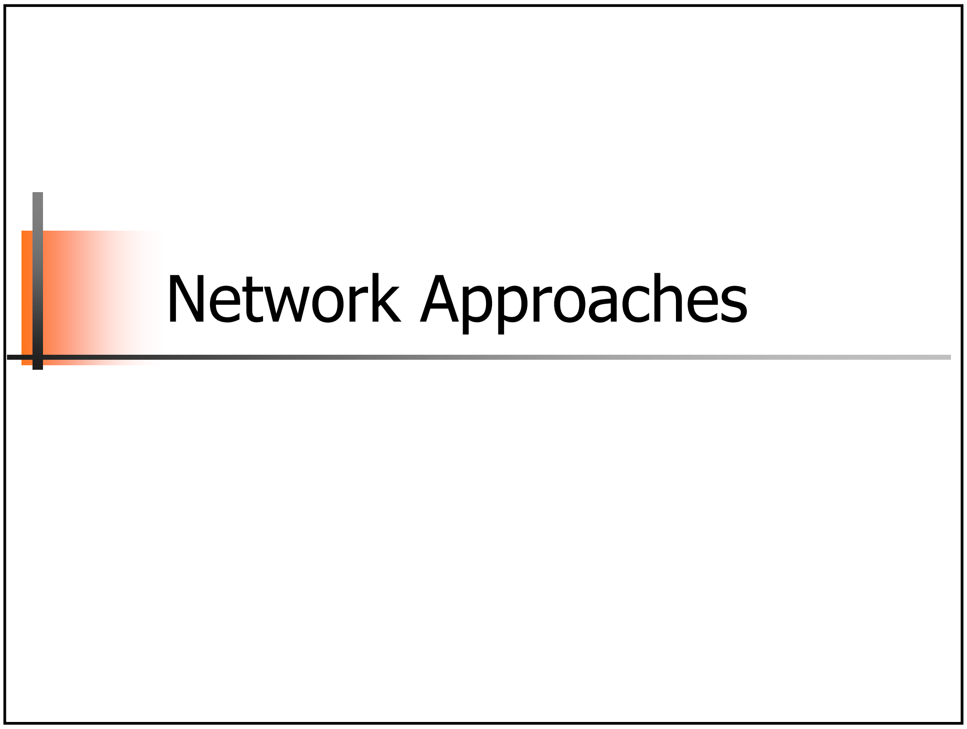

AMD Opteron™ Processor-based 4P Server
HyperTransport™ Technology Buses Enable Glueless Expansion for up to 8-way Servers
HyperTransport™ Technology Buses for Glueless I/O or CPU Expansion
Memory Capacity Scales w/ Number of Processors
DDR 144-bit
DDR 144-bit
DDR 144-bit
DDR 144-bit
AMD Opteron Processor
AMD Opteron Processor
AMD Opteron Processor
Separate Memory and I/O Paths Eliminates Most Bus Contention
Processors are Directly Connected to Processors; Cores are Connected On-die
HyperTransport Link Has Ample Bandwidth For I/O Devices
PCI-X
PCI-X Bridge
GBE, SATA
PCI-E Bridge
PCI Express
IDE, USB, LPC, Etc.
I/O Hub
Key
Memory Traffic
I/O Traffic
IPC Traffic
Intel Xeon MP Processor-based 4P Server
Maximum of Four Processors per Memory Controller Hub
Processors Share FSB Bus Bandwidth
Intel
Intel
Intel
Intel MP
application
file system
communication system
disk
network card
PCI-X Bridge
GBE (x2)
North Bridge
DDR2 144-bit
DDR2 144-bit
XMB
XMB
PCI Express
PCI Express Expansion
Fibre Channel
SCSI (x2)
PCI-X
IDE, SIO, USB, Etc.
I/O Hub
IOP
Bandwidth Bottlenecks: Link Bandwidth < Attached Device Bandwidth

☞ Know your hardware –
different configuration may have different bottlenecks

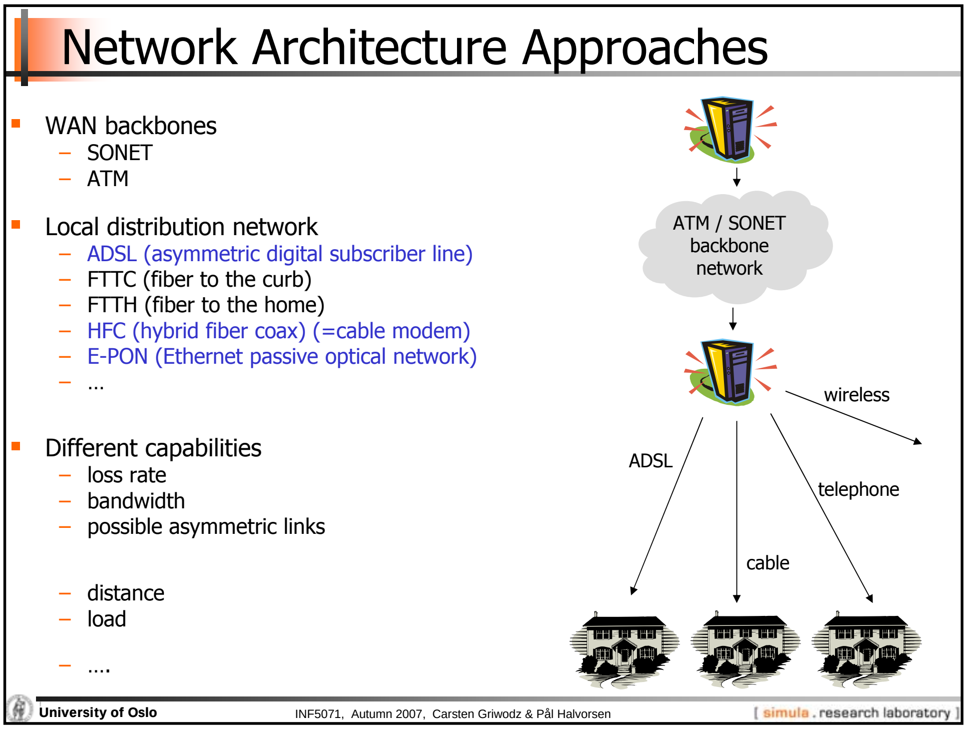
Server Internals

- Data retrieval from disk and push to network
  - buffer requirements
  - bus transfers
  - CPU usage

  - concurrent users can be merged?
  - storage (disk) system:
    - scheduling – ensure that data is available in time
    - block placement – contiguous, interleaving, striping
  - …
- Stable operations:
  - redundant HW
  - multiple nodes
- Much more, e.g., caching/prefetching, admission control, …
- We will in later lectures look at several of these

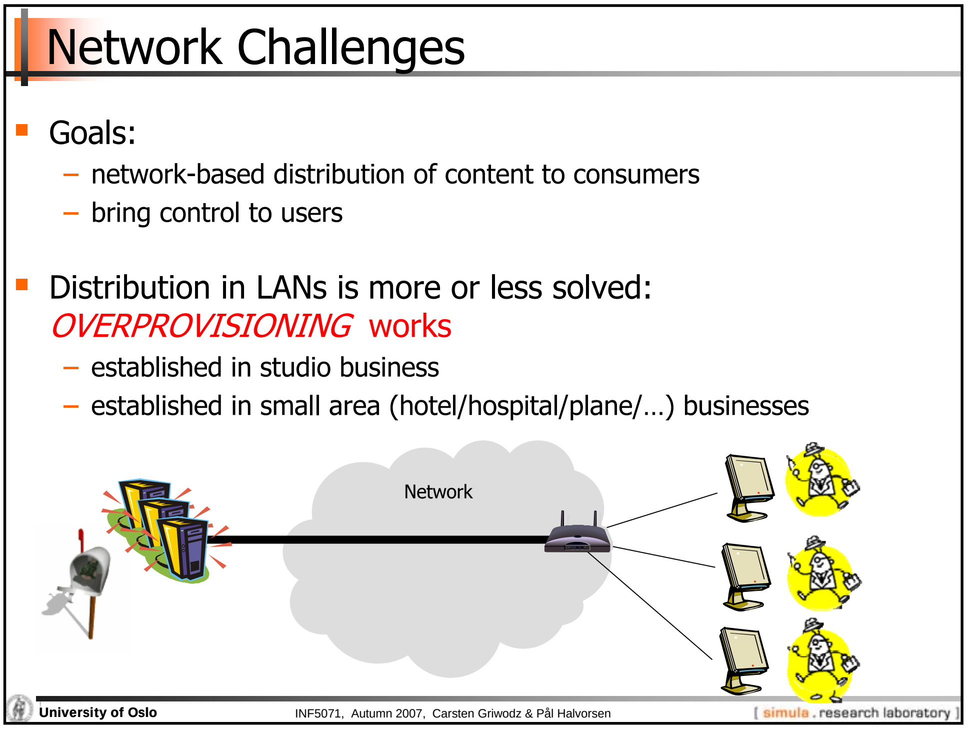


Network Approaches



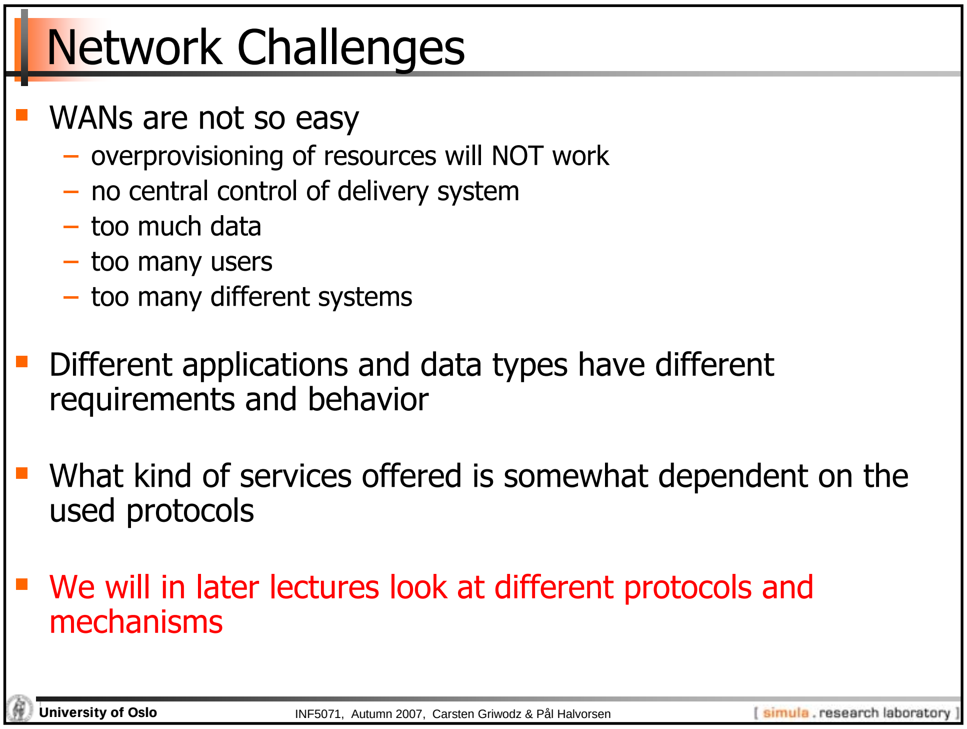
Network Architecture Approaches

- WAN backbones
  - SONET
  - ATM
- Local distribution network
  - ADSL (asymmetric digital subscriber line)
  - FTTC (fiber to the curb)
  - FTTH (fiber to the home)
  - HFC (hybrid fiber coax) (=cable modem)
  - E-PON (Ethernet passive optical network)
  - …
- Different capabilities
  - loss rate
  - bandwidth
  - possible asymmetric links
  - distance
  - load
  - ….



Network Challenges

- Goals:
  - network-based distribution of content to consumers
  - bring control to users
- Distribution in LANs is more or less solved: *OVERPROVISIONING* works
  - established in studio business
  - established in small area (hotel/hospital/plane/…) businesses

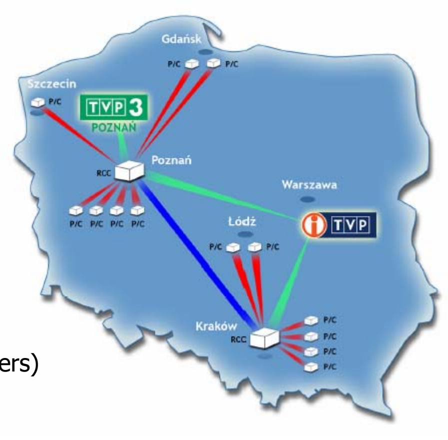


Network Challenges

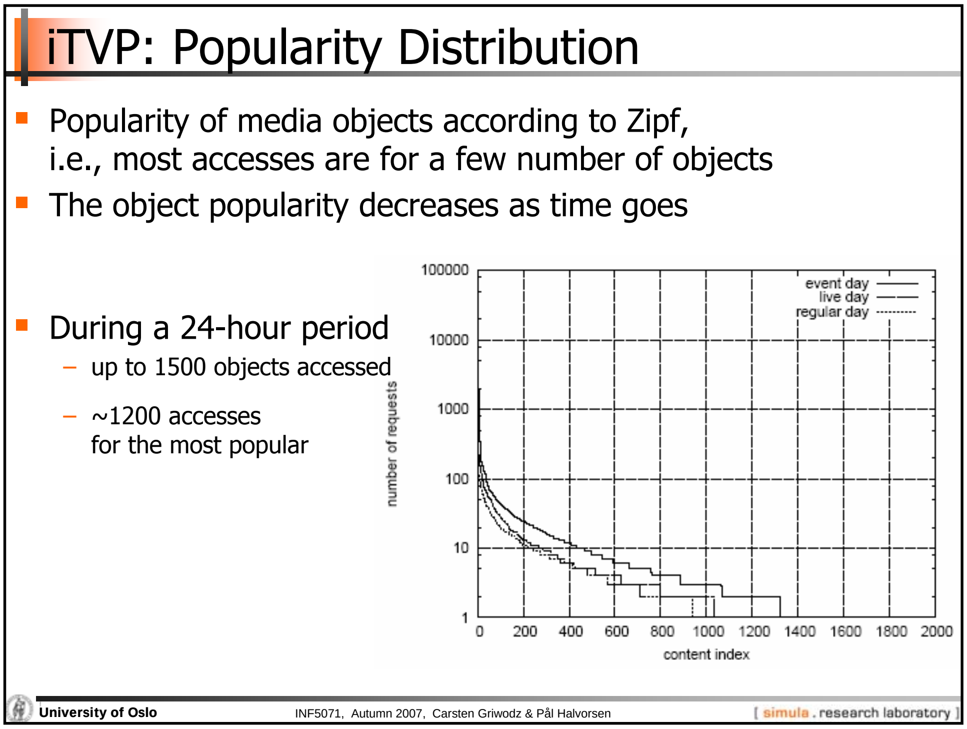
- WANs are not so easy
  - overprovisioning of resources will NOT work
  - no central control of delivery system
  - too much data
  - too many users
  - too many different systems
- Different applications and data types have different requirements and behavior
- What kind of services offered is somewhat dependent on the used protocols
- We will in later lectures look at different protocols and mechanisms

Case Studies: Application Characteristics

iTVP



- Country-wide **IP TV and VoD** in Poland
  - live & VoD
  - hierarchical structure with caching
    - regional content centers (receiving data from content providers)
    - a number of proxy caches below (handling requests from users)
  - different quality levels of the video – up to 700 Kbps
  - observations over several months

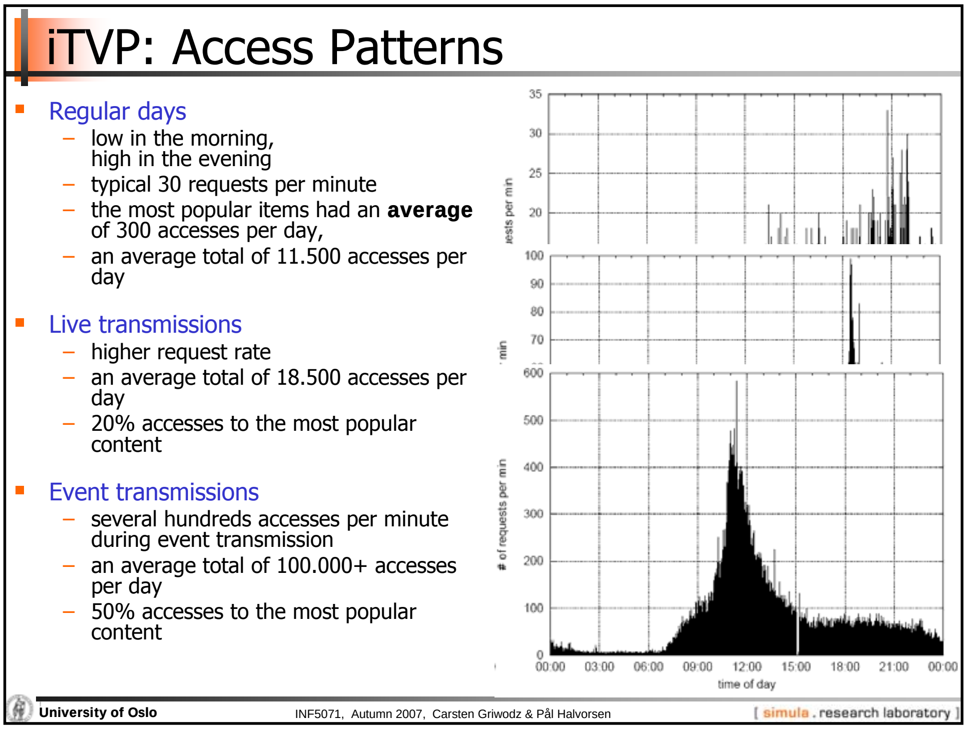


iTVP: Popularity Distribution

- Popularity of media objects according to Zipf, i.e., most accesses are for a few number of objects
- The object popularity decreases as time goes

- During a 24-hour period
  - up to 1500 objects accessed
  - ~1200 accesses for the most popular

100000
10000
1000
100
10
1
number of requests
0
200
400
600
800
1000
1200
1400
1600
1800
2000
content index
event day
live day
regular day

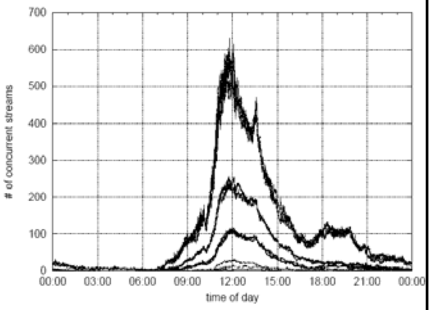


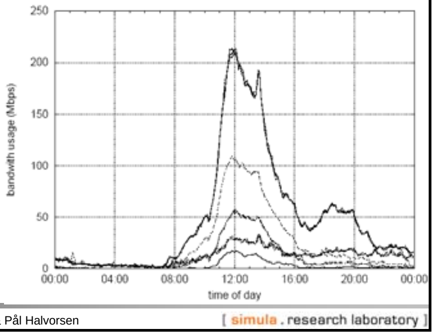
iTVP: Access Patterns

- Regular days
  - low in the morning, high in the evening
  - typical 30 requests per minute
  - the most popular items had an **average** of 300 accesses per day,
  - an average total of 11.500 accesses per day
- Live transmissions
  - higher request rate
  - an average total of 18.500 accesses per day
  - 20% accesses to the most popular content
- Event transmissions
  - several hundreds accesses per minute during event transmission
  - an average total of 100.000+ accesses per day
  - 50% accesses to the most popular content

iTVP: Concurrency and Bandwidth

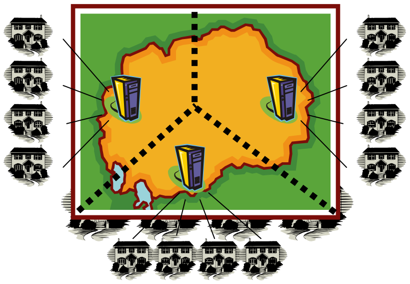
- The number of concurrent users vary, e.g., for a single proxy cache
  - event: up to 600
  - regular: usually less than 20
- Transfers between nodes are on the order of several Mbps, e.g.,
  - event:
    - single proxy: up to 200 Mbps
    - whole system: up to 1.8 Gbps
  - regular:
    - single proxy: around 60 Mbps
    - whole system: up to 400 Mbps





Funcom’s Anarchy Online

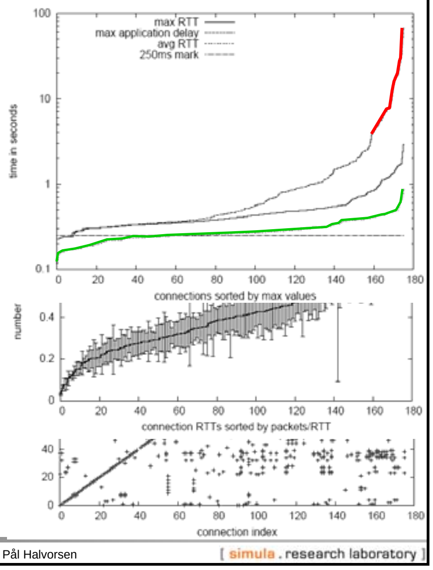
- World-wide **massive multiplayer online roleplaying game**
  - client-server
    - point-to-point TCP connections



- virtual world divided into many regions
- one or more regions are managed by one machine

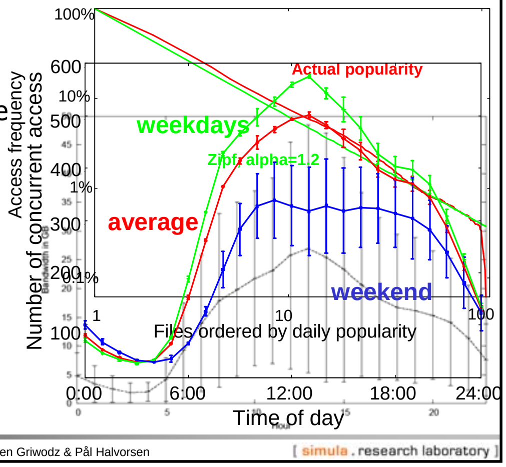
Funcom's Anarchy Online

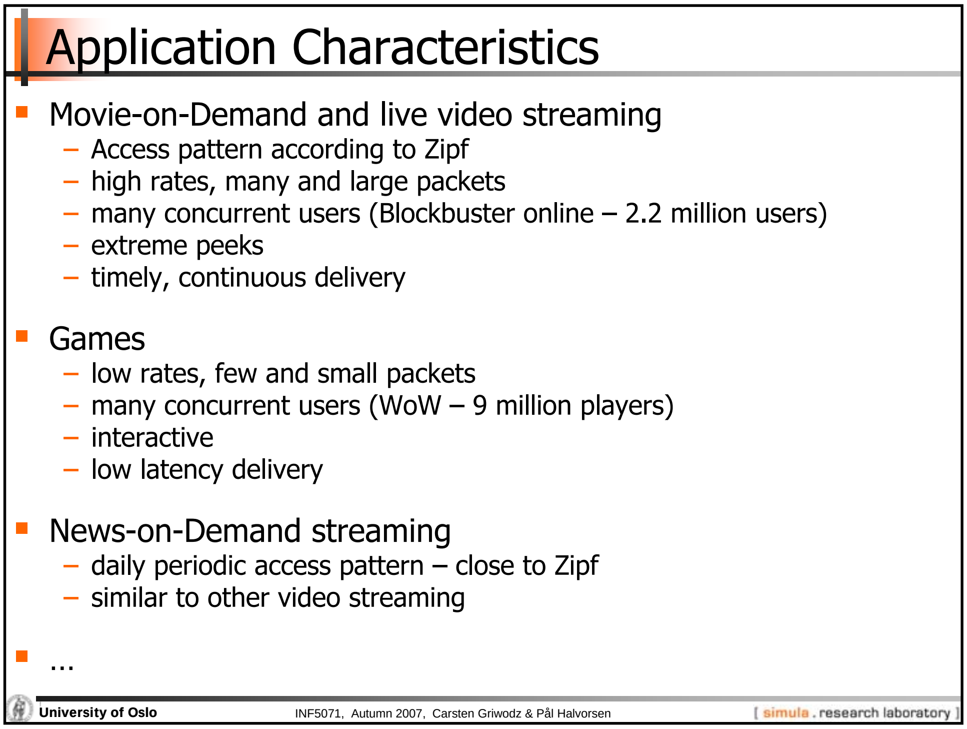
- For a given region in an one hour trace we found
  - ~175 players
  - average **layer 3** RTT somewhat above 250 ms
    ↳ OK
  - a worst-case **application** delay of 67 s (!)
    ↳ loss results in a players nightmare
  - less than 4 packets per second
  - small packets: ~120 B
    ↳ thins streams
  - Sharing/competing for both server and network resources



Verdens Gang (VG) TV: News-on-Demand

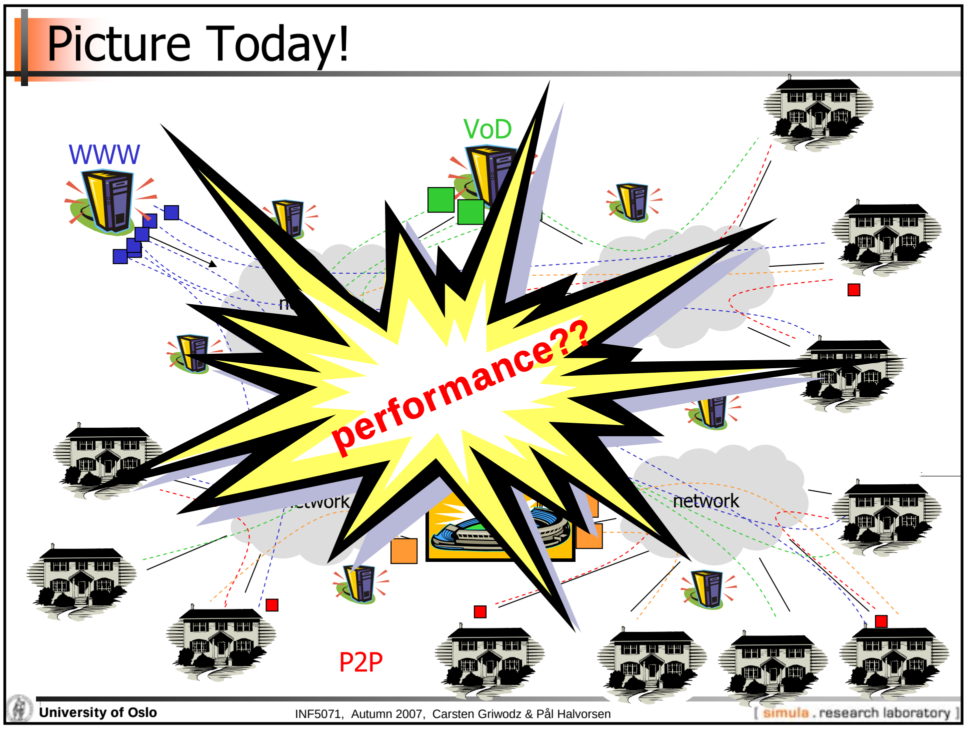
- Client-server
- Microsoft Media Server protocol (over UDP, TCP or HTTP)
- From a 2-year log of client accesses for news videos Johnsen et. al. found
  - Approximated Zipf distributed popularity, but more articles are popular
  - Access pattern dependent on time of day and day of week
  - Large bandwidth requirements, i.e., several GBs per hour



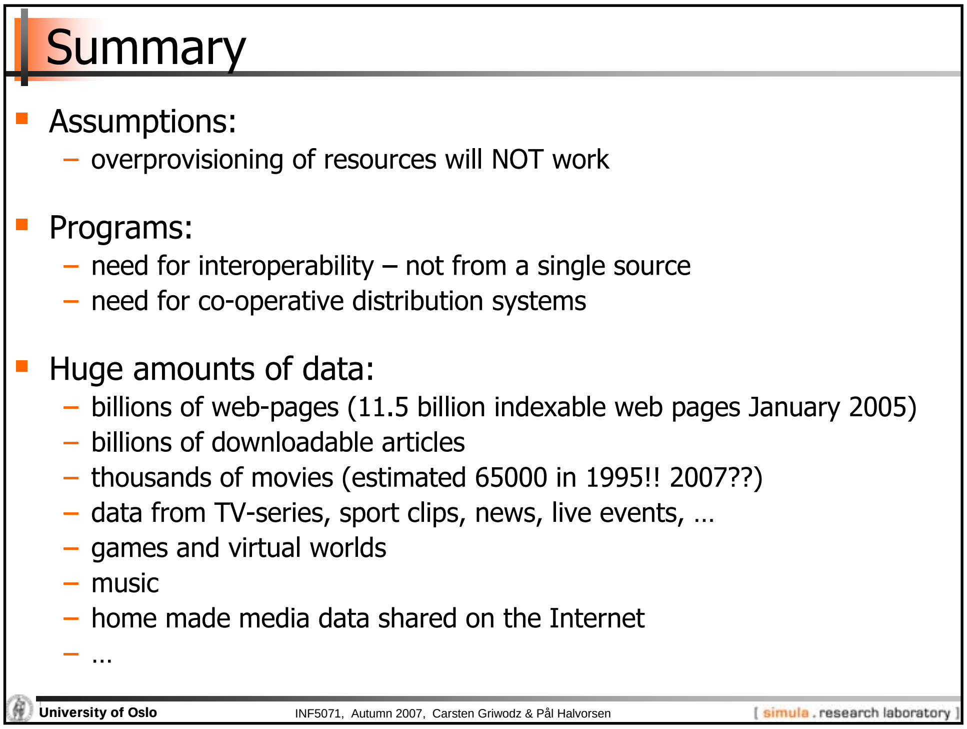


Application Characteristics

- Movie-on-Demand and live video streaming
  - Access pattern according to Zipf
  - high rates, many and large packets
  - many concurrent users (Blockbuster online – 2.2 million users)
  - extreme peeks
  - timely, continuous delivery
- Games
  - low rates, few and small packets
  - many concurrent users (WoW – 9 million players)
  - interactive
  - low latency delivery
- News-on-Demand streaming
  - daily periodic access pattern – close to Zipf
  - similar to other video streaming
- …

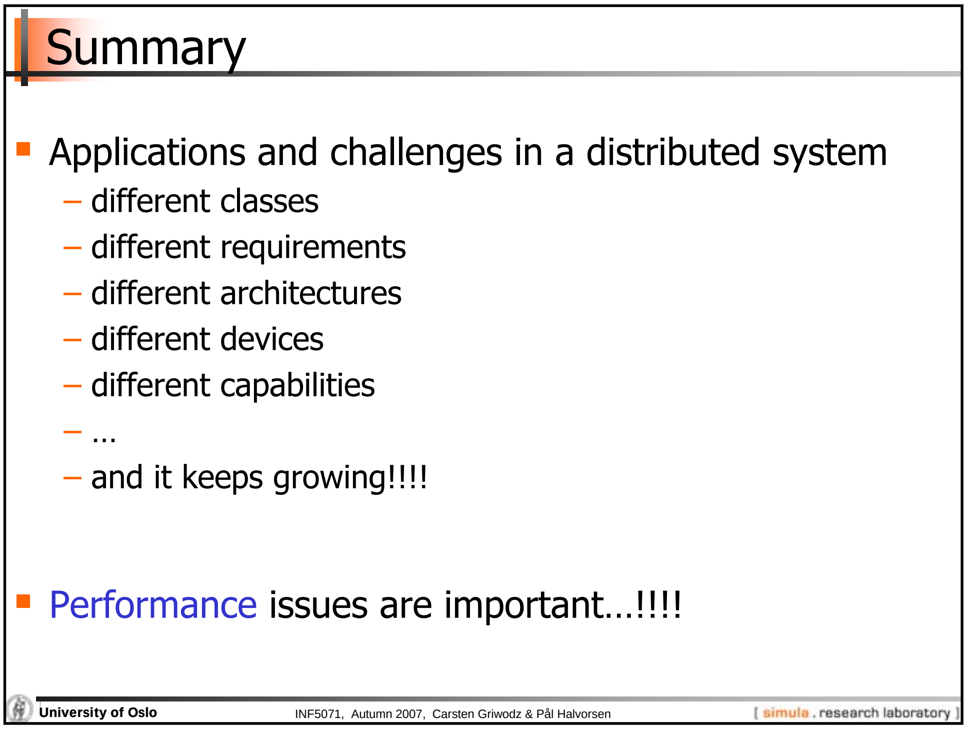


Picture Today!



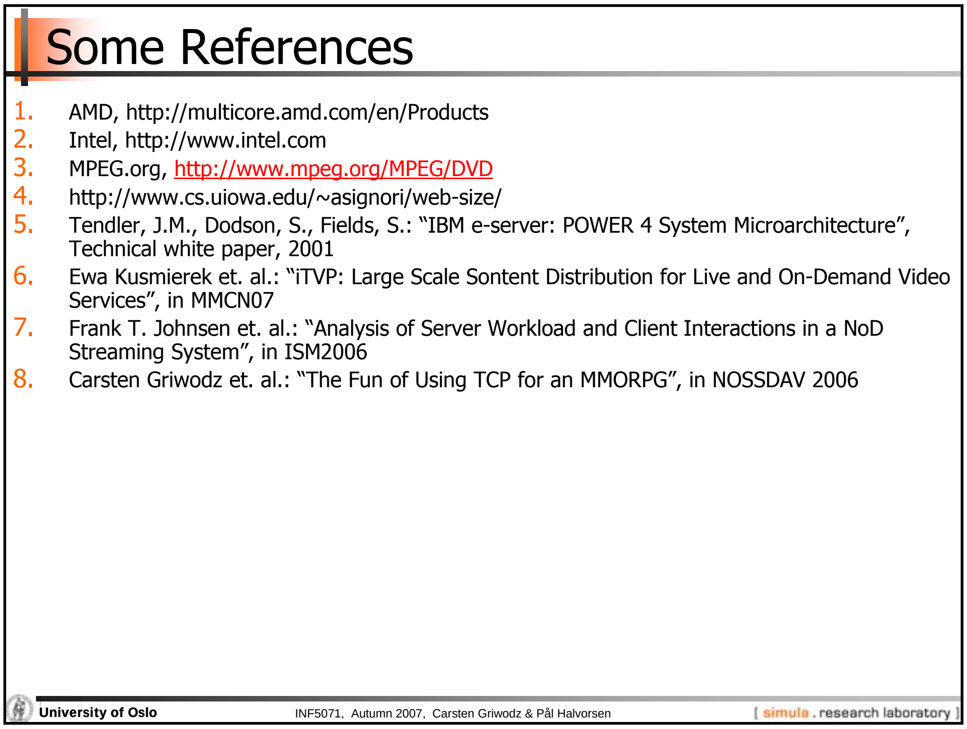
Summary

- Assumptions:
  - overprovisioning of resources will NOT work
- Programs:
  - need for interoperability – not from a single source
  - need for co-operative distribution systems
- Huge amounts of data:
  - billions of web-pages (11.5 billion indexable web pages January 2005)
  - billions of downloadable articles
  - thousands of movies (estimated 65000 in 1995!! 2007??)
  - data from TV-series, sport clips, news, live events, …
  - games and virtual worlds
  - music
  - home made media data shared on the Internet
  - …



Summary

- Applications and challenges in a distributed system
  - different classes
  - different requirements
  - different architectures
  - different devices
  - different capabilities
  - …
  - and it keeps growing!!!!
- Performance issues are important…!!!!



Some References

1. AMD, http://multicore.amd.com/en/Products
2. Intel, http://www.intel.com
3. MPEG.org, http://www.mpeg.org/MPEG/DVD
4. http://www.cs.uiowa.edu/~asignori/web-size/
5. Tendler, J.M., Dodson, S., Fields, S.: "IBM e-server: POWER 4 System Microarchitecture", Technical white paper, 2001
6. Ewa Kusmierek et. al.: "iTVP: Large Scale Sontent Distribution for Live and On-Demand Video Services", in MMCN07
7. Frank T. Johnsen et. al.: "Analysis of Server Workload and Client Interactions in a NoD Streaming System", in ISM2006
8. Carsten Griwodz et. al.: "The Fun of Using TCP for an MMORPG", in NOSSDAV 2006