Multimedia Data Distribution and Processing in IP Networks

Eva Hladká, CESNET & Masaryk University, Czech Republic

The Third International Conference on Advances in P2P Systems, AP2PS 2011

Lisbon, Portugal 23. 11. 2011

Lecture Content

1. Data distribution in IP networks
2. Virtual multicast
3. Active networks
4. Programmable router $\longrightarrow$ Active element
5. Data processing on AE
6. Active Elements with P2P Control Plane
7. CoUniverse
8. Split Streaming
9. Demonstrations
10. Conclusion and Future work

Data distribution in IP networks

- Generally: data transport from source to n destinations
- from 1 source to 1 destination
- from 1 source to n destinations
  - IP multicast
  - Virtual multicast

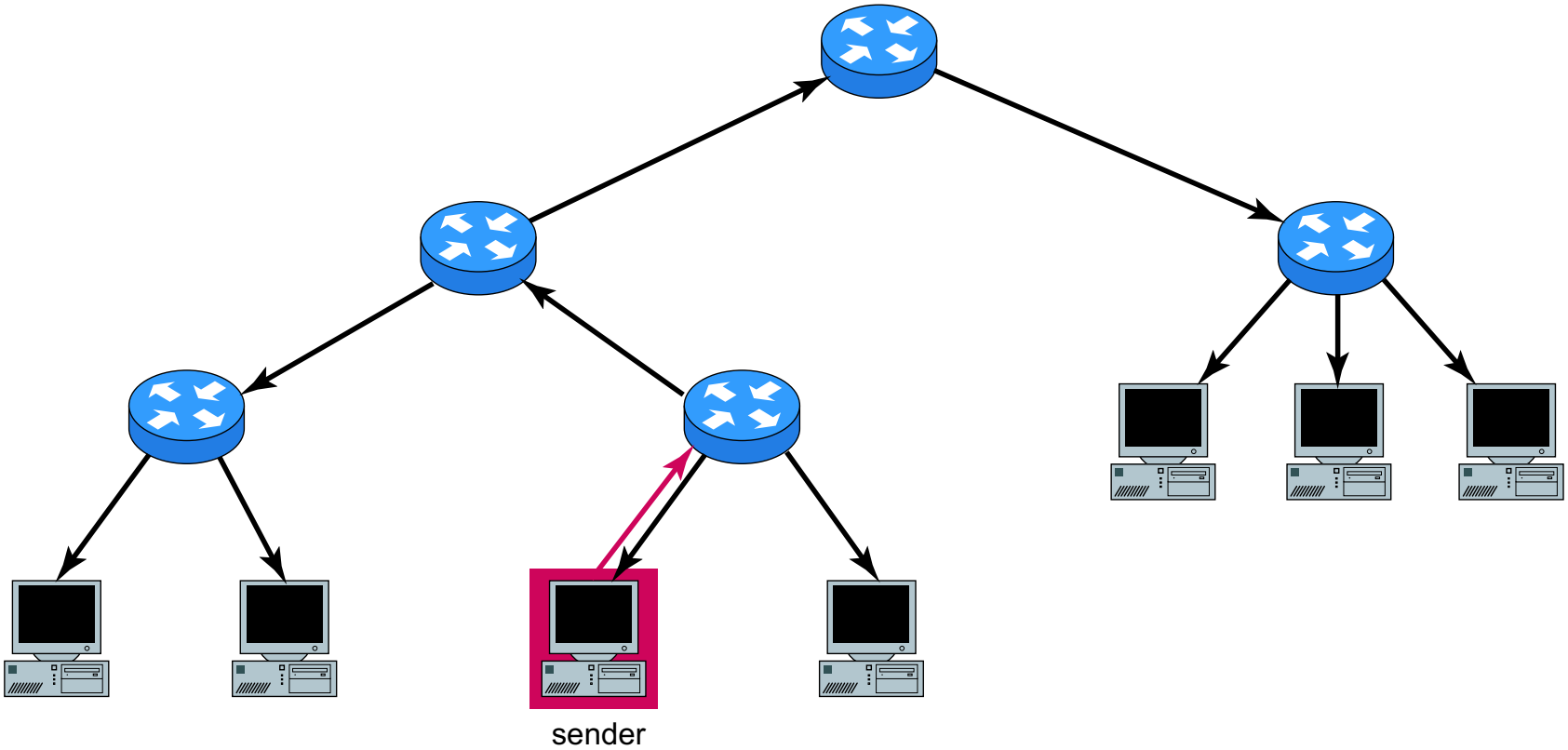
IP multicast

- At most one data copy per link
- Network property (hop by hop, not end-to-end service)
- Not reliable (best effort, UDP, group address)
- Range of spread is limited by TTL (Time To Live) in packet
- Protocols
  - Group management – Internet Group Management Protocol (RFC 1112), IGMPv2 (RFC 2236)
  - Routing – Source Based Tree, Shared Tree (Core Based Tree)
- Properties: scalability, problematic accounting, not reliable service, easy attack target

IP multicast – schema

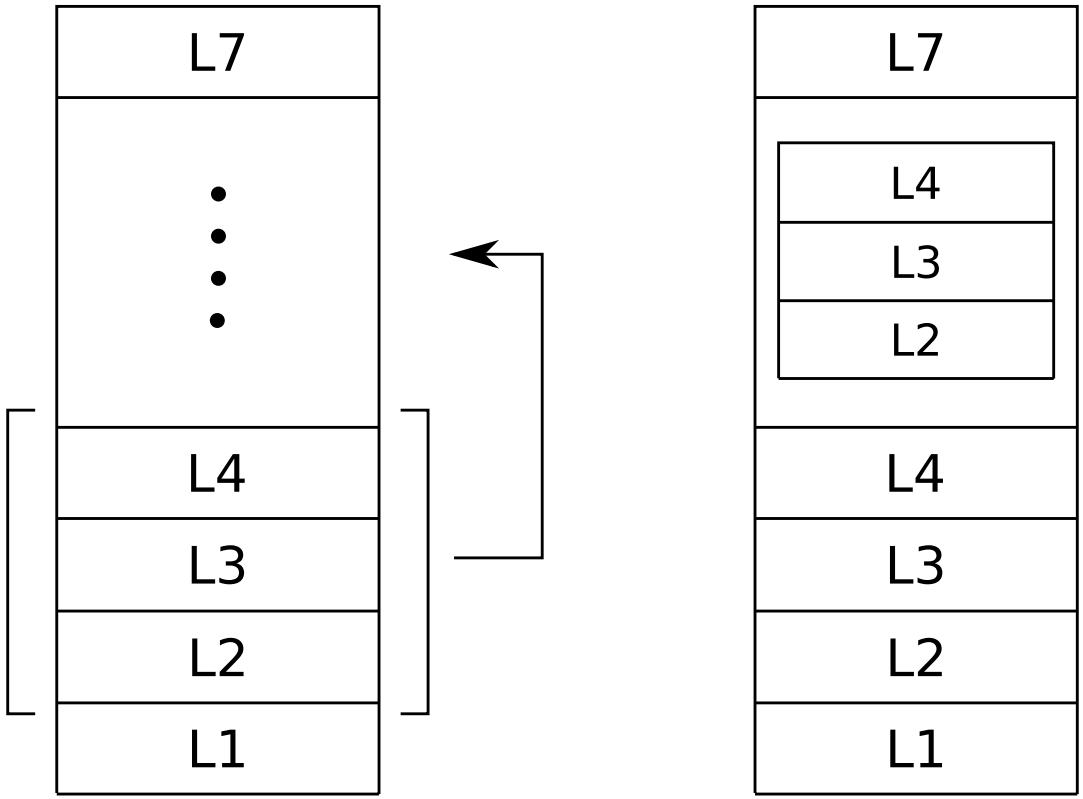
Multicast distribution tree

– At most one data copy per link.



Virtual multicast

Virtual network is an overlay network with functionality demanded by application and mapping to interconnecting network.

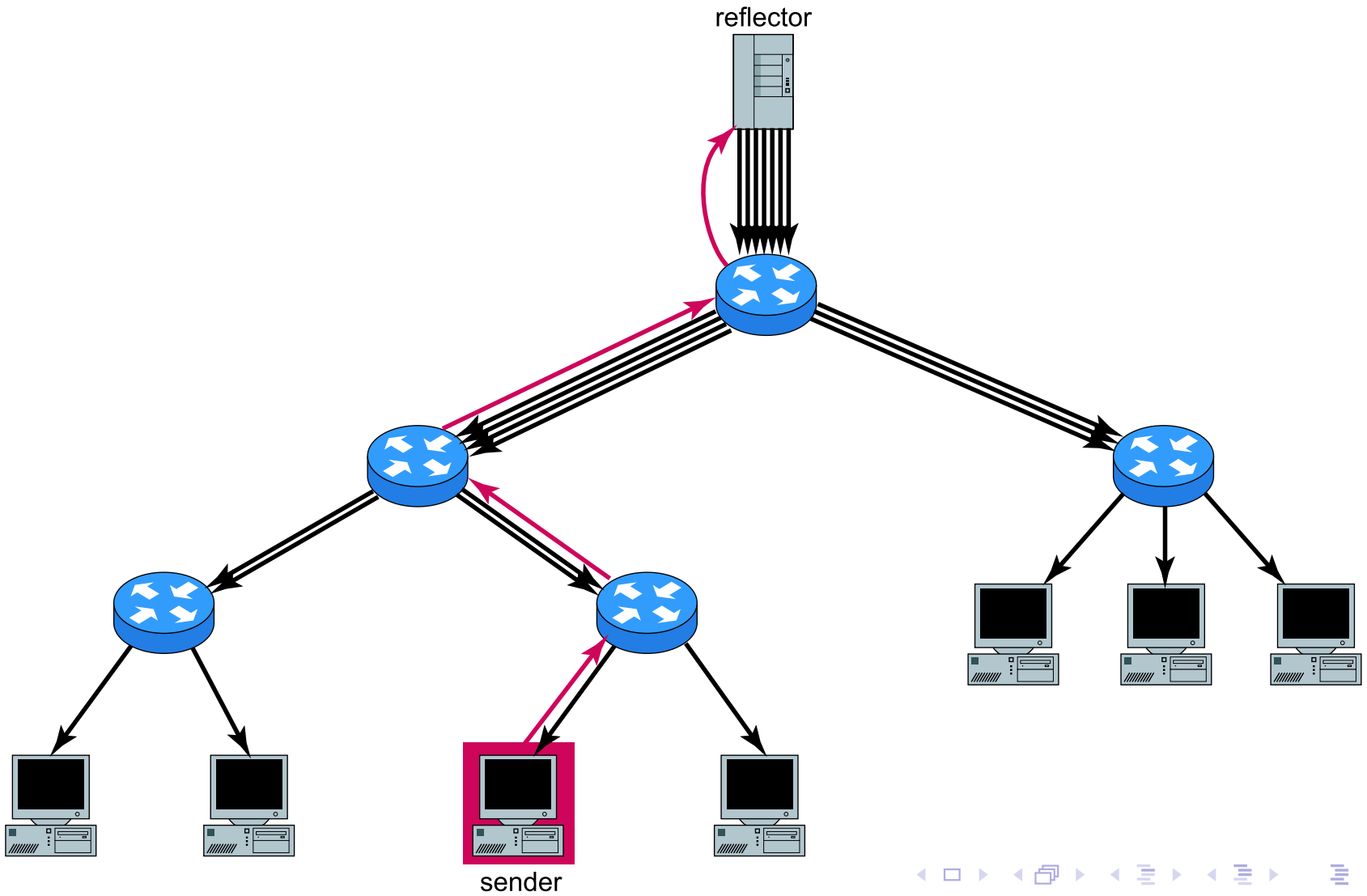


Virtual multicast is a realization of data distribution $1 : n$ in a virtual network.

Virtual multicast – schema

Virtual multicast distribution tree

– One data copy per host.



Advantages × disadvantages of virtual multicast

- efficiency, higher network load
- scalability
+ independency on network services
+ individual transport by end-client demands
+ managing during the transfer
+ security

Active networks

Passive transport medium $\longrightarrow$ distributed computing environment

- interior nodes provide user managed data processing
- passive links + active (programmable) nodes
- application examples: caching, video processing, reliable multicast, ...

Active networks - architectures

- Active packet
  - program code is inside of each packet
  - packet programming language NetScript
  - flexible, limited, big overhead
- Active nodes
  - program is injected into the node before data transfer
  - usual programming languages
  - statefull, security
- Combination of active packets and active nodes

Active networks

- new concept in networking
- 1995–2004
- a way to realize virtual/overlay networks
- applications

Virtual network and Active elements

Virtual network construction:

- On application level – tunnelling
- Overlay network based on replication elements
  - Active elements (AE) as a replication elements

Active element is a programmable network element

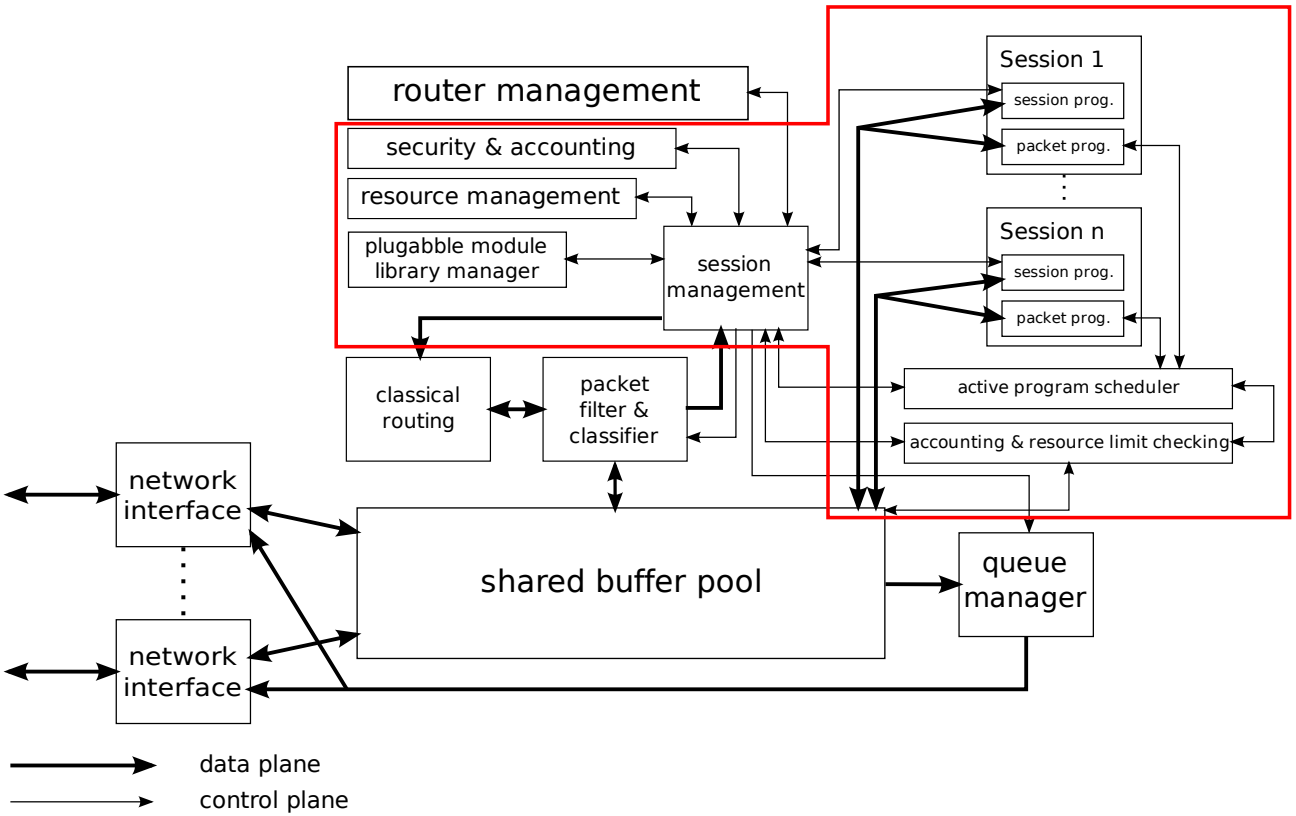
- AE works on application level and could be managed by user
- AE processes and forwards data
- AE is programmable on application level
  - AE does not intervene to networking stack on standard networks levels

Examples of the AE functionality

- Data replication
- Transport through firewalls
- Data formats translation
- Security of transferred data
- Data monitoring
- Logging and accounting
- Caching
- Multiple streams synchronisation

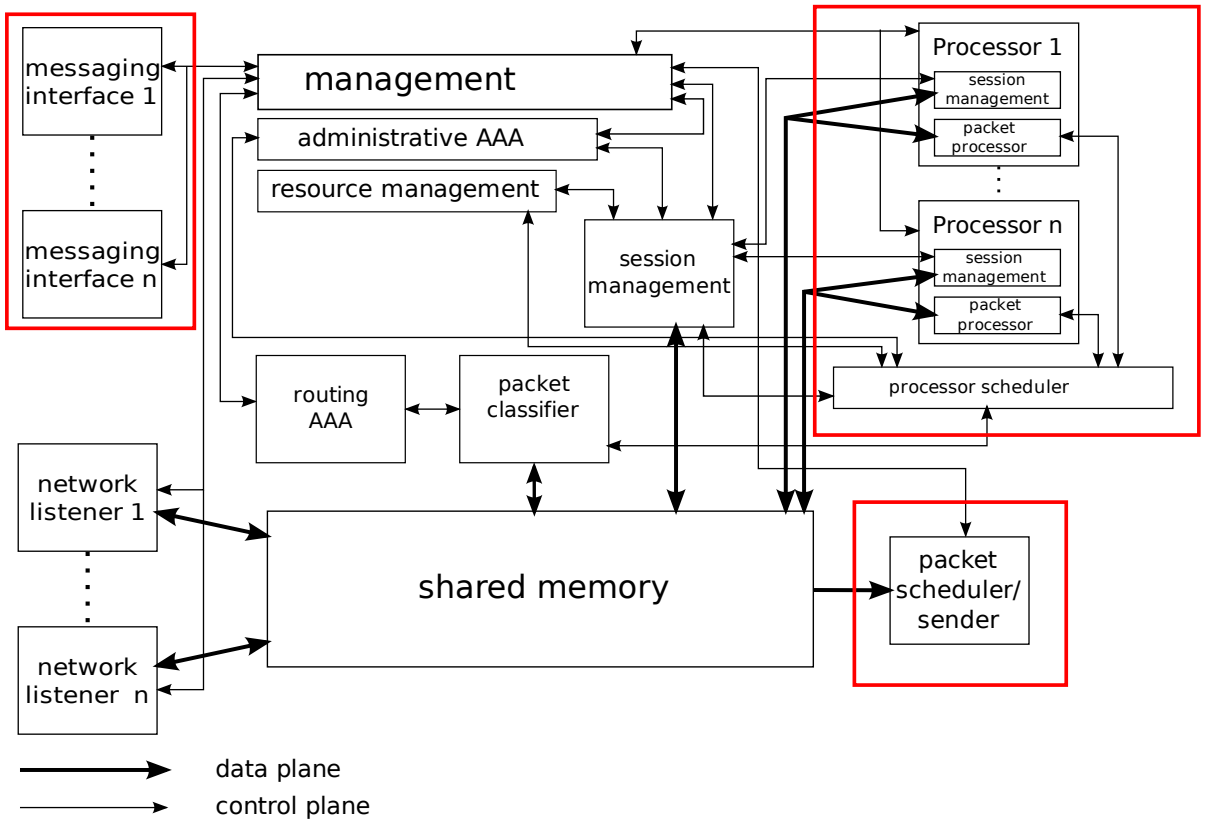
Active element evolution 1

- First step: General active router
  - Concept of programmable network on level network elements (L2, L3, L4)
  - Only prototypes on L7
  - Left due to complexity, low stability and price



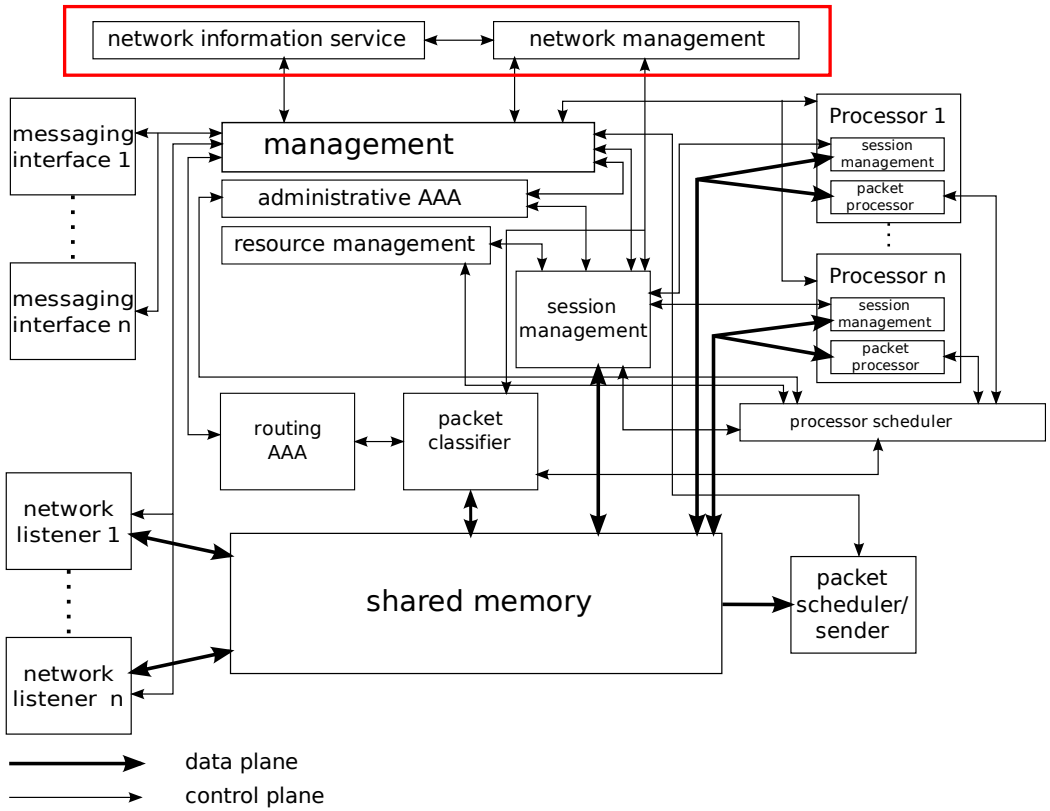
Active element evolution 2

- Second step: move to application level – active element
  - Independence on network elements, flexibility
  - Lower efficiency



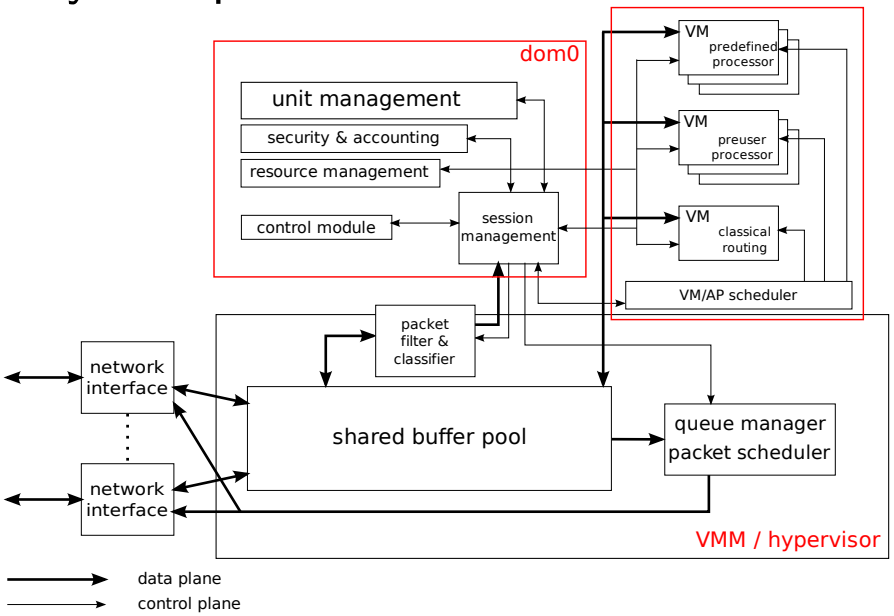
Active element evolution 3

- Third step: scalability
  - Active elements network
  - Distributed active element
  - Better efficiency



Active element evolution 4

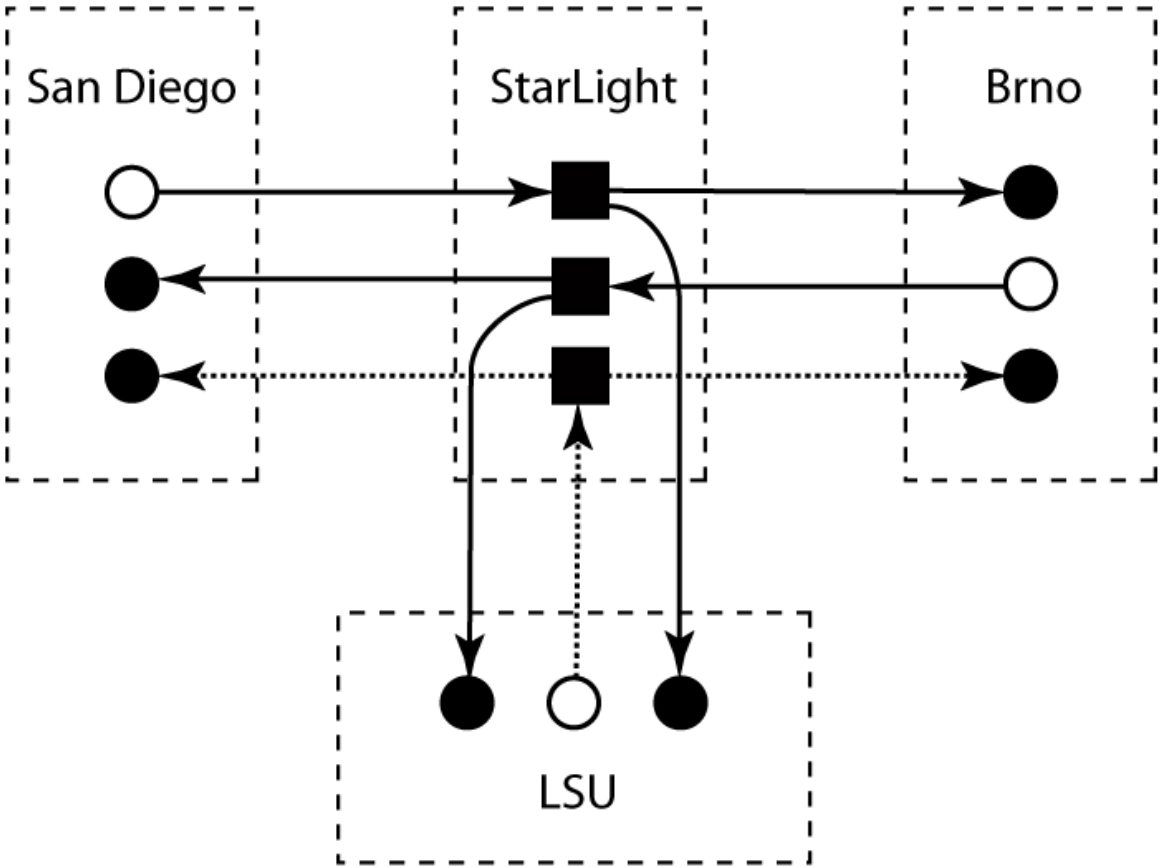
- Fourth step: Virtualisation AE
  - Better efficiency in return on network elements?
  - Complexity and price?



Active element performance – I

- Active elements used for replication 1,5 Gbps streams
- Dual AMD64 Opteron 250 (2,4 GHz CPU, 4 GB RAM)
Throughput [Gbps]	Packetloss [%]	CPU load [%]
1.8	0	52
1.9	0	55
2.0	0.01	60
2.1	0.04	76
2.2	1.7	80
2.3	7.1	84

Active element performance – topology

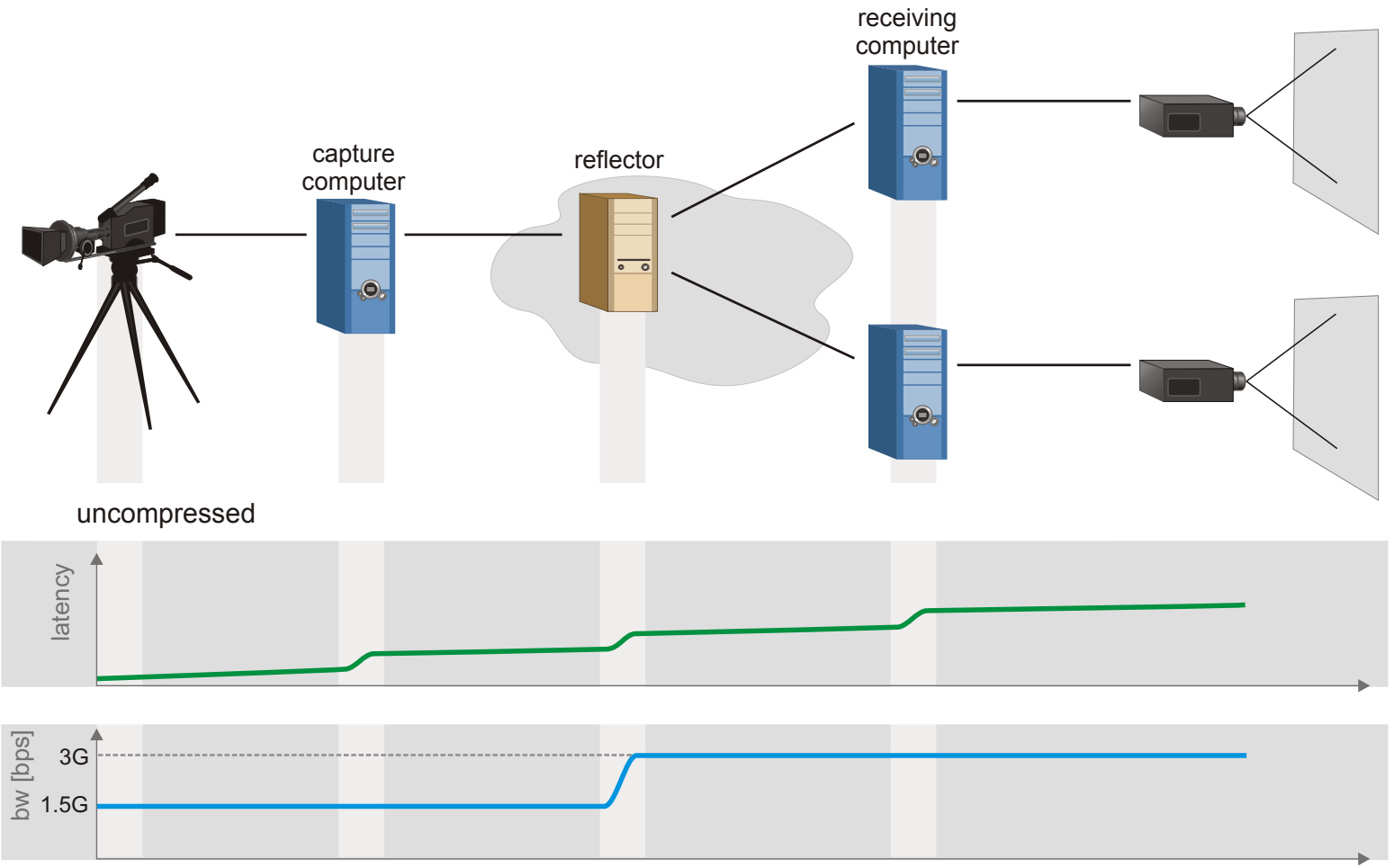


Active element performance – II

- Active element delay: 13 ± 2 ms
- Circuit delay:
  - San Diego $\longleftrightarrow$ StarLight: 78.2 ± 0.2 ms (routed)
  - Louisiana $\longleftrightarrow$ StarLight: 31.09 ± 0.04 ms (switched)
  - Brno $\longleftrightarrow$ StarLight: 126.7 ± 0.3 ms (routed)

Active element performance – III

- Connectivity scheme with time axis



Efficiency upgrade – Distributed AE

- Data stream is divided to substreams and each of them is processed separately
- Distributed AE could be part of an AE network
- Distributed AE is still user controlled
- It can fill line of any capacity

Active Elements with P2P Control Plane

- User-empowered approach to synchronous data distribution
- Separation of control and data plane
- P2P control plane
  - monitoring of the network status for data layer organization
    - aggressive monitoring for smaller networks
  - information on available content
  - clients of the network may or may not be able for data forwarding
- Pluggable data distribution plane (various models)
  - from full mesh (relying entirely on IP routing)...
    ...to multiple (as minimum as possible) spanning trees

jSon – Prototype Implementation

- Based on JXTA-C implementation of JXTA 2.0 standard
  - uses C following the performance-optimized AE implementation
- Server side
  - modules for the AEs
  - JXTA interface
  - network monitoring
  - information service
- Client side
  - JXTA interface
  - control of legacy applications

jSon – Evaluation

- Modifications in JXTA parameters reduces failure detection and recovery from 60 s to 1 s
  - default JXTA setup is designed for traditional applications, where such a fast reactions are not necessary/desirable
- Scales well
  - linear growth of total number of messages w.r.t. number of nodes
  - constant number of messages per node

Active Elements with P2P Control Plane Applications

- Synchronous data distribution for high-bandwidth streams

MBone Tools	more than 10 Mbps/stream
DV or stereo DV	30/60 Mbps/stream
uncompressed HD	1.5 Gbps/stream

- iGrid 2005, SuperComputing 2005
  - global uncompressed HD over 10GE/lambda network infrastructure
- Interesting for emergency situation support

Active element – secure communication

- IP Multicast – shared key
- Virtual multicast with AE
  - serial distribution and processing on AE – individual key
  - solution with VPN, virtual traffic division

	no VPN	UDP VPN	TCP VPN	TCP VPN + HTTP proxy
pchar latency [ms]	3.51	3.69	3.94	3.93
iperf jitter [μs]	6	6	9	13
pchar capacity est. [Mb/s]	39.8	35.2	20.1	19.8
iperf packet loss @ 30 Mb/s [%]	0.0	0.0	0.0	0.0
iperf CPU idle @ 30 Mb/s [%]	48.9±0.2	41.7±0.4	44.5±0.4	42.6±0.4

  - usage of federations for authentication and administration of admitting points

Applications

- Simple videoconference support
  - Since Y2K many groups, regularly
- HD videoconference
  - iGrid 2005 – demonstration of the first uncompressed HD multipoint videoconference
- Reliable and secure videoconferencing for medical consultations
  - Project Ithanet (6th EU Framework Programme)

Applications 2

- Advanced videoconferencing environment
  - Subgroup communication
  - Moderating
  - Video stream composition
- Stereoscopic video
  - Point-to-point or multipoint transfer
  - AE is used for stream synchronisation

Special applications – visualisation

- More computing (Grid) oriented
- Combination of high volume real-time visualization in real time with collaborative environment
  - HDTV stream generated in Baton Rouge and transported to Brno and San Diego
  - In parallel, videodata are transferred in internal format of used visualisation protocol
  - Data for visualisation were generated in Baton Rouge and in Brno, or at other places in USA/Europe
    - Data streams in Gbps
  - Goal: to explore possibilities of computer visualisation with HDTV, data replication in network, interaction with the computation from the place of visualisation (San Diego)

Applications – visualisation

Applications 3

Lecture PV177 - Introduction to High Performance Computing

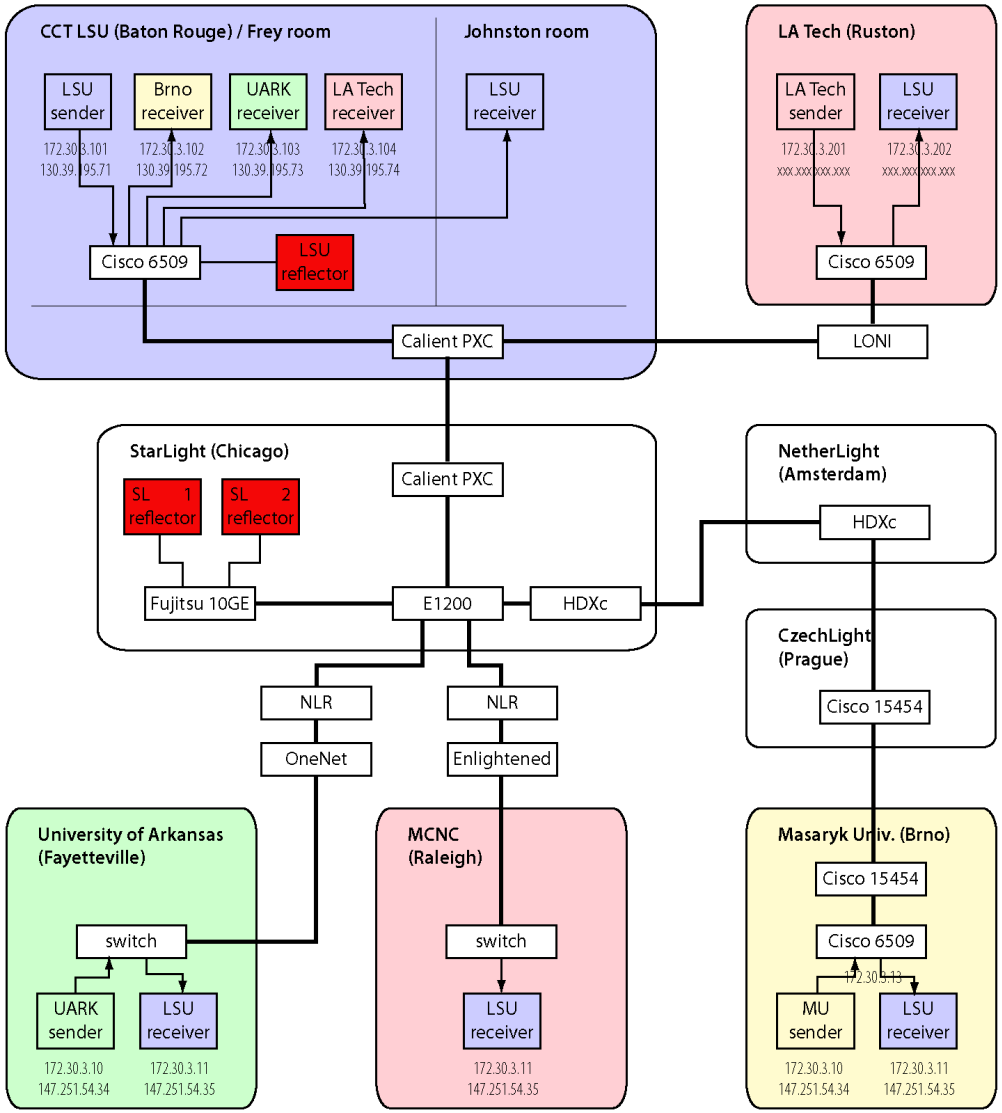
- Prof. Thomas Sterling, LSU, LA, USA

Five remote organisations (year 2007):

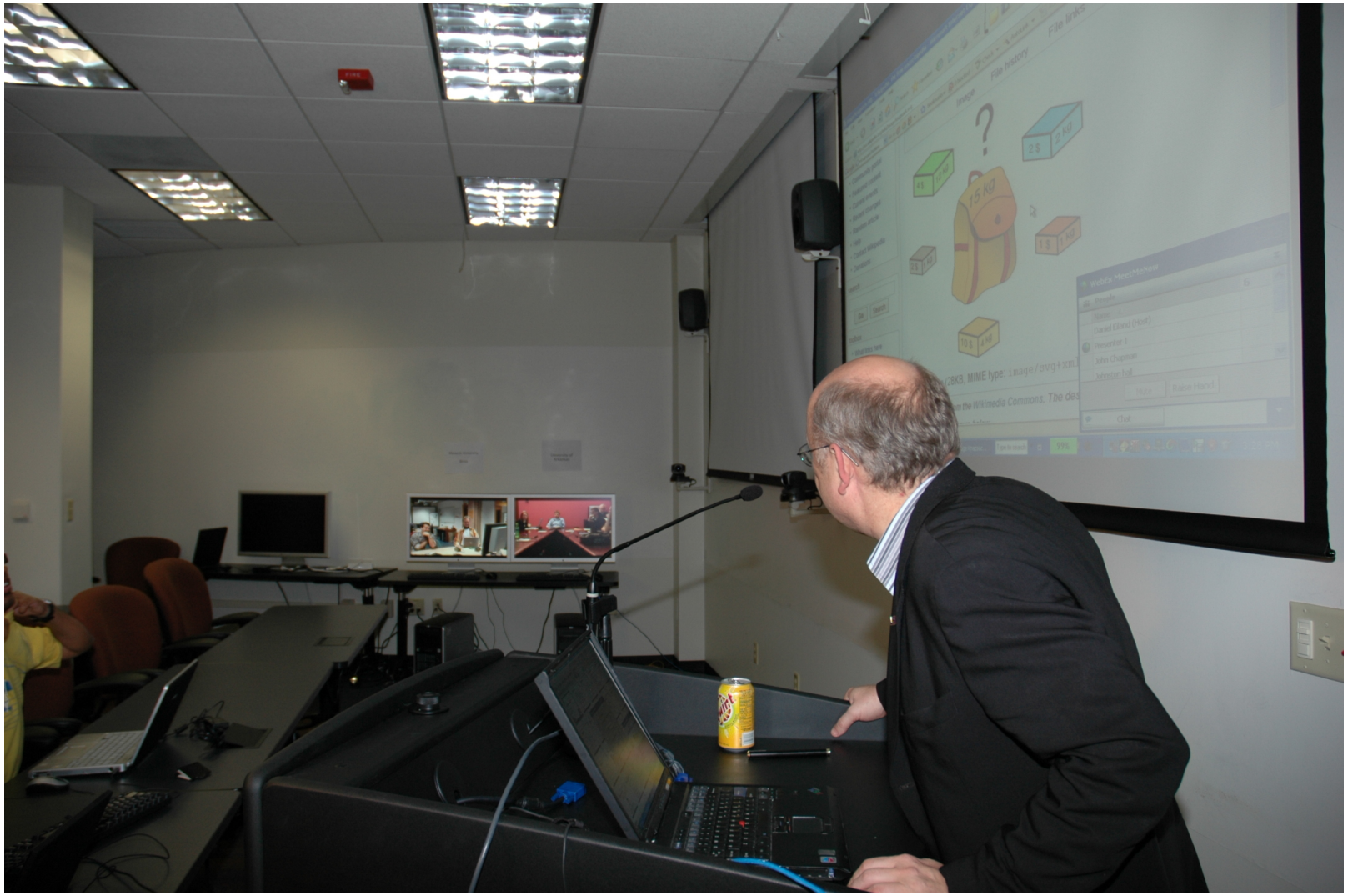
- Faculty of Informatics, Masaryk University
- University of Arkansas
- Louisiana Technical University
- MCNC, North Carolina
- North Carolina State University

Spring semester (January – June) 2007, 2008, 2009, 2010

Data Distribution for PV177



Prof. Thomas Sterling – lecture



Demonstrations

- iGrid 2005, 26. – 29. 9. 2005, Callt2, University of California, San Diego
  - HD Multipoint Conference
  - Interactive Remote Visualisation across the LONI and the National LambdaRail
- SC'05
- SC'06
- Glif 2007 – CoUniverse: Self-Organising Collaborative Environment
- SC'07 – CoUniverse demo
- RedStick2007
- I2 Fall 2008 member meeting – Dynamic Circuit Networking-enabled HD UltraGrid Videoconferencing

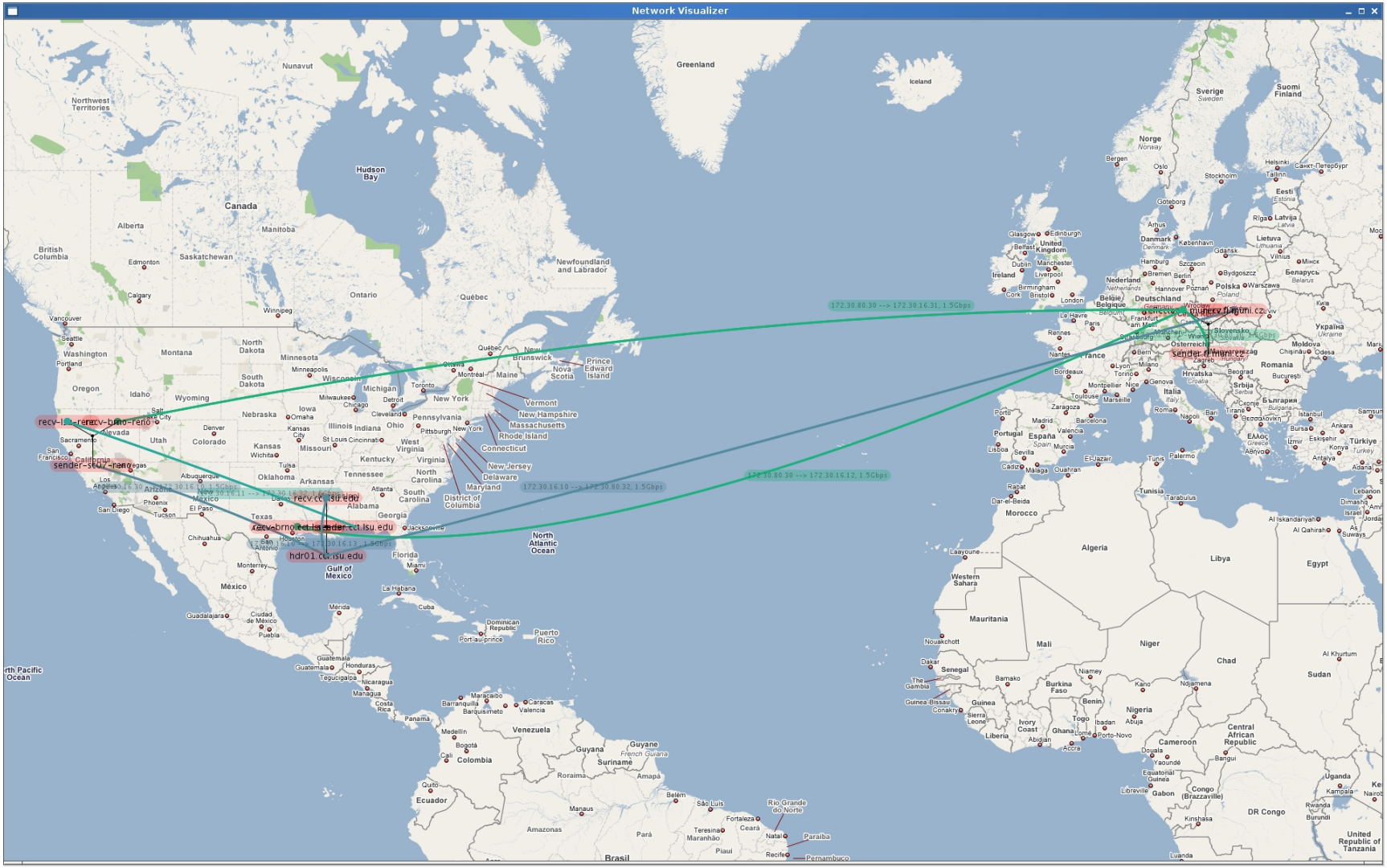
CoUniverse

- Self-organizing collaborative environment
- Pioneering middleware for orchestration of interactive collaborative environments
- Motivation: everyday experience with high-quality collaborative environments supported by AEs
  - High number of components: network nodes (AEs, workstations), network links, applications
  - Too many applications to be configured and started manully, one by one
  - Difficult management of network links and interfaces
  - Slow reaction on dynamic events in network (namely failures)
  - High risk of configuration errors
- Daily manual configuration is intractable for end-users, who are not experts in underlying technologies

CoUniverse

- Configuration and running of media-processing applications
  - Producers, consumers, and distributors (Active Elements)
  - Direct support for the frequently used applicaitions (e. g., UltraGrid, UCL Media Tools, RumHD) and hardware devices (Polycom)
  - Indirect support for other tools
- Dynamic configuration and allocation of network links
- Network topology visualization
- Peer-to-peer control plane

CoUniverse – topology



I2 Fall 2008 member meeting – I

I2 Fall 2008 member meeting – press release

The data for the UltraGrid demo will be distributed by application-level modular programmable UDP packet reflectors that have been developed over the past five years by CESNET and Laboratory of Advanced Networking Technologies. **This technology allows for independence on network-native multicast, while it is possible to process the data in per-user specific way**.
Both iHDTV and UltraGrid technologies are under active development by the research and education community. Through the iHD DevCore partnership, the community is currently investigating how to create interoperability between these platforms to enable more widespread adoption of uncompressed high-definition video technology.

Data transfer scheduling in CoUniverse

- High requirements on media quality in some applications – healthcare (surgery video transmissions), movie industry (remote postprocessing)
- Bandwidth of applicable media streams is comparable to capacities of links (9 Gbps for uncompressed 4K video)
- Risk of data losses due to network components overload, congestion of links, massive reordering
- Sophisticated planning of stream routes is required instead of standard hop-by-hop best-effort approach

Data transfer scheduling in CoUniverse

- Requirements on routing schedule:
  - Place all data distribution trees in the network, respecting all applications' requirements
  - Maintain capacity restrictions on links
  - Apply optimization criterion, e. .g, overall latency
- Physical network topology is not known to CoUniverse. End-to-end topologies of subnetworks are used instead.
- Uncertainty introduced by possible sharing of physical links by virtually independent end-to-end links
- Scheduling may consider features of AEs (format translation)
- Originally constraint programming approach, newer integer programming approach improves scalability

Split streaming

- Only integral flows have been considered so far
- Legacy applications cannot handle splitting of data at intermediate nodes (AEs)
  - Massive packet reordering in real-time applications incurrs data loss
- Split streaming could enable efficient scheduling and bandwidth utilization
- How should future media tools be built to benefit from split streaming?

Challenges in using split streaming

- Quality of parallel data transfer is determined by the weakest subpath (e. g., in terms of latency)
- The problem of creating a splitting scheme which is optimal (i. e., with respect to latency) whithout the knowledge of physical topology
- Splitting scheme may depend on the media content or specific applications' requirements
- Synchronization mechanisms should be employed at end nodes to reconstruct the stream with required quality.

Future work

- CoUniverse
  - Links planning, dealing with discrepancies between physical and virtual links, both in integral and split streaming
- Is virtualisation a solution of efficiency problem?
  - Return back to the lower network levels
- Next work on applications
  - Higher security level
  - Scalability on higher speed
  - Medical applications and their specific demands
- mobile collaborative environment
- Protocols with explicit latency compensation to support collaborative environments

Conclusion

- 1 : n data distribution – challenge for network protocols
- Native solution vs. virtual networks
- Active element is a key stone of virtual networks
  - From concept to implementation
- Ideas confirmed by applications
  - Administrations and reliability of use
  - Extreme traffic demands (demonstrations and routine traffic)
- Identifying the requirements on future high-quality media applications
- CoUniverse – Self-organizing collaborative environment

Thanks to

- Employees and students from Laboratory of Advanced Network Technologies

- VZ Optical network for national research MŠM 6383917201
- VZ Large scale parallel and distributed systems MŠM 0021622419
- Ithanet–Electronic Infrastructures for Thalassemia Research Network (RI-2004-026539)
- MediGrid – methods and tools for use of Grids in biomedicine AV ČR T2 0209 0537

Thank for your attention

questions?