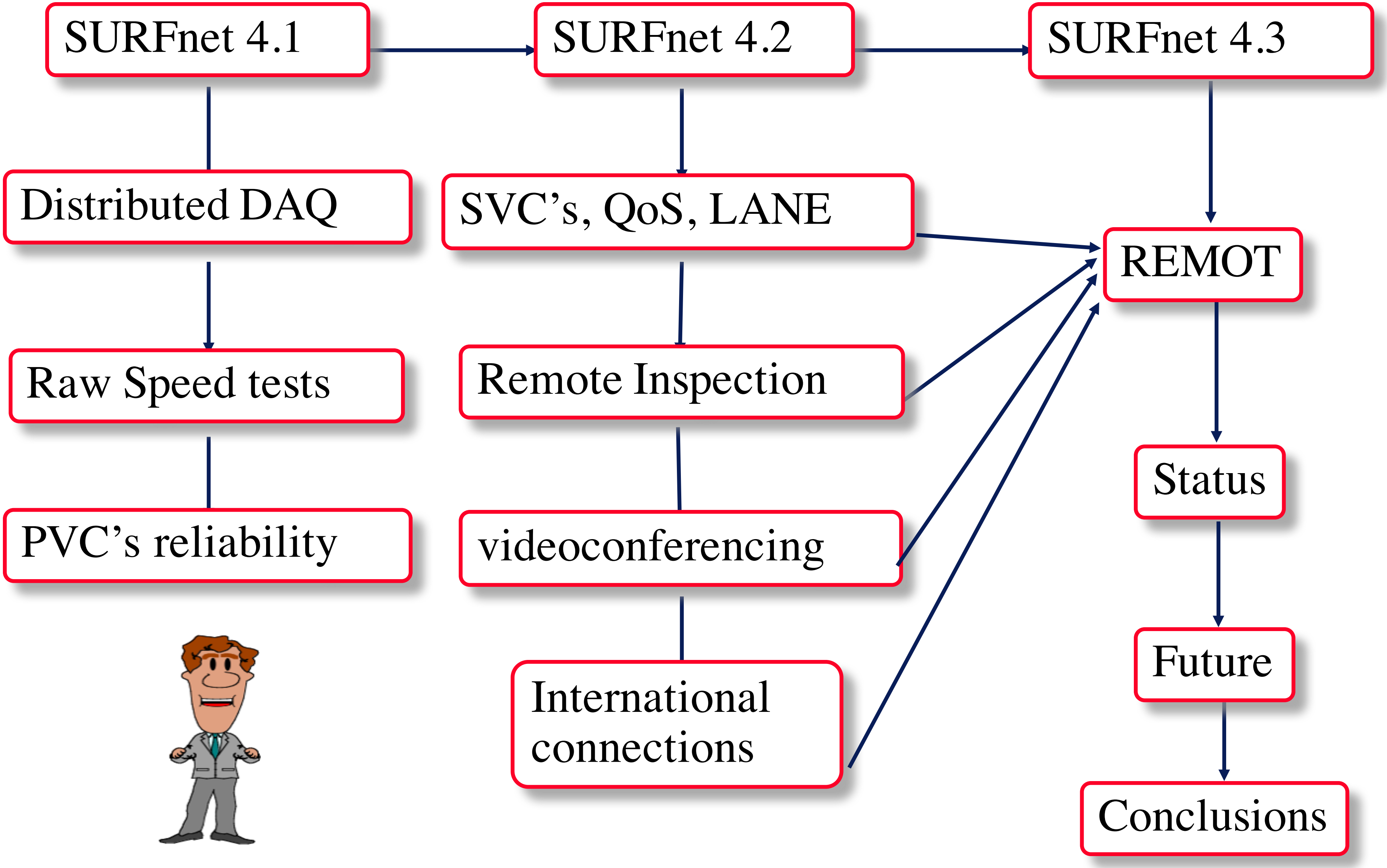
Experiences with the application of ATM network technology in experimental physics.

C.Th.A.M. de Laat, P.G. Kuijer, V. Giesing, P. Olthuis, J. Venema

Physics and Astronomy Department

Universiteit Utrecht

Outline of this talk



SURFnet 4.1

Distributed DAQ

Raw Speed tests

PVC's reliability

SURFnet 4.2

SVC's, QoS, LANE

Remote Inspection

videoconferencing

International connections

SURFnet 4.3

REMOT

Status

Future

Conclusions

SURFnet4 project

SURFnet bv is the Dutch research network organisation

SURFnet4 is a joint project of SURFnet bv and PTT Telecom

Aim:

- **bring research network backbone on ATM technology to cope with yearly doubling of traffic**
- **Introduce new application specific services**

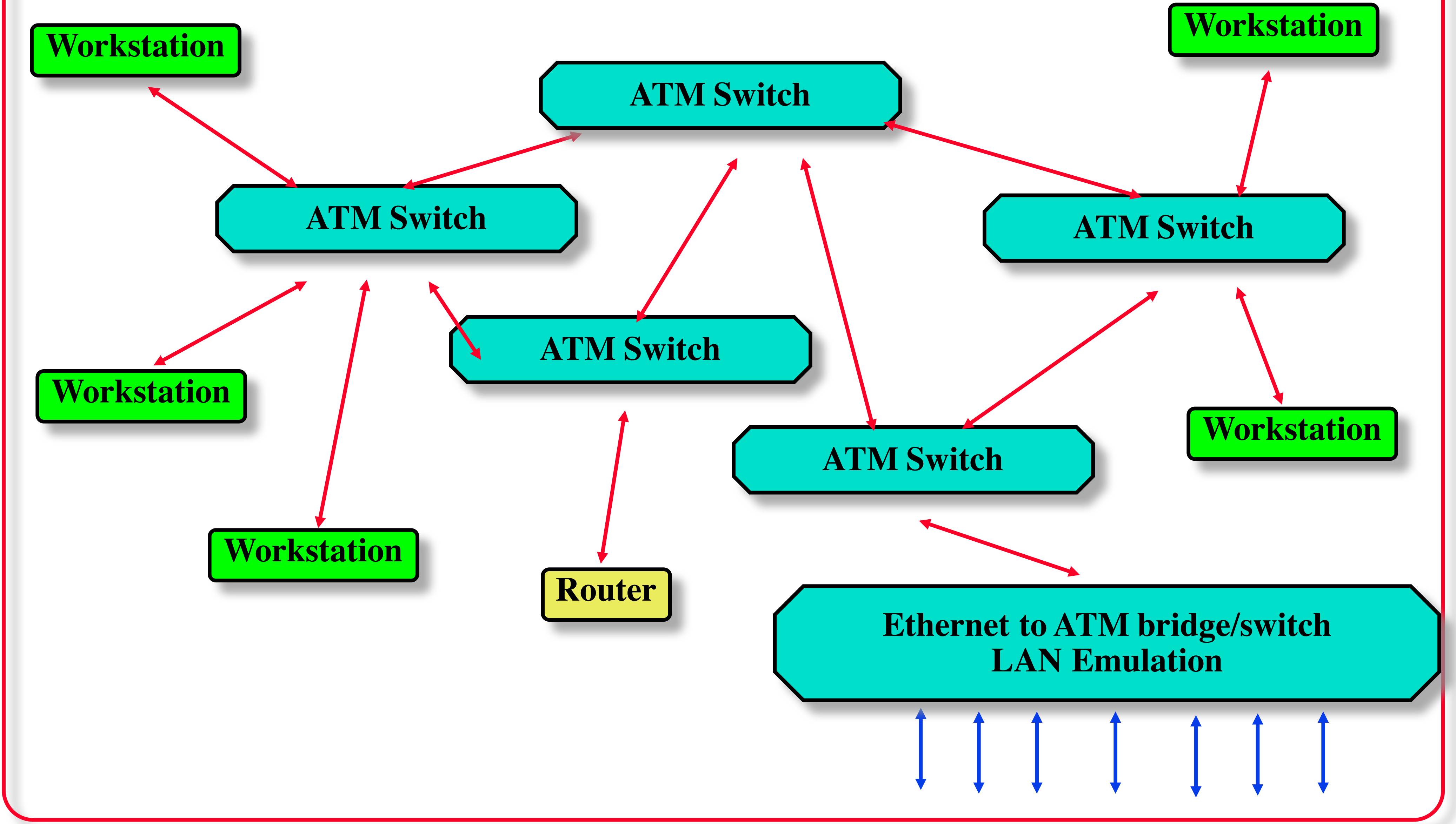
Why ATM:

- **more bandwidth (4 -> 34 -> 155 Mbit in 1996)**
- **scalable (34, 100, 155, 622, 2400 Mbit)**
- **fixed length cells with addresstion -> hw-routing**
- **Quality of Service per connection**
- **allows LAN-services mixed with sound/video channels**
- **allocatable <-> shared bandwidth**
- **billing possibilitiesBECAUSE HOLLYWOOD WANTS IT!!**

Asynchronous Transfer Mode (ATM)

- ATM
  - Fixed sized cells containing addresses
  - Proccessing optimized
    - » size and location of cell known
    - » flexible since each cell knows where it is going
  - combines STM and PTM
    - » Cell synchronous
    - » Cell Addressing
    - » Scalable Bandwidth
    - » Flexible bandwidth
  - Cell layout is independent of physical layer transport -> cell format does not change when going to other speeds
  - B-ISDN standard (Broadband Integrated Services Digital Network)
  - ATM-Forum
    - » UNI and NNI specifications

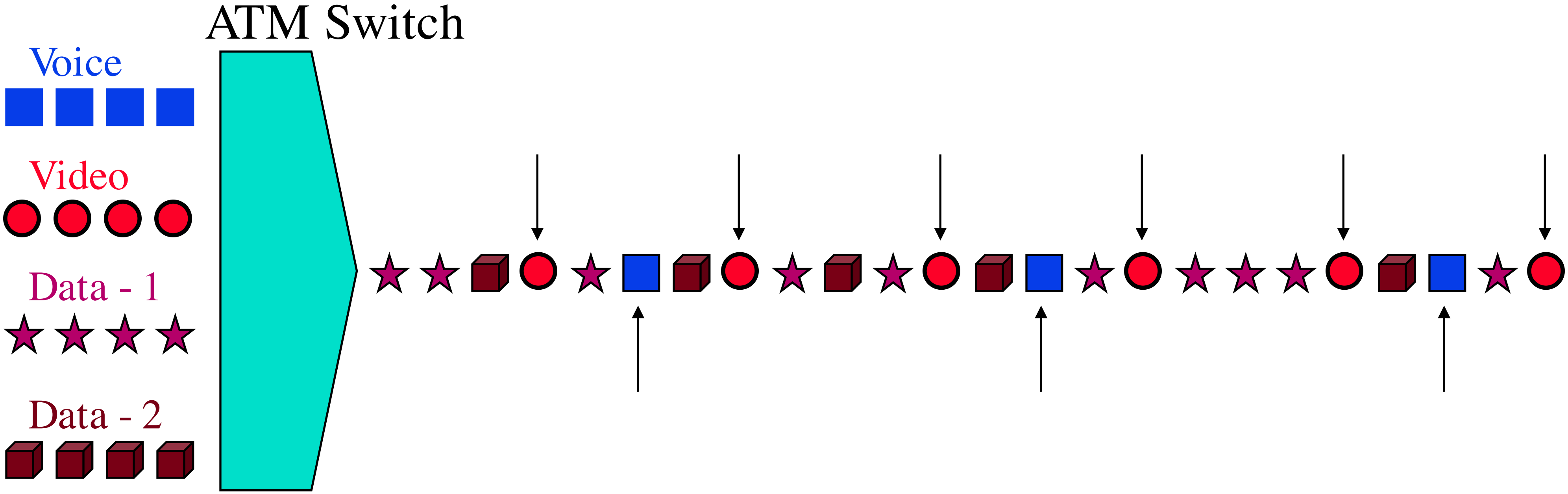
ATM Network


Workstation
Workstation
ATM Switch
ATM Switch
ATM Switch
Workstation
ATM Switch
ATM Switch
Workstation
Workstation
Router
Ethernet to ATM bridge/switch
LAN Emulation

ATM Layer

- **ATM-cell**
  - 53 bytes consisting of 48 bytes data and 5 bytes header
  - Header
    - 4 bits => Generic Flow Control (GFC) or Virtual Path Identifier (VPI)
    - 8 bits => Virtual Path Identifier (VPI)
    - 16 bits => Virtual Connection Identifier (VCI)
    - 3 bits => Payload Type Indicator (PTI)
    - 1 bit => Cell Loss Priority (CLP)
    - 8 bits => Header Error Check
    - 40 bits

Cell Multiplexing



- **Connections**
  - end to end (like telephone system), end can be a router
  - Constant Bit Rate (CBR) -> every nth cell
  - Variable Bit Rate (VBR) -> Guaranteed mean which may be exeeded
  - Unspecified Bit Rate (UBR) -> idle cells, no guarantee at all
  - Available Bit Rate (ABR) -> idle cells with flow control to minimize cell loss

Cell Switching

- ATM Switch
  - VC's are multiplexed in VP's
  - Switching
    - » look at incoming cell's port number and VP/VC
    - » table lookup gives destination port and VP/VC
    - » insert in output que taking into account type of connection and bit rate
  - Switching can be done in hardware lookup tables, ass. memory
    - » fast
    - » fixed cell format -> cell header inspection forwarding while receiving
  - Signaling
    - » User Network Interface (UNI 3.0 & 3.1)
    - » Network to Network Interface (NNI)
    - » Connection setup for Switched Virtual Circuits (SVC's)
    - » Error recovery, Resilient Virtual Circuits (RVC's)
    - » Management

Flow Control

- Flow control systems
  - credit based flow control
    - receiver has room in input buffers
    - receiver sends credits for those free buffers to sender
    - sender may send as much as it has credits for
    - works point to point for each link
  - rate based flow control
    - depending on clp bits in header and congestion information in network the endnodes have to calculate the available bandwidth
    - statistical calculation which can go wrong
    - works end to end over a network
    - part of UNI 4.0

ATM Adaption Layer

- Convergence Sublayer (CS), prepares for segmentation
- Segmentation and Reassemble Sublayer (SAR)
- AAL 1 -> Voice/Video
  - CBR, connection oriented, timing relation source and destination
  - compensation for delay variation
- AAL 2 (RIP)
  - VBR, Connection Oriented -> Packet/Video
- AAL 3/4 -> Data
  - VBR, Connection Less/Oriented, no timing relation required
- AAL 5 -> Data
  - VBR, Connection Less/Oriented, no timing relation required
  - SEAL (Simple and Efficient Adaption Layer)
  - used in IP over ATM via RFC 1577, Lan Emulation

SURFnet4 Phase 1, 2 and 3

Phase 1: ATM test

- Test ATM technology
- Start with two sites: Amsterdam and Utrecht
- Pilot applications testing ATM from desk to desk
- Test and exploit specific capabilitiesavailable bandwidth - constant bit rate
- 34 Mbit backbone
- Timeframe: Aug

Phase 2: Services test

- Seven more research sites connected
- International connections
- Timeframe: 1995

Phase 3: Expansion

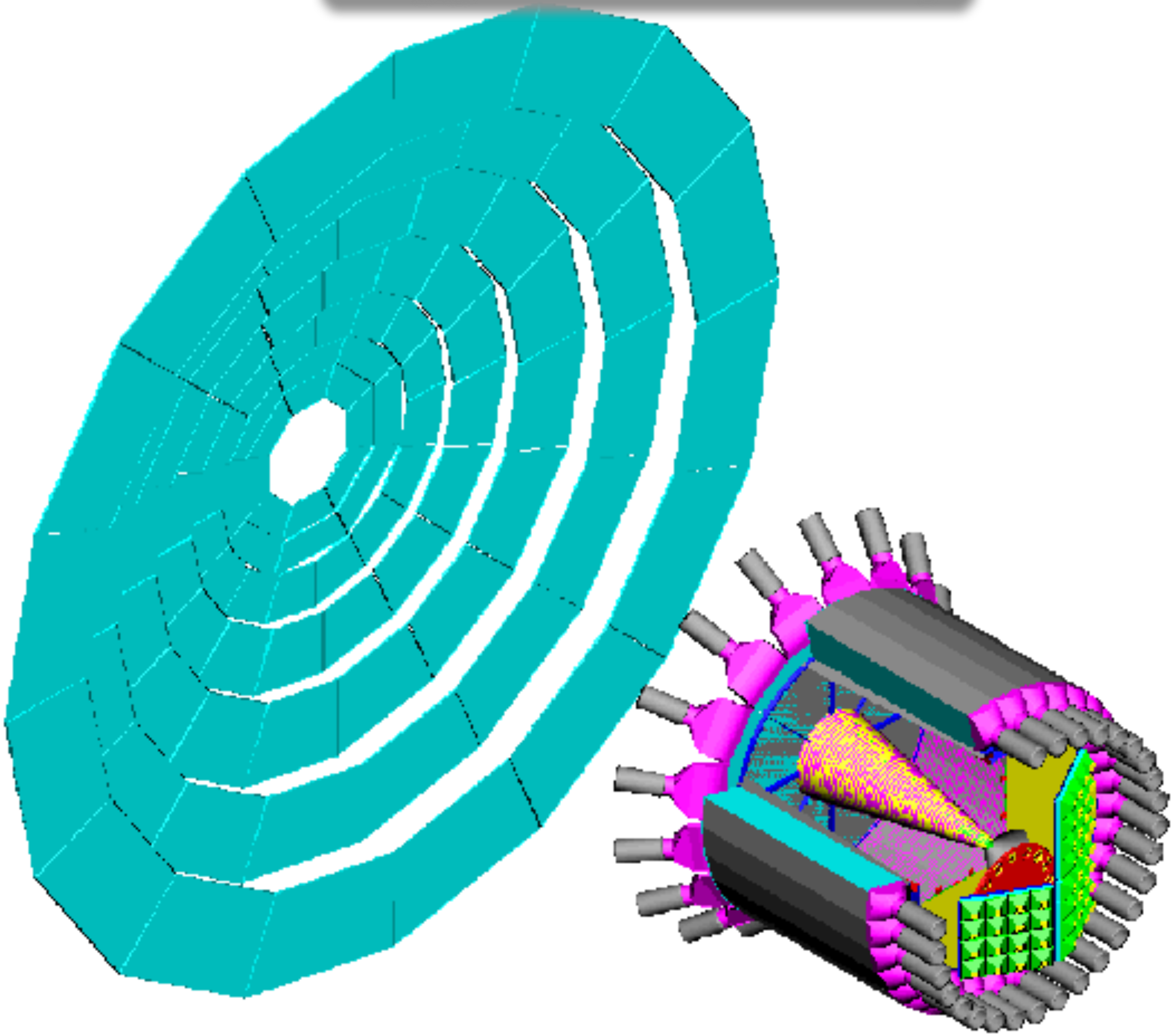
- operational ATM services
- 155 Mbit backbone
- Timeframe: 1996 -

□◀◀◆▶

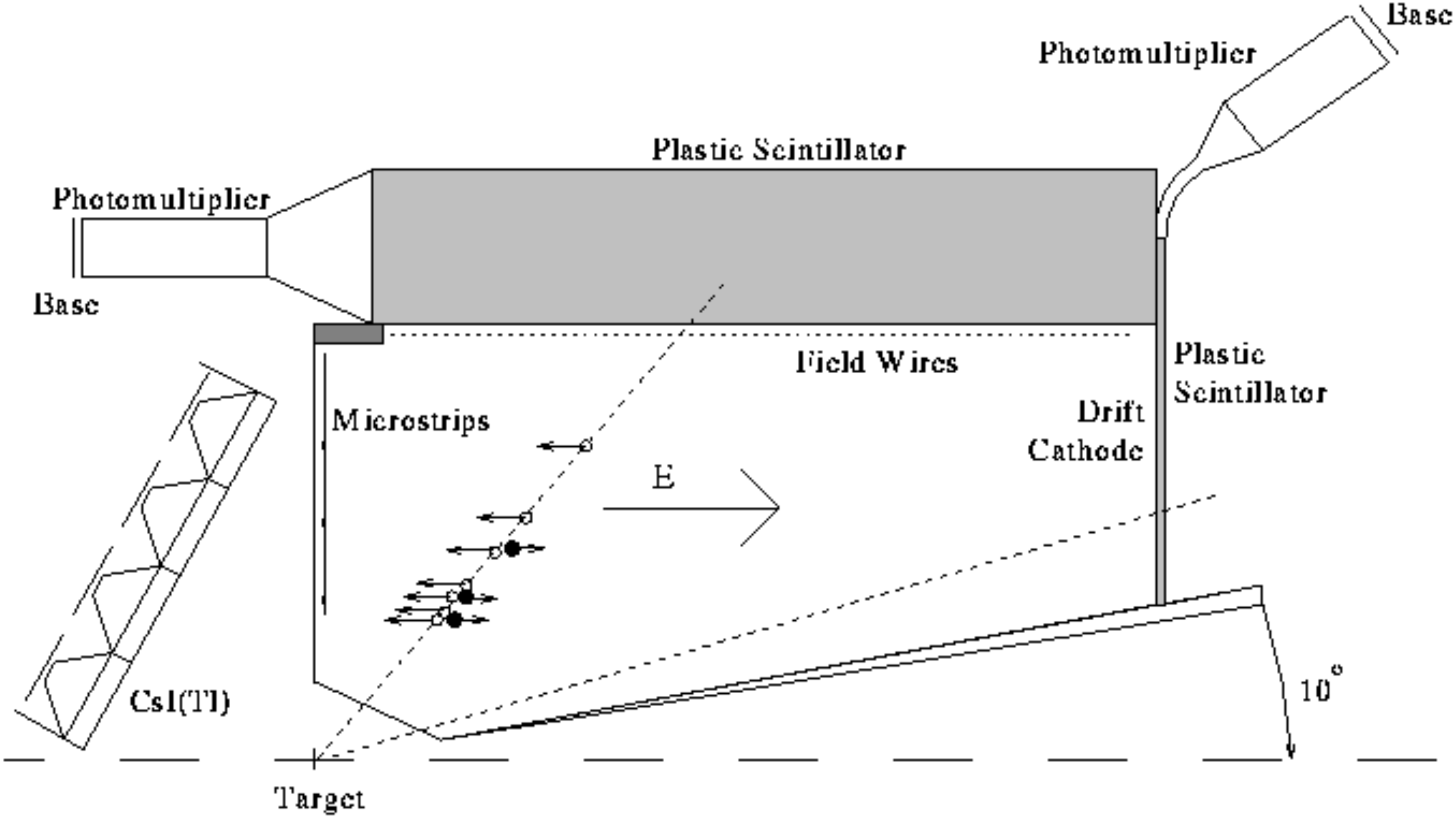
SURFnet 4.1

- **Remote data acquisition and analysis.**
  - » Allow a physicistsing videoconferencing tools.
  - » Adapt data acquisition system to exploit Quality of service opportunities of ATM.
  - » Test ATM on various protocol levels under varying loads using the DAQ.

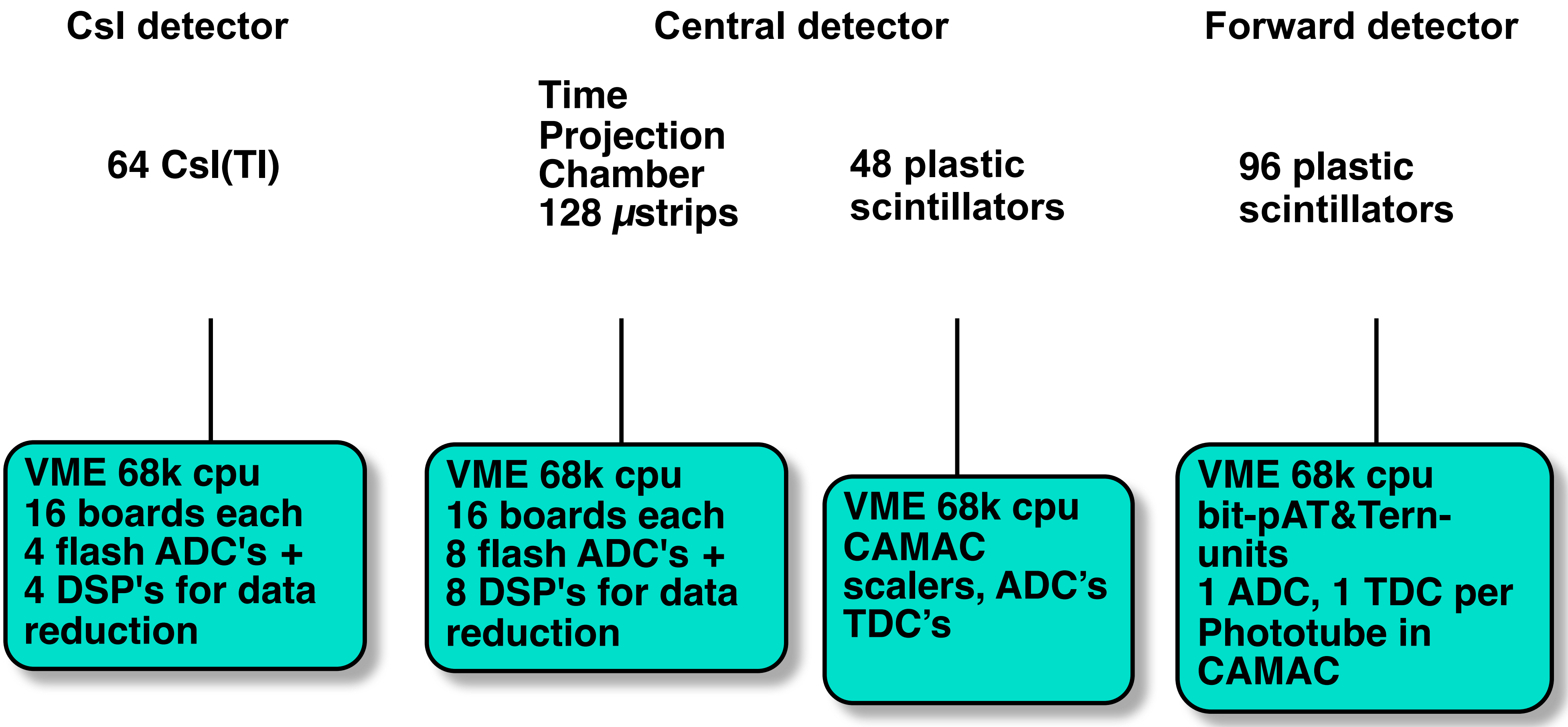
Huygens detector setup



Detector principle



Huygens Electronics



CsI detector

64 CsI(TI)

VME 68k cpu
16 boards each
4 flash ADC's +
4 DSP's for data reduction

Central detector

Time Projection Chamber
128 *μ*strips

VME 68k cpu
16 boards each
8 flash ADC's +
8 DSP's for data reduction

48 plastic scintillators

VME 68k cpu
CAMAC
scalers, ADC's
TDC's

Forward detector

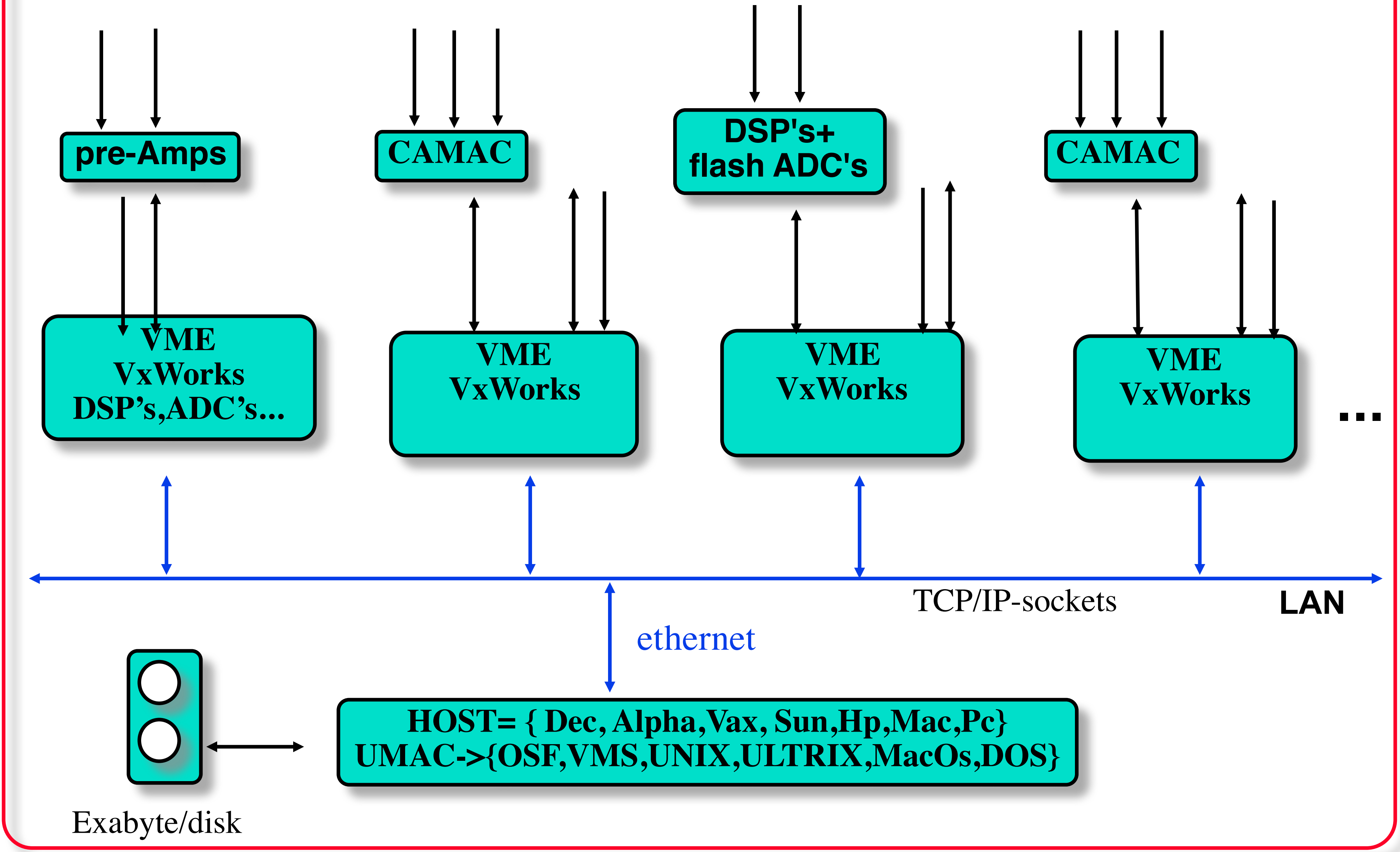
96 plastic scintillators

VME 68k cpu
bit-pAT&Tern-units
1 ADC, 1 TDC per Phototube in CAMAC

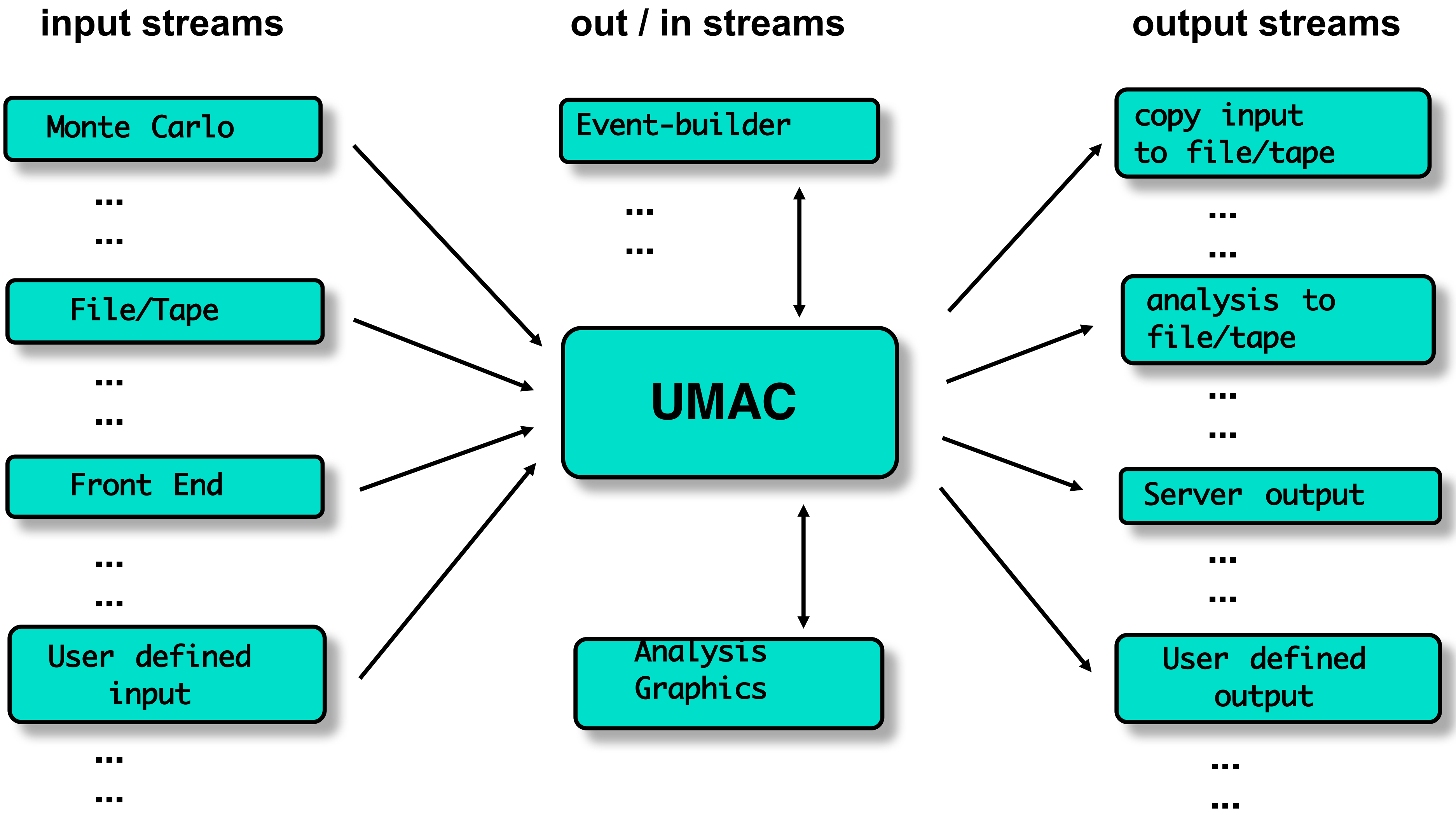
<Event size> 150 + 400 + 100 + 400

1050 bytes/event ✸ ± 500 events/sec ≈ 500 kByte/sec = 4 Mbit/sec

Hardware

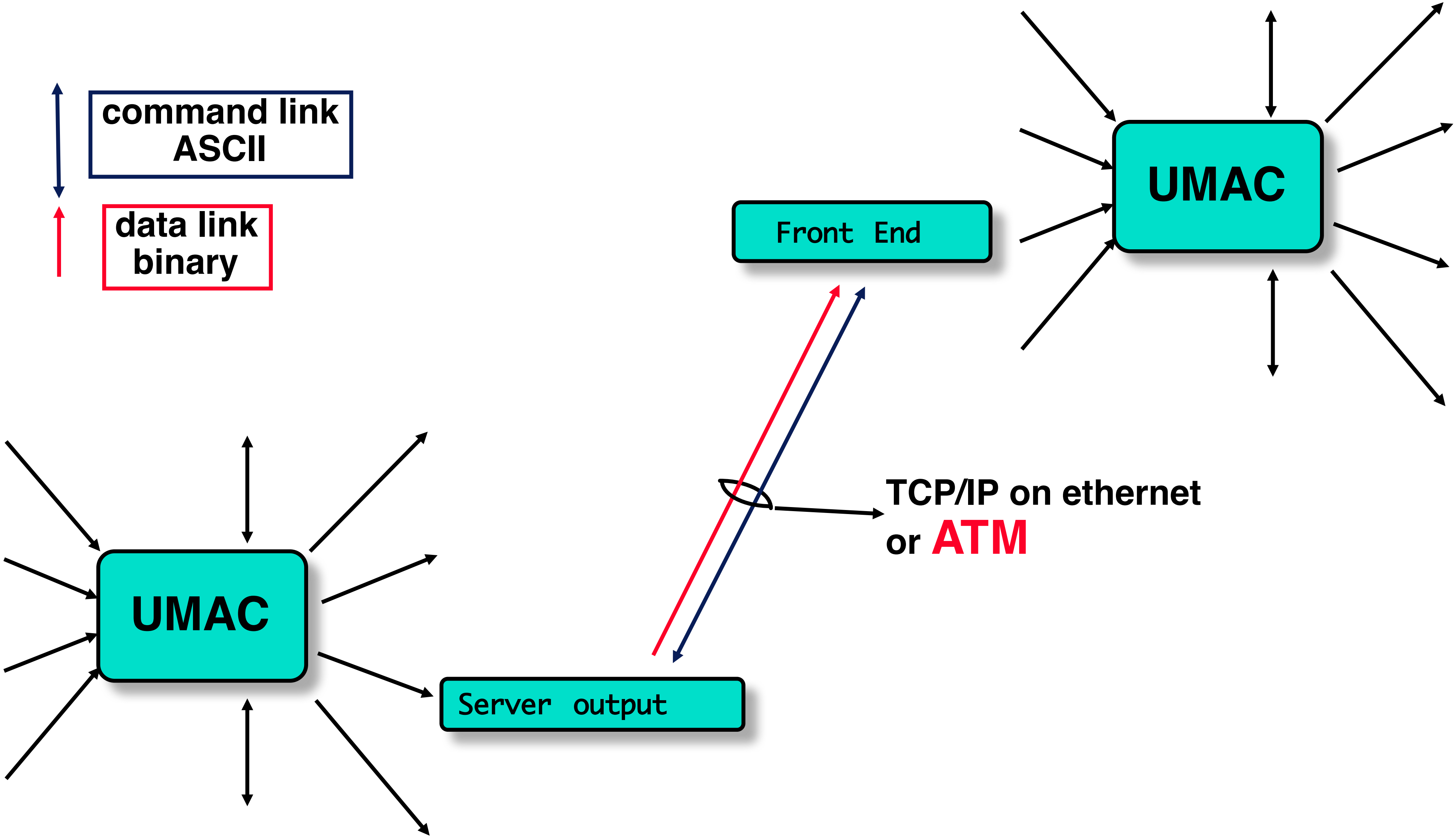

pre-Amps
CAMAC
DSP's+
flash ADC's
CAMAC
VME
VxWorks
DSP’s,ADC’s...
VME
VxWorks
VME
VxWorks
VME
VxWorks
...
TCP/IP-sockets
LAN
ethernet
HOST= { Dec, Alpha,Vax, Sun,Hp,Mac,Pc}
UMAC->{OSF,VMS,UNIX,ULTRIX,MacOs,DOS}
Exabyte/disk

Data Stream Management

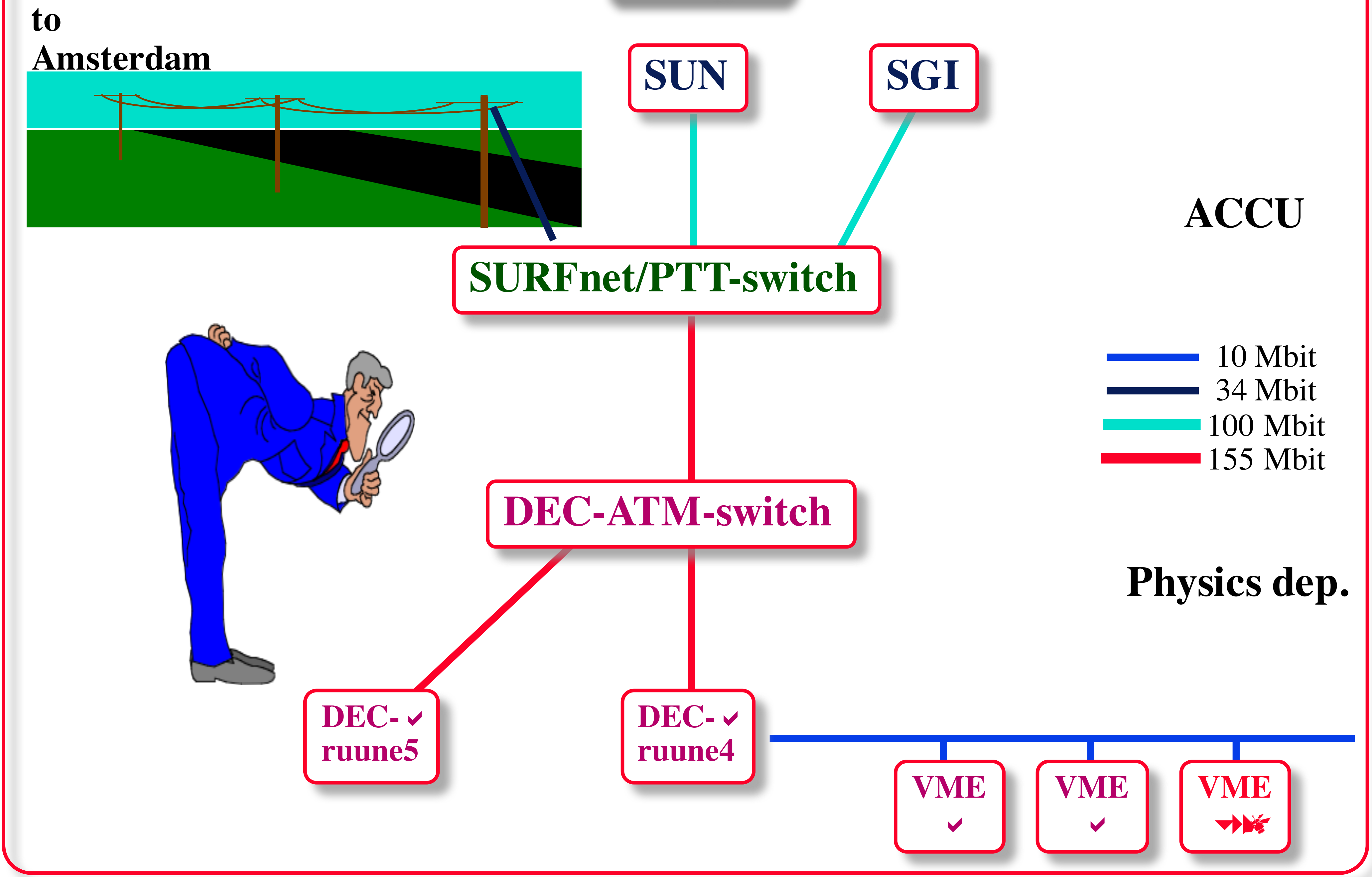


- 16 data streams (input+event-builders+output) simultaneously

Server-mode


command link
ASCII
data link
binary
Front End
UMAC
TCP/IP on ethernet
or ATM
UMAC
Server output

Setup-1



to Amsterdam

SUN

SGI

ACCU

SURFnet/PTT-switch

DEC-ATM-switch

Physics dep.

DEC- ✓
ruune5

DEC- ✓
ruune4

VME
✓

VME
✓

VME

🟦	10 Mbit
▬	34 Mbit
🟩	100 Mbit
🟥	155 Mbit

Test results 1

Data transfer rates from 4 VME 68k processors to a host on a silent ethernet:

Sender	Receiver	Kbytes/s	Mbits/s
4 VME 68k	Alpha 3000-400	908	7.3
4 VME 68k	Sun Sparc SLC	508	4.1

Data transfer rates from 1 VME processor to a host on a silent ethernet:

VME processor	Host Workstation	Kbytes/s	Mbits/s
Force 68k@25mc	Alpha 3000-400	550	4.4
AXP@160mc	Alpha 3000-400	1050	8.4

Data transfer rates between 2 UMAC programs:

Server Workstation	Client Workstation		Kbytes/s	Mbits/s
Alpha 3000-400	Alpha 3000-400	(ethernet)	1100	8.8
Alpha 3000-400	Alpha 3000-400	(ATM-encap-IP)	9400	75.2
Alpha Station-600	Alpha Station-600	(ATM-encap-IP)	16000	128

- •conclusion: if network is not busy, mean transfer rate good for our purposes.
- •Our aim is throughput, not response time (hard real time is handled by the FE’s).

Know Your Cell

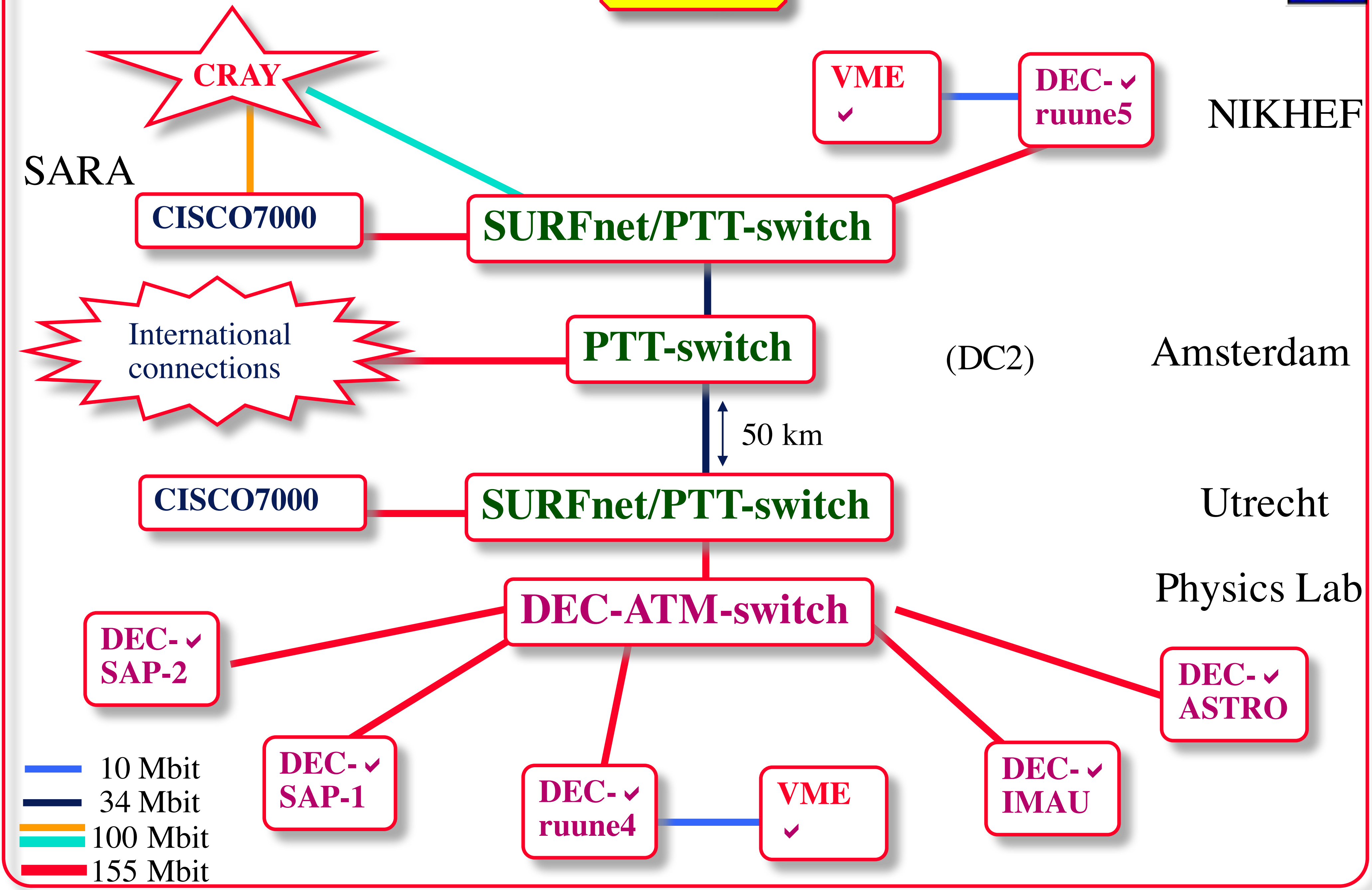
- SONET/SDH OC3 connection = 155.840 Mbit/s
- 53 bytes/cell * 8 bits/byte = 424 bits/cell
- 155.840 Mbit/s ÷ 424 bits/cell = 367547 cells/s
- 1 ÷ 367547 cells/s = 2.72 μsec/cell
- Lightspeed in fiber = c/n = 299792458 m/s ÷ 1.5 ≈ 2. 10^{8} m/s
- Length of a cell = 2.72 μsec/cell * 2. 10^{8} m/s = 544 m/cell
- Length of a byte = 544 m ÷ 53 bytes/cell = 10.26 m/byte
- 1 sec of traffic contains 155.840 Mbit/s ÷ 8 = 19.48 Mbyte
- useful with rtt: per millisec per megabit:
  1.10^{-3} msec * 1.10^{6} Mbit ÷ 8 bits/byte = 0.125 Kbytes
- Sliding window size for 20 msec on 6 megabit = 15 Kbytes for zero length MTU
- Costs: 100 kf/y/(31536000 s/y *367547 cells/s * 544 m/cell) =

  1.59 nano-cent per meter ATM cell !!

SURFnet 4.2

- **Package of 4 (inter)national proposals:**
  - » Magnetohydrodynamics of astrophysical and thermonuclear plasmas. Cray-Amsterdam and Minnesota in Minneapolis (USA).
  - » Institute for Marine and Atmospheric research Utrecht (Cray-Amsterdam).
  - » Remote data analysis and data base access. CERN, Geneva, heavy ION collaboration WA98.
  - » Remote database management and analysis L3. CERN, Geneva, LEP experiment L3.
- **Migration ethernet to ATM**
  - » Lan bridging and Lan Emulation
- **SVC's, QoS, ATM-API**
  - » Automatic connection management

Setup-2


CRAY
SARA
CISCO7000
SURFnet/PTT-switch
VME ✓
DEC- ✓ ruune5
NIKHEF
International connections
PTT-switch
(DC2)
Amsterdam
50 km
CISCO7000
SURFnet/PTT-switch
Utrecht
Physics Lab
DEC-ATM-switch
DEC- ✓ SAP-2
DEC- ✓ SAP-1
DEC- ✓ ruune4
VME ✓
DEC- ✓ IMAU
DEC- ✓ ASTRO
10 Mbit
34 Mbit
100 Mbit
155 Mbit

Test results 2-1

- **Cell loss can kill performance**

  1 out 10000 cells data error -> 20 % throughput. Problem was identified and corrected by PTT Telecom, at this moment no cell loss if supplied bandwidth does not exceed

- **Round trip times e4 - e5 < 1 ms, light speed in 100 km fiber 📻600 *μ*s**

- **Data transfer rates between two UMAC programs.**

ATM PVC setting	Throughput
8/12 Mbit/s	920 Kbytes/s = 7.4 Mbit/s
15/20 Mbit/s	1518 Kbytes/s = 12.1 Mbit/s

- **Video conferencing**

  nv and sd (public domain) used.

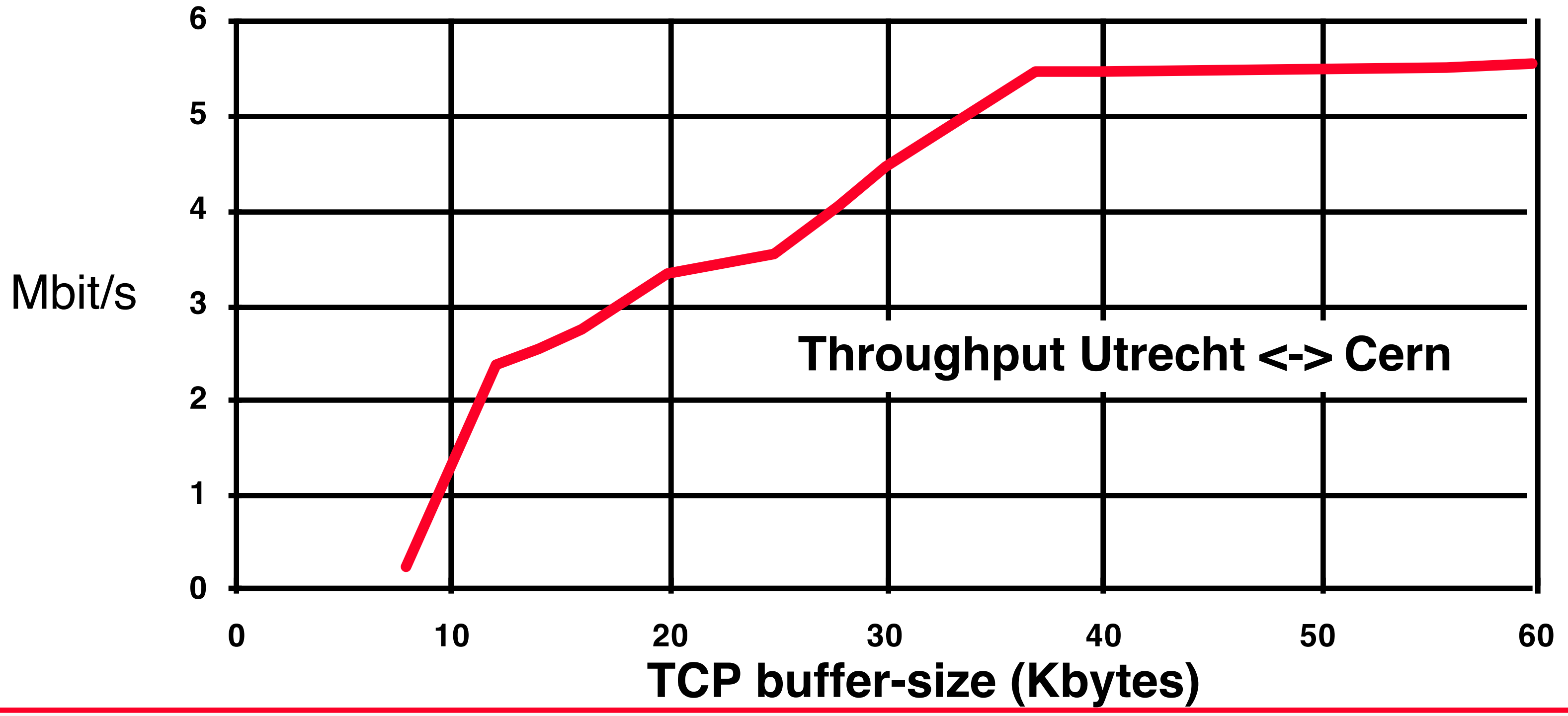
» nv:	ftp.parc.xerox.com
» sd,vat,wb,vic:	ftp.ee.lbl.gov

- **NFS**

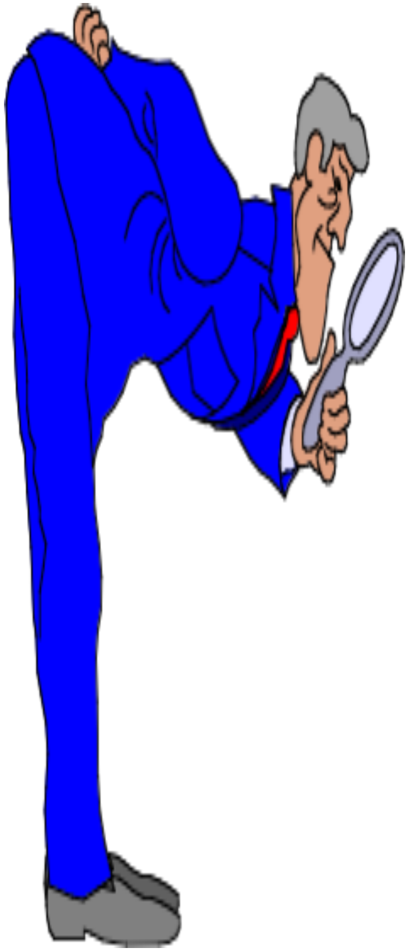
  NFS performance on ethernet ± 500 Kbytes/s, on ATM only 600 Kbytes/s -> NFS is RPC based which is the bottleneck.

Test results 2-2

- ATM connection is stable -> Eastern: in 4 days 360 GByte transfer on 8 Mbit
- CERN connection up and running for a few weeks since may 19th 1995
  - –round trip time 20 ms ≈ 4000 km fiber
  - –throughput UMAC: 5.5 Mbit on a 6 Mbit connection
  - –long fat network syndrome when tcp-buffer size below 36 Kbytes
  - –Throughput < TCP buffer size / round trip time

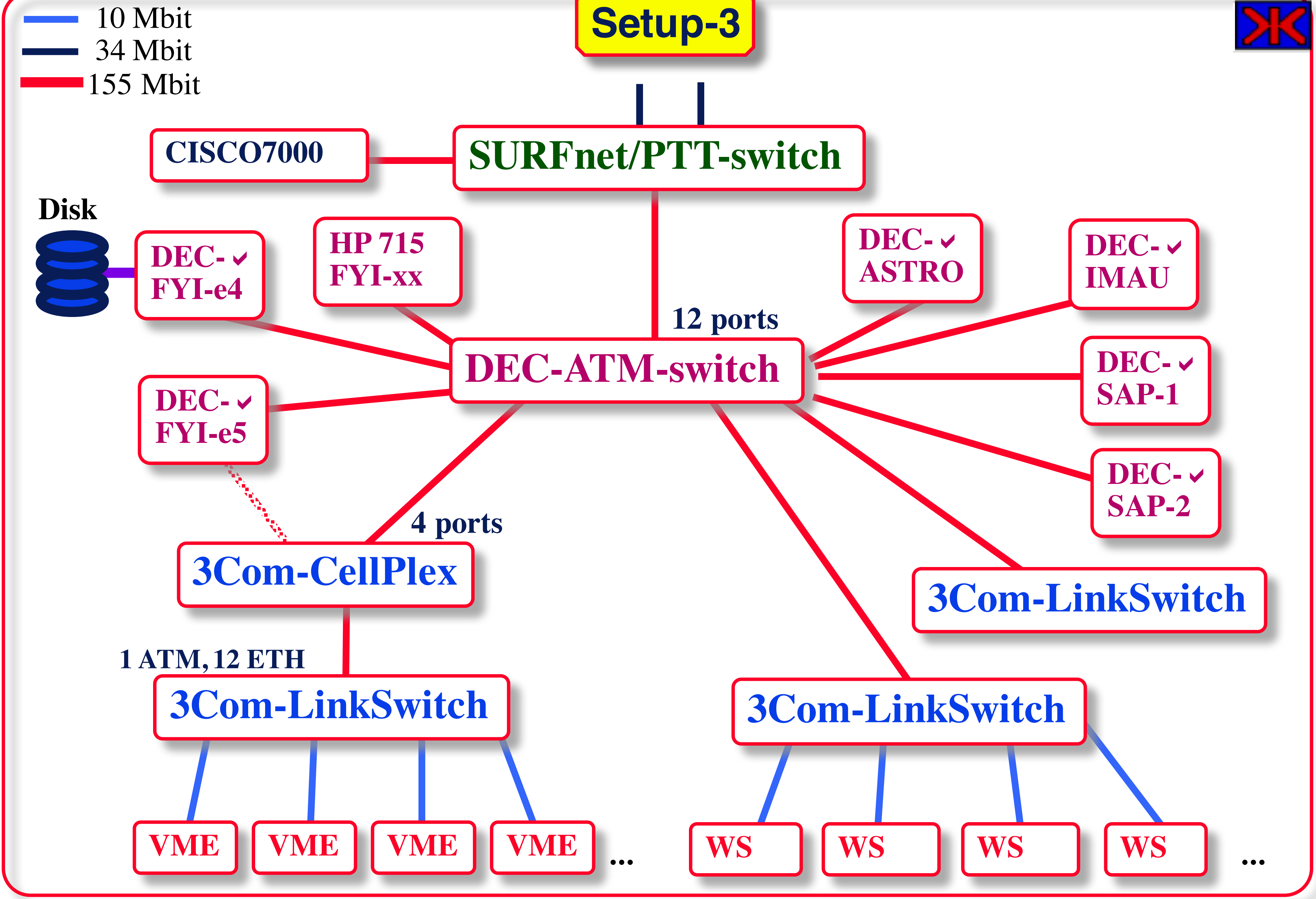


SURFnet 4.3



- REMOT
  - Remote Experiment control
- Migration ethernet to ATM
  - LAN bridging and LAN Emulation
- SVC’s, QoS, ATM-API
  - Automatic connection management

Setup-3



10 Mbit

34 Mbit

155 Mbit

CISCO7000

SURFnet/PTT-switch

Disk

DEC- ✓
FYI-e4

HP 715
FYI-xx

DEC- ✓
ASTRO

DEC- ✓
IMAU

12 ports

DEC-ATM-switch

DEC- ✓
SAP-1

DEC- ✓
FYI-e5

DEC- ✓
SAP-2

4 ports

3Com-CellPlex

3Com-LinkSwitch

1 ATM, 12 ETH

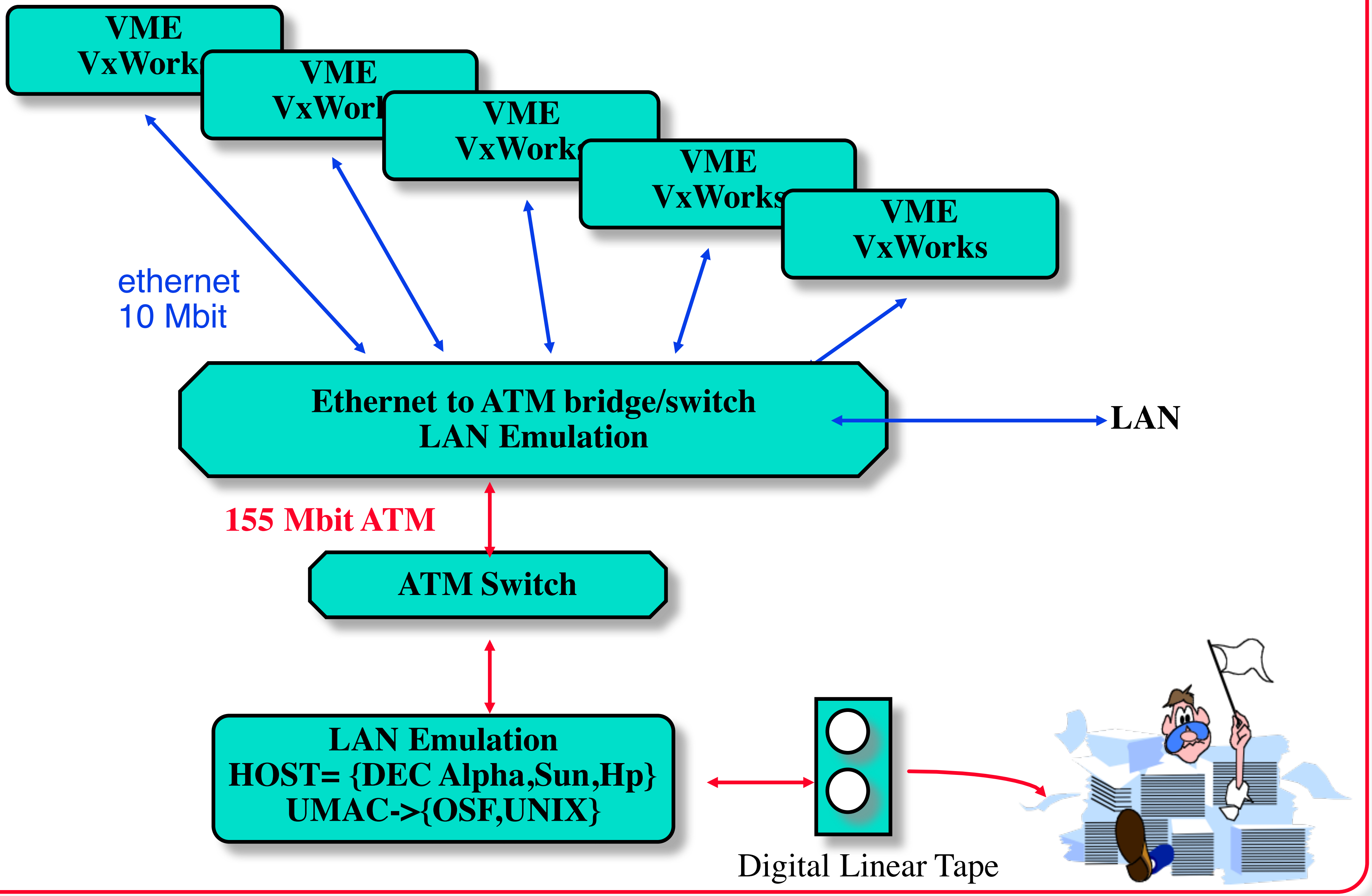
3Com-LinkSwitch

3Com-LinkSwitch

VME VME VME VME ...

WS WS WS WS ...

Lan Emulation based



Test results 3-1

- SVC's
  - –started to use ABR-SVC's UNI 3.0, NSAP addresses locally in September 1995
  - –after some beta testing now stable
  - –national and international connections still use CBR-PVC's
  - –participated in SVC tests at site of AT&T in Hilversum
- QoS
  - –participated in august 1995 and april 1996 in VBR tests, traffic shaping and CLP understood
- LANE
  - –first tests on 3Com Linkswitches in November 1995
  - –work together with Digital switch, LES, BUS and SVC's
- Data transfer rates with LANE:

Sender	Receiver		Kbytes/s	Mbits/s
4 VME 68k + Sun + Pc	Alpha 3000-400	(Linkswitch)	3500	28
1 VME + 8 Workstations	Alpha Station-600	(Linkswitch)	6888	55
Alpha Station-600	Alpha Station-600	(ATM-LANE)	6250	50

Raw LANE Results

```
umac>req
 Request: date and time  : 10-06-96  20:28:36
 umac program status     : running,     runnumber :     13
     serial measurement : disabled
 elapsed time,  this run :    181.89     sec, total:   290.51     sec.
 ==========================================================================
 ##   stream     #events  #analysed   #skipped  events/s    #MByte    kByte/s
 -------------------------------------------------------------------------
  0 Analysis=          0          0          0  0.000E+00  0.000E+00  0.000E+00
  1 fyscb::a<    942840          0          0  5.227E+03   184.      1.045E+03
  2 fysaz::a<    934335          0          0  5.206E+03   182.      1.041E+03
  3 fysav::a<    588195          0          0  3.479E+03   115.      696.
  4 fysau::a<    608310          0          0  3.972E+03   119.      794.
  5 ruunya::<    914490          0          0  5.161E+03   179.      1.032E+03
  6 fyscp::a<    371250          0          0  2.209E+03   72.5      442.
  7 ruund0::<    700650          0          0  4.103E+03   137.      821.
  8 ruuny5::<    403110          0          0  2.310E+03   78.7      462.
  9 vme05::a<     94338          0          0   555.       92.1      555.
 -------------------------------------------------------------------------
    total in<   5557518          0          0  3.222E+04  1.159E+03  6.888E+03
 ==========================================================================
umac>
```

```
[ruunf7][ROOT]{local/sbin}>./top

load averages:   1.16,  1.20,  1.11                            20:39:35
38 processes:  2 running, 8 sleeping, 28 idle
Cpu states: 22.1% user,  0.0% nice, 62.4% system, 15.4% idle
Memory: Real: 20M/121M act/tot  Virtual: 14M/151M use/tot  Free: 84M
```

Results-3-2

- Done:
  - 3Com LAN Emulation works (one vendor)
  - 3Com linkswitches can use LES+BUS in Digital switch
  - 2 logical groups, one using LES+BUS Digital , other LES+BUS 3Com works
  - SVC's DEC- ✓ <-> 3Com CellPlex work
  - No Flow Control via 3Com --> Cell Loss + TCP/IP - connections lost
  - Bridging and traffic separation confirmed
- At this moment
  - LAN Emulation Client software for Digital- ✓ and Hp
- To do:
  - Connection to CISCO router
  - Broadcast overload
  - Congestion tests
  - Wide area connections with LAN Emulation

Future

- QoS, low level ATM connection setup --> API
- Lan Emulation
- Switched Virtual Circuits
- Virtual Control Room research (REMOT)
    - specification of set-ups
    - procedures for access
    - negotiation Remote high volume data access and distributed analysi
- Distributed computing

Acknowledgments

- This work is supported by
  - SURFnet bv
    - » proposals CD1 and CD29
  - Digital Equipment
    - » European External Research Proposal
  - 3Com
    - » ATM LANE Project R.U.U.