# SNR-2006: Very Dynamic LightPath Applications in DAS3 & StarPlane.


**Cees de Laat**

**University of Amsterdam**

SARA
TI
TNO
NCF


**SURFnet**

**BSIK**

**EU**

SURFnet

Advanced School for Computing and Imaging

# History - 1

DAS = Distributed ASCI Supercomputer

- Project DAS-1 started in 1997 by Andrew Tanenbaum
- To prove distributed clusters were as effective as super...
- 4-5 clusters connected via high speed links
  - DAS-1 -> 6 Mbit/s full mesh ATM
  - DAS-2 -> Gbit/s L3
  - DAS-3 -> StarPlane

- DAS-1 ran BSD, changed to Linux (Andrew... :-)
- DAS-1 and 2 uniform architecture, not so in DAS-3
- Over 200 users, 25 Ph.D. theses
- http://www.cs.vu.nl/das/

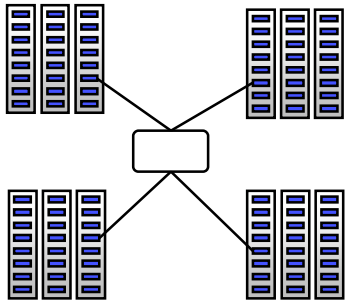
Advanced School for Computing and Imaging

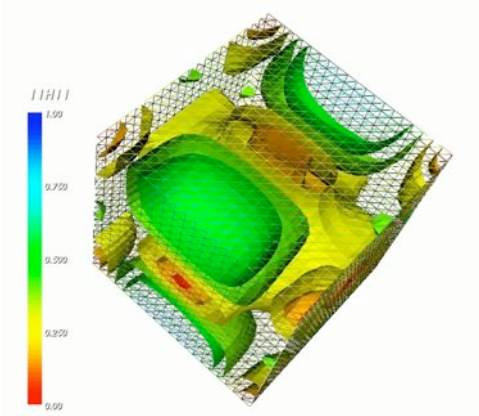
# Examples cluster computing

- **Communication protocols for Myrinet**
- **Parallel languages (Orca, Spar)**
- **Parallel applications**
  - PILE: Parallel image processing
  - HIRLAM: Weather forecasting
  - Solving Awari  (3500-year old game)
- **GRAPE: N-body simulation hardware**

# Distributed supercomputing on DAS

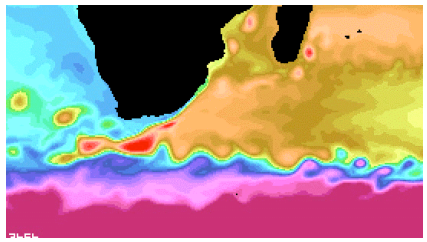
- **Parallel processing on multiple clusters**
- **Study non-trivially parallel applications**
- **Exploit hierarchical structure for locality optimizations**
  - latency hiding, message combining, etc.
- **Successful for many applications**
  - E.g. Jem3D in ProActive [F. Huet, SC'04]





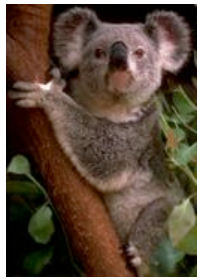
# Example projects

- **Albatross**
  - Optimize algorithms for wide area execution
- **MagPIe:**
  - MPI collective communication for WANs
- **Manta: distributed supercomputing in Java**
- **Dynamite: MPI checkpointing & migration**
- **ProActive (INRIA)**
- **Co-allocation/scheduling in multi-clusters**
- **Ensflow**
  - Stochastic ocean flow model

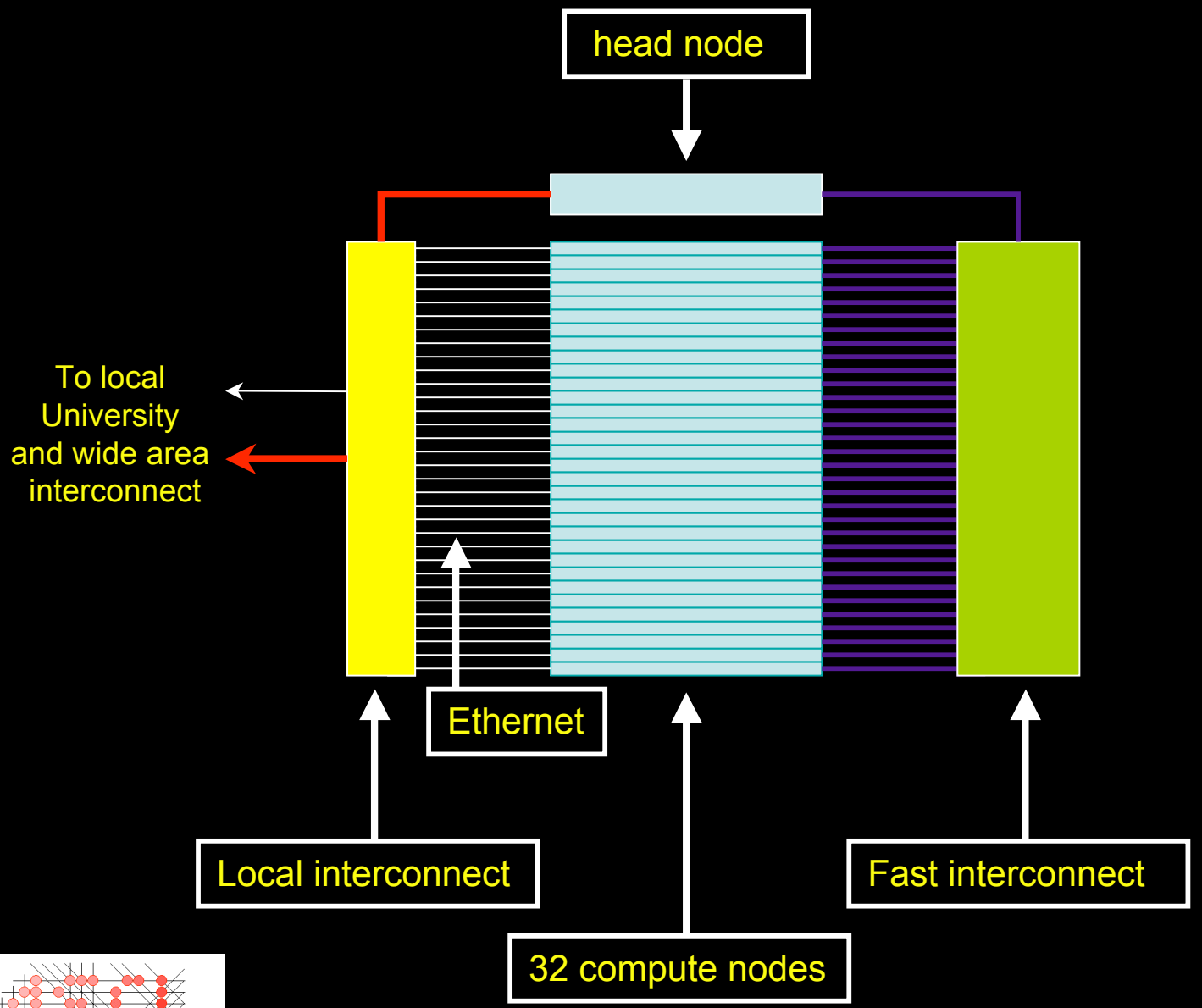



# Grid & P2P computing: using DAS-2 as part of larger heterogeneous grids

- Ibis: Java-centric grid computing
- Satin: divide-and-conquer on grids
- Zorilla: P2P distributed supercomputing
- KOALA: co-allocation of grid resources
- Globule: P2P system with adaptive replication
- CrossGrid: interactive simulation and visualization of a biomedical system



# DAS 1 - 2 Cluster

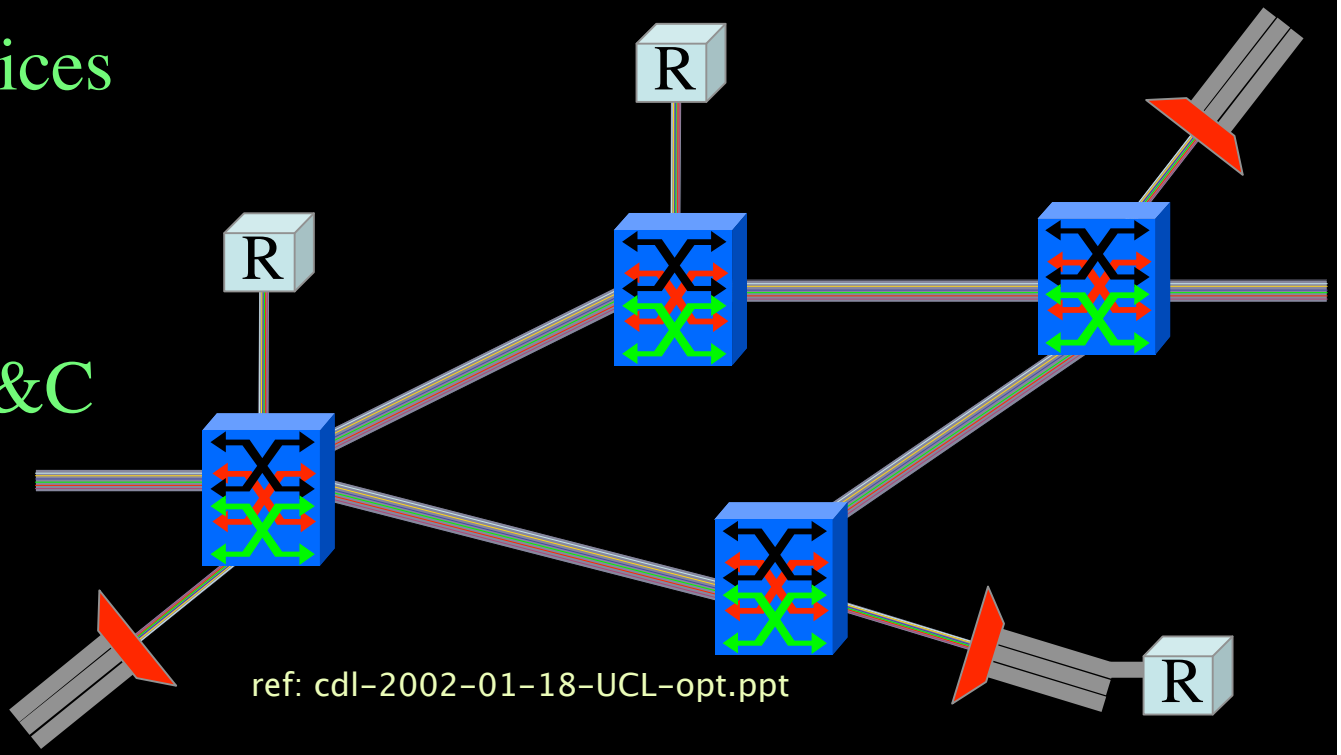



Advanced School for Computing and Imaging

# History - 2

SURFnet6 Architecture discussions 2001-2002

- photonic backbone
- (L2 and) L3 services
- NORTEL
- Static
- Summer 2004 K&C
- NWO-GLANCE
- StarPlane
- PHD-PD-SP



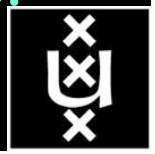

- Start 1-feb-06, Li Xu, Jan Philip Velders, Jason Maasen
  - Henri Bal, Paola Grosso, Herbert Bos, CdL, SN-folks.

SURFnet

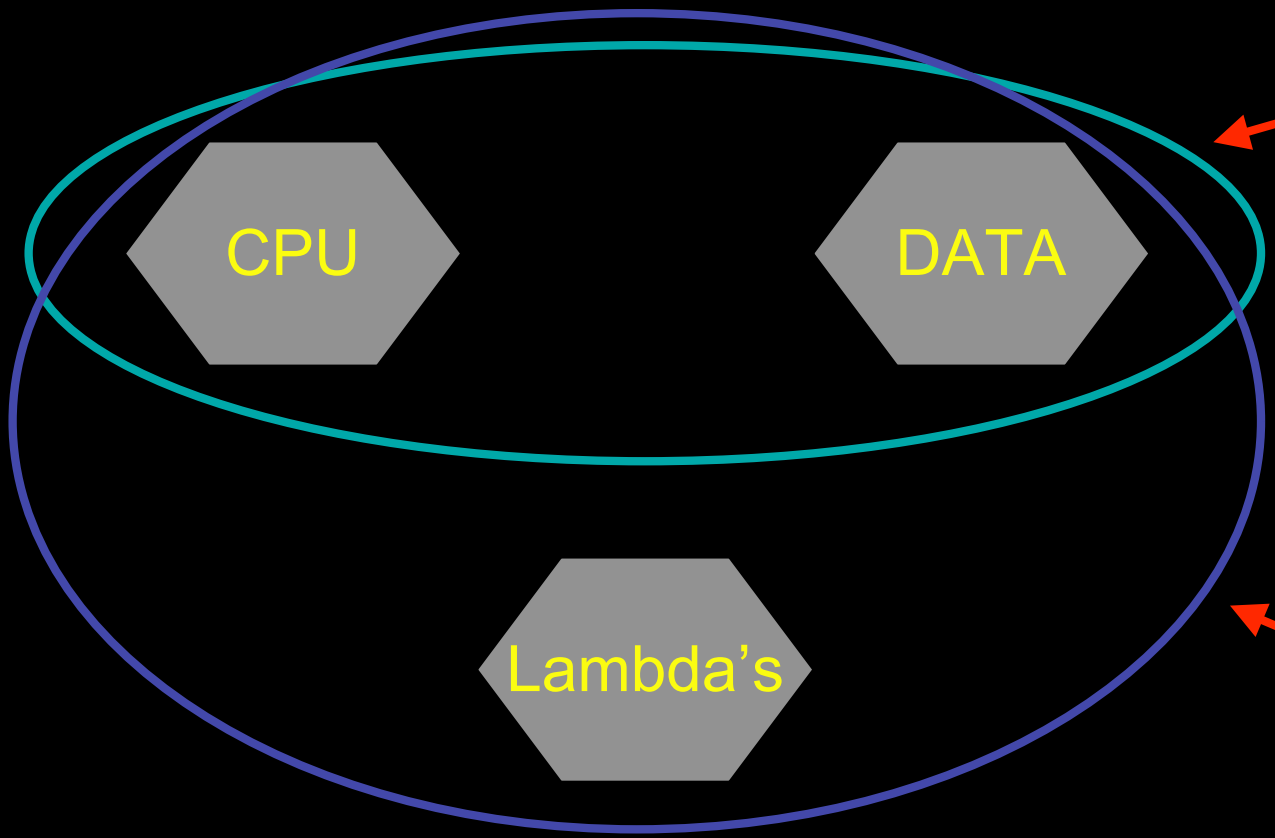
# StarPlane Approach

(PG)

- StarPlane is a NWO funded project with major contributions from SURFnet and NORTEL.
- The vision is to allow part of the photonic network infrastructure of SURFnet6 to be manipulated by Grid applications to optimize the performance of specific e-Science applications.
- StarPlane will use the physical infrastructure provided by SURFnet6 and the distributed supercomputer DAS-3.
- The novelty: to give flexibility directly to the applications by allowing them to choose the logical topology in real time, ultimately with subsecond lambda switching times.



# GRID-Colocation problem space




Extensively under research

New!

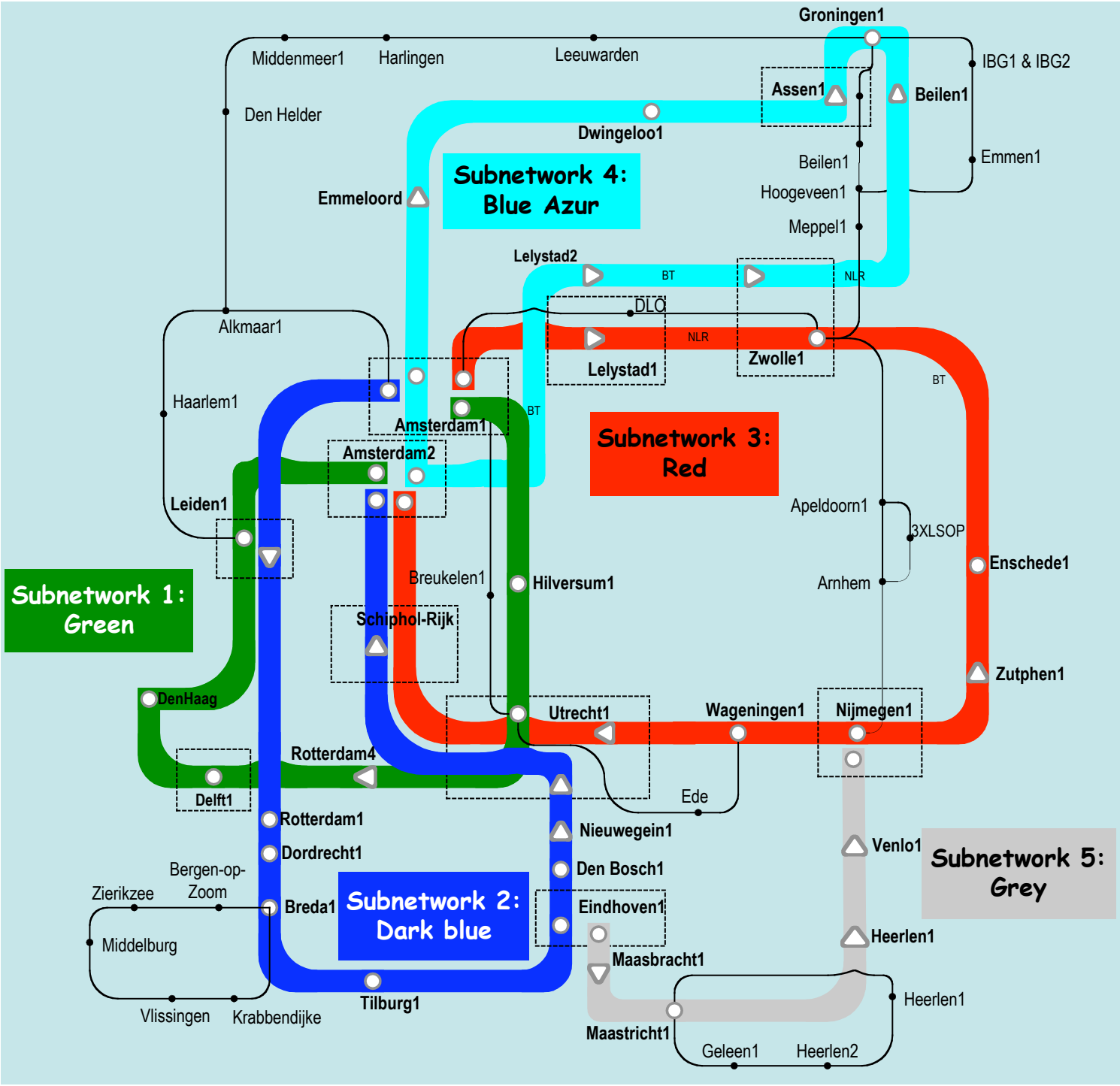

In The Netherlands SURFnet connects between 180:

- universities;
- academic hospitals;
- most polytechnics;
- research centers.

with a user base of ~750K users

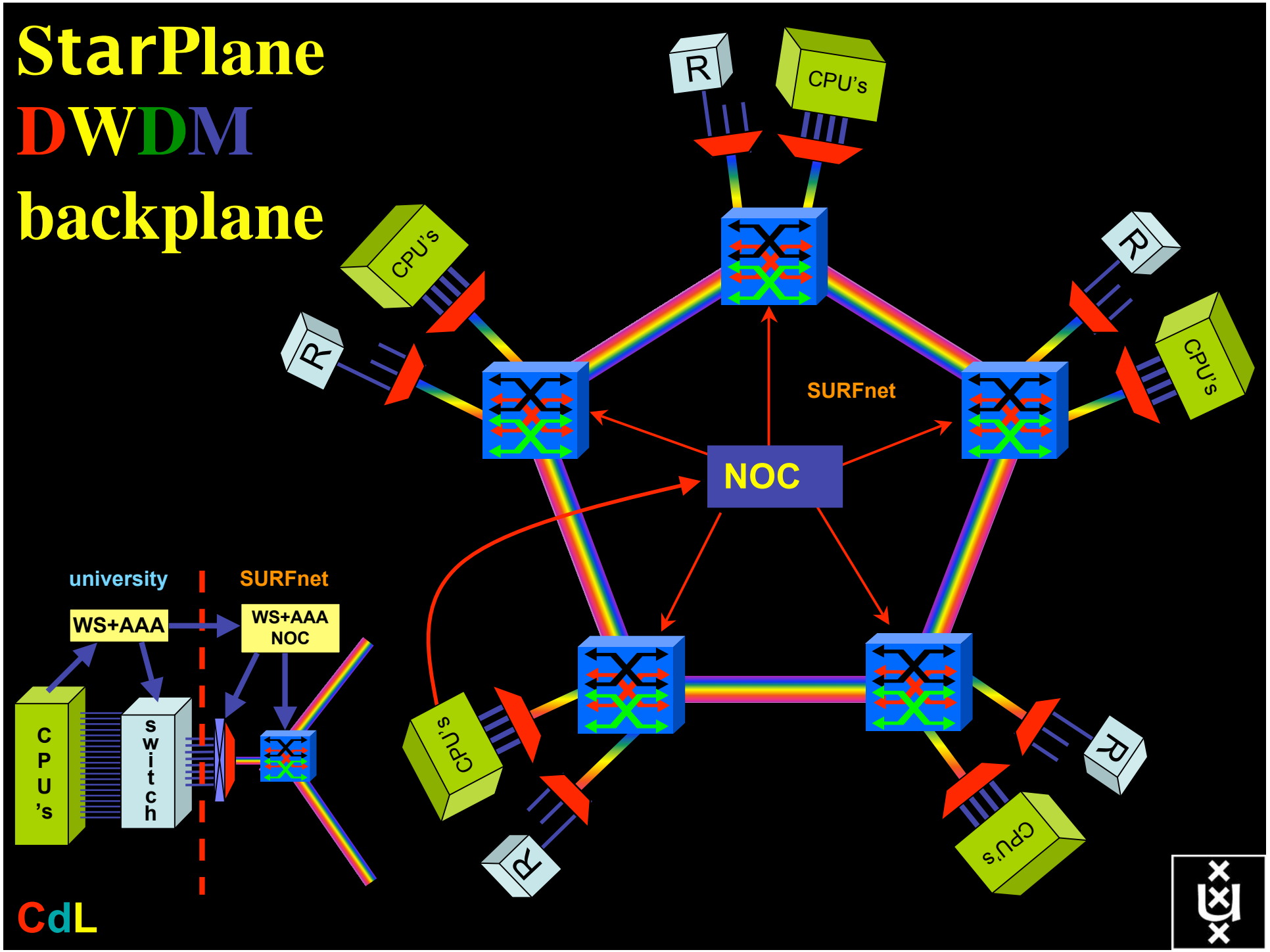
> 6000 km

comparable to railway system



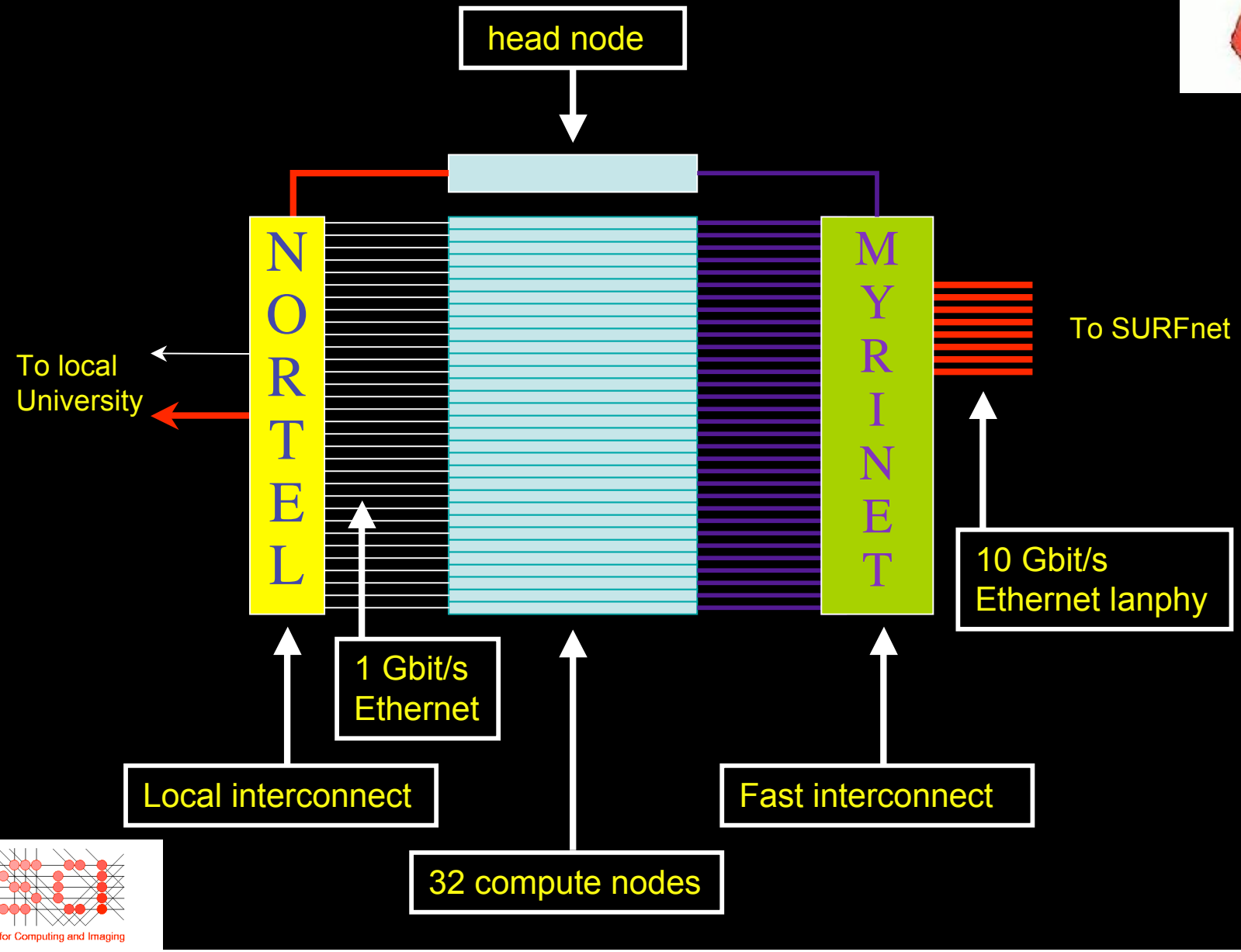

# Common Photonic Layer (CPL) in SURFnet6

supports up to
72 Lambda's of
10 G each
40 G soon.


StarPlane
DWDM
backplane
R
CPU's
SURFnet
NOC
university
SURFnet
WS+AAA
WS+AAA
NOC
CPU's
switch
CdL

# DAS-3 Cluster Tender

http://www.clustervision.com/pr_das3_uk.html

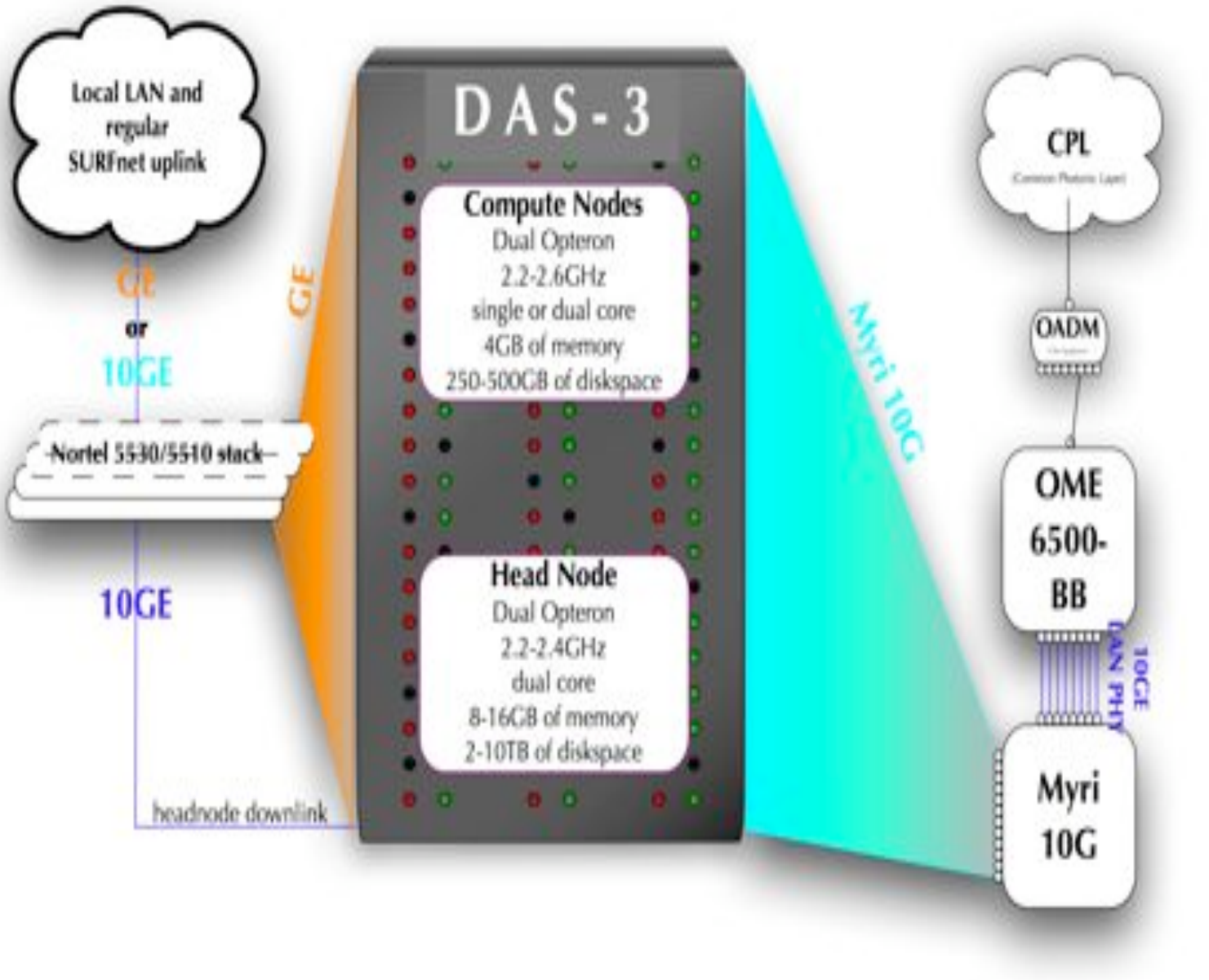



Advanced School for Computing and Imaging

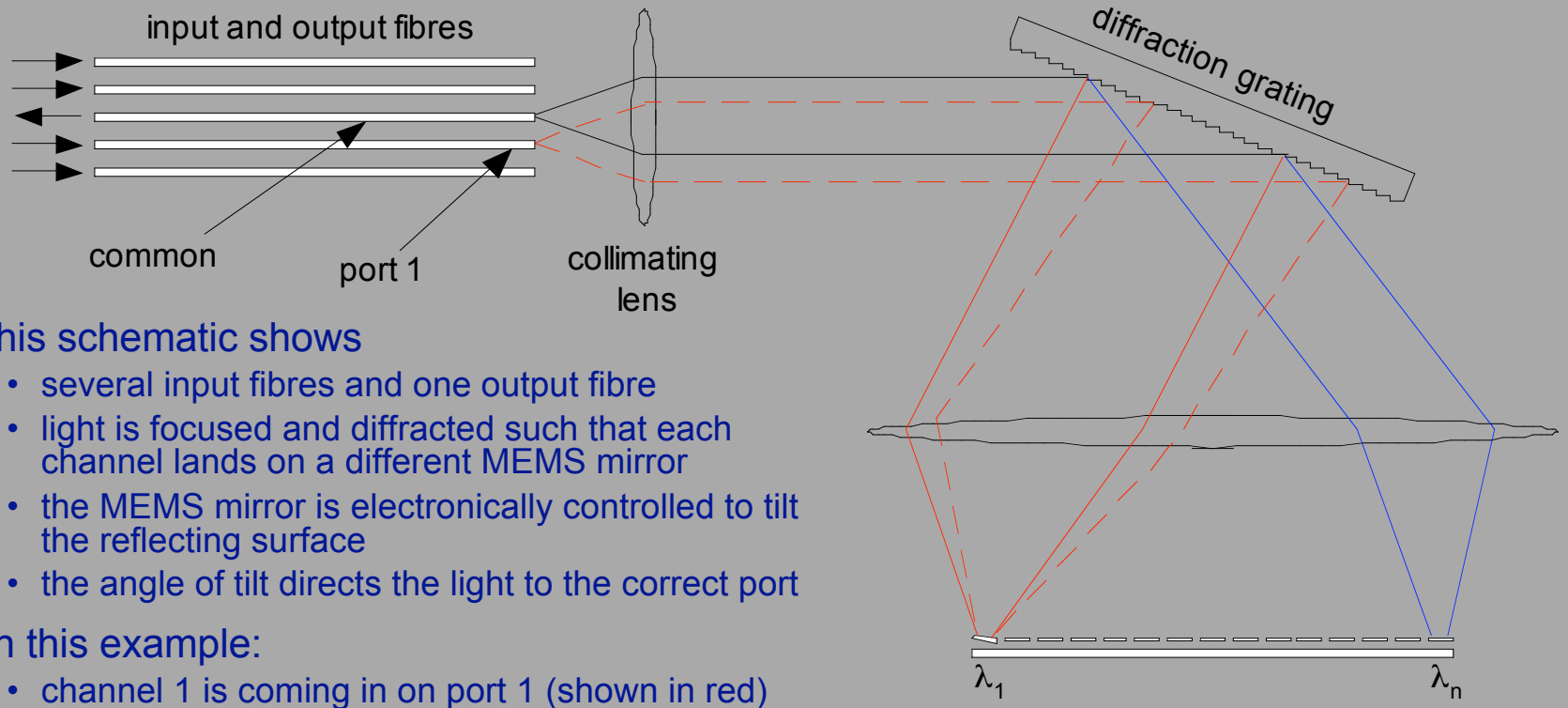
# Heterogeneous clusters

| | LU | TUD | UvA | UvA-MN | VU | TOTALS |
|---|---|---|---|---|---|---|
| **Head** | | | | | | |
| * storage | 10TB | 5TB | 2TB | 2TB | 10TB | 29TB |
| * CPU | 2x2.4GHz DC | 2x2.4GHz DC | 2x2.2GHz DC | 2x2.2GHz DC | 2x2.4GHz DC | |
| * memory | 16GB | 16GB | 8GB | 16GB | 8GB | 64GB |
| * Myri 10G | 1 | | 1 | 1 | 1 | |
| * 10GE | 1 | 1 | 1 | 1 | 1 | |
| | | | | | | |
| **Compute** | 32 | 68 | 40 (1) | 46 | 85 | 271 |
| * storage | 400GB | 250GB | 250GB | 2x250GB | 250GB | 84 TB |
| * CPU | 2x2.6GHz | 2x2.4GHz | 2x2.2GHz DC | 2x2.4GHz | 2x2.4GHz DC | 1.9 THz |
| * memory | 4GB | 4GB | 4GB | 4GB | 4GB | 1048 GB |
| * Myri 10G | 1 | | 1 | 1 | 1 | |
| | | | | | | |
| **Myrinet** | | | | | | |
| * 10G ports | 33 (7) | | 41 | 47 | 86 (2) | |
| * 10GE ports | 8 | | 8 | 8 | 8 | 320 Gb/s |
| | | | | | | |
| **Nortel** | | | | | | |
| * 1GE ports | 32 (16) | 136 (8) | 40 (8) | 46 (2) | 85 (11) | 339 Gb/s |
| * 10GE ports | 1 (1) | 9 (3) | 2 | 2 | 1 (1) | |

# Photonics


Local LAN and
regular
SURFnet uplink
GE
or
10GE
Nortel 5530/5510 stack
10GE
headnode downlink
GE
DAS-3
Compute Nodes
Dual Opteron
2.2-2.6GHz
single or dual core
4GB of memory
250-500GB of diskspace
Head Node
Dual Opteron
2.2-2.4GHz
dual core
8-16GB of memory
2-10TB of diskspace
Myri 10G
CPL
OADM
OME
6500-
BB
10GE
LAN PHY
Myri
10G

# Module Operation




> this schematic shows
  - several input fibres and one output fibre
  - light is focused and diffracted such that each channel lands on a different MEMS mirror
  - the MEMS mirror is electronically controlled to tilt the reflecting surface
  - the angle of tilt directs the light to the correct port

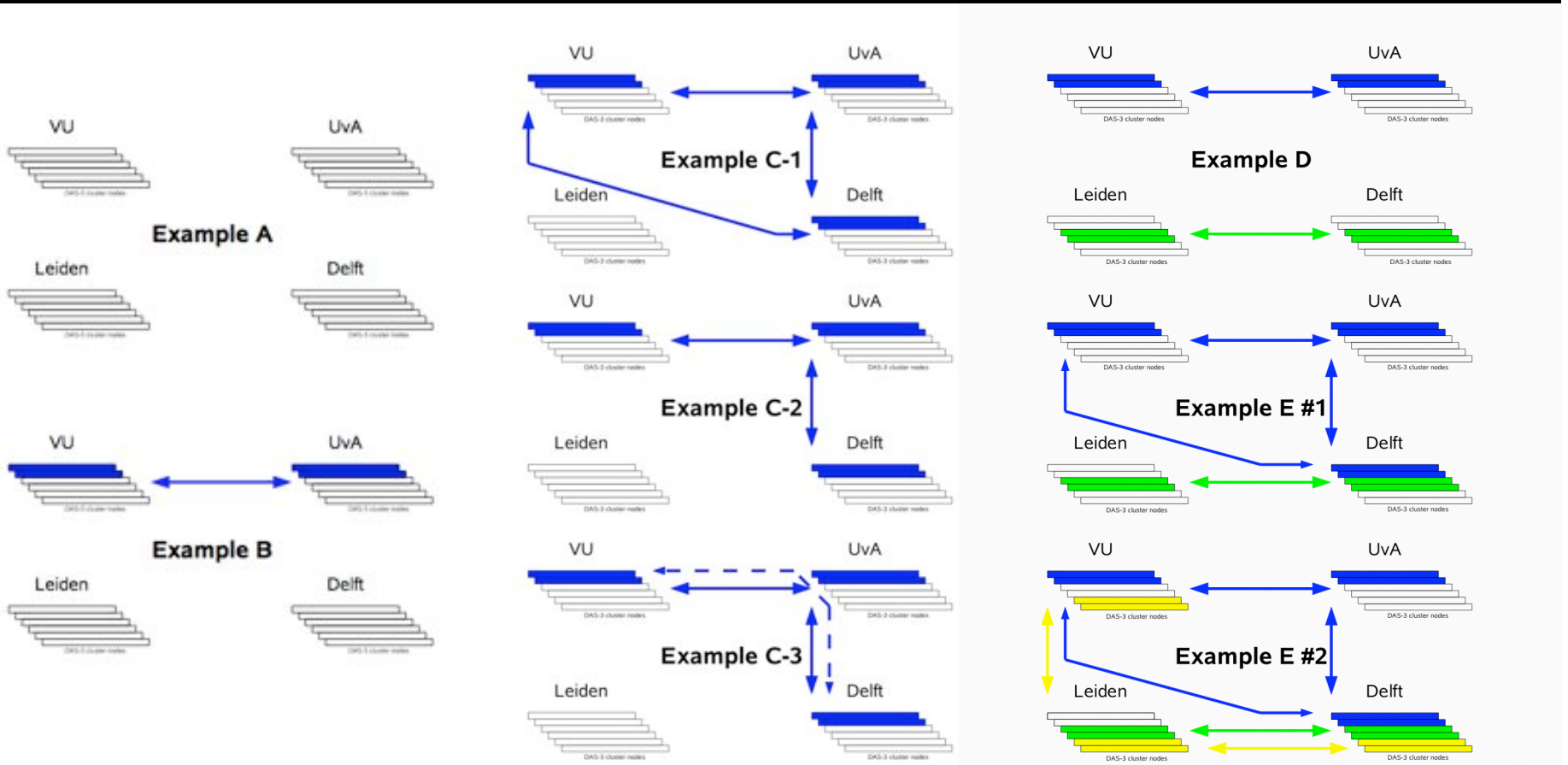
> in this example:
  - channel 1 is coming in on port 1 (shown in red)
  - when it hits the MEMS mirror the mirror is tilted to direct this channel from port 1 to the common
  - only port 1 satisfies this angle, therefore all other ports are blocked

ref Eric Bernier, NORTEL

# What makes StarPlane possible

- Wavelength Selective Switches
- Sandbox by confining StarPlane to a band
- Optimization of the controls to turn on/off a Lambda
- electronic Dynamically Compensating Optics (eDCO)
- traffic engineering

# Traffic engineering



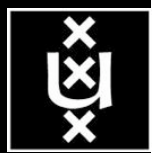
# What do we need

- vlan's
- trunking
- spanning tree modified?
- mac in mac?
- source routing modified
- Policy interfaces
- AAA interaction (EduRoam, Shibboleth)

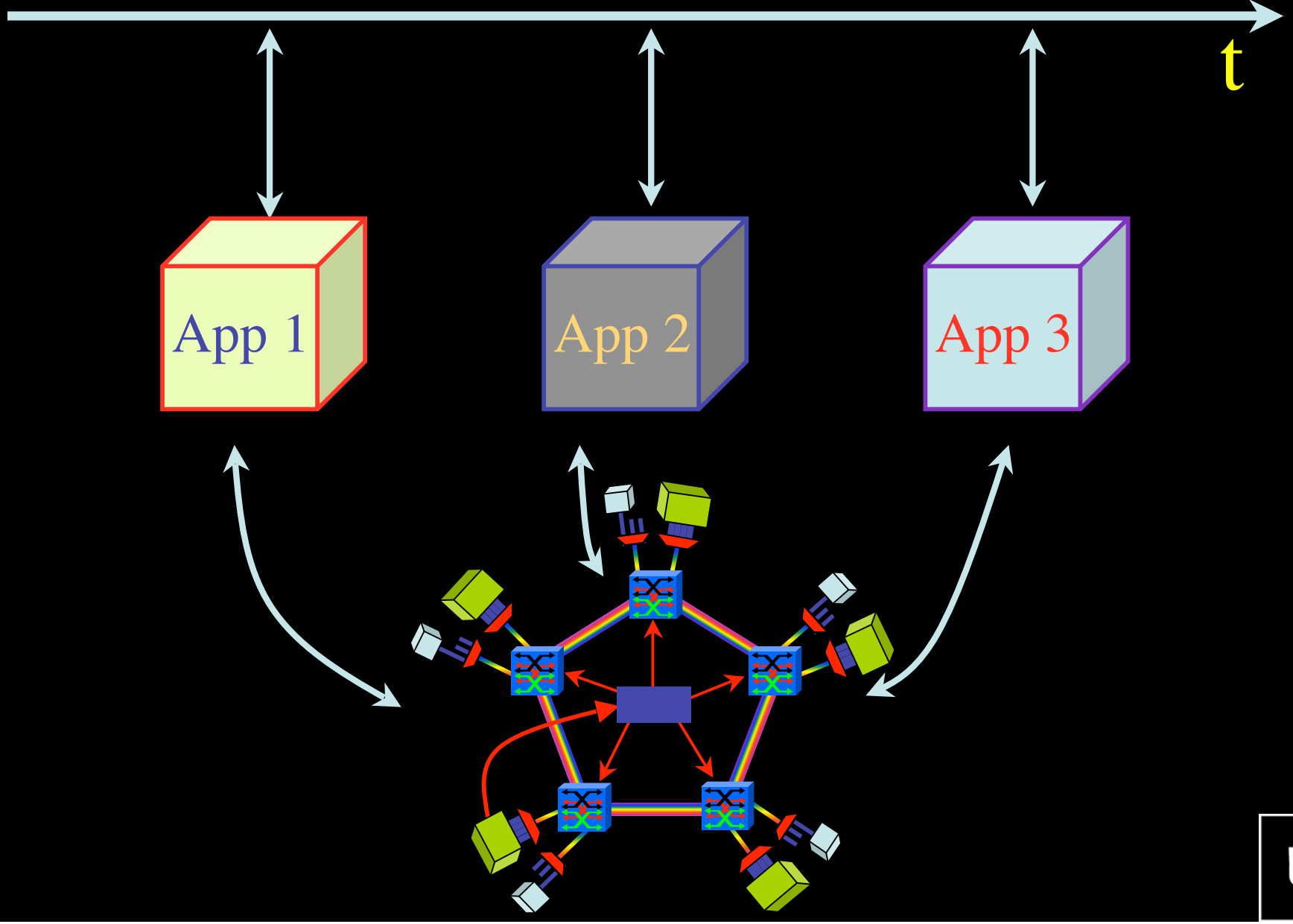
# StarPlane applications

- Large 'stand-alone' file transfers
  - User-driven file transfers
  - Nightly backups
  - Transfer of medical data files (MRI)
- Large file (speedier) Stage-in/Stage-out
  - MEG modeling (Magnetoencephalography)
  - Analysis of video data
- Application with static bandwidth requirements
  - Distributed game-tree search
  - Remote data access for analysis of video data
  - Remote visualization
- Applications with dynamic bandwidth requirements
  - Remote data access for MEG modeling
  - SCARI

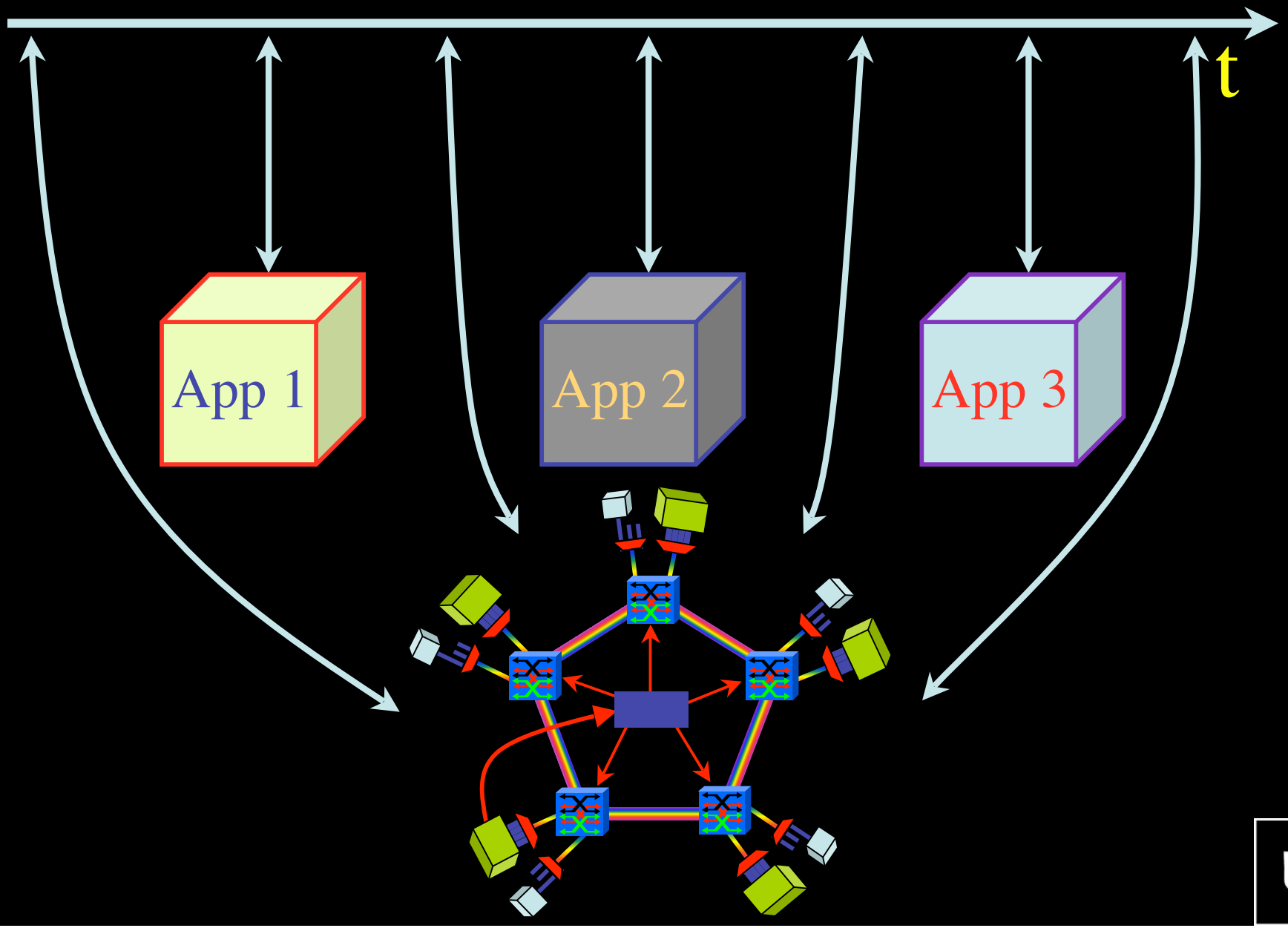




# Application - Network interaction



# Workflow based App.



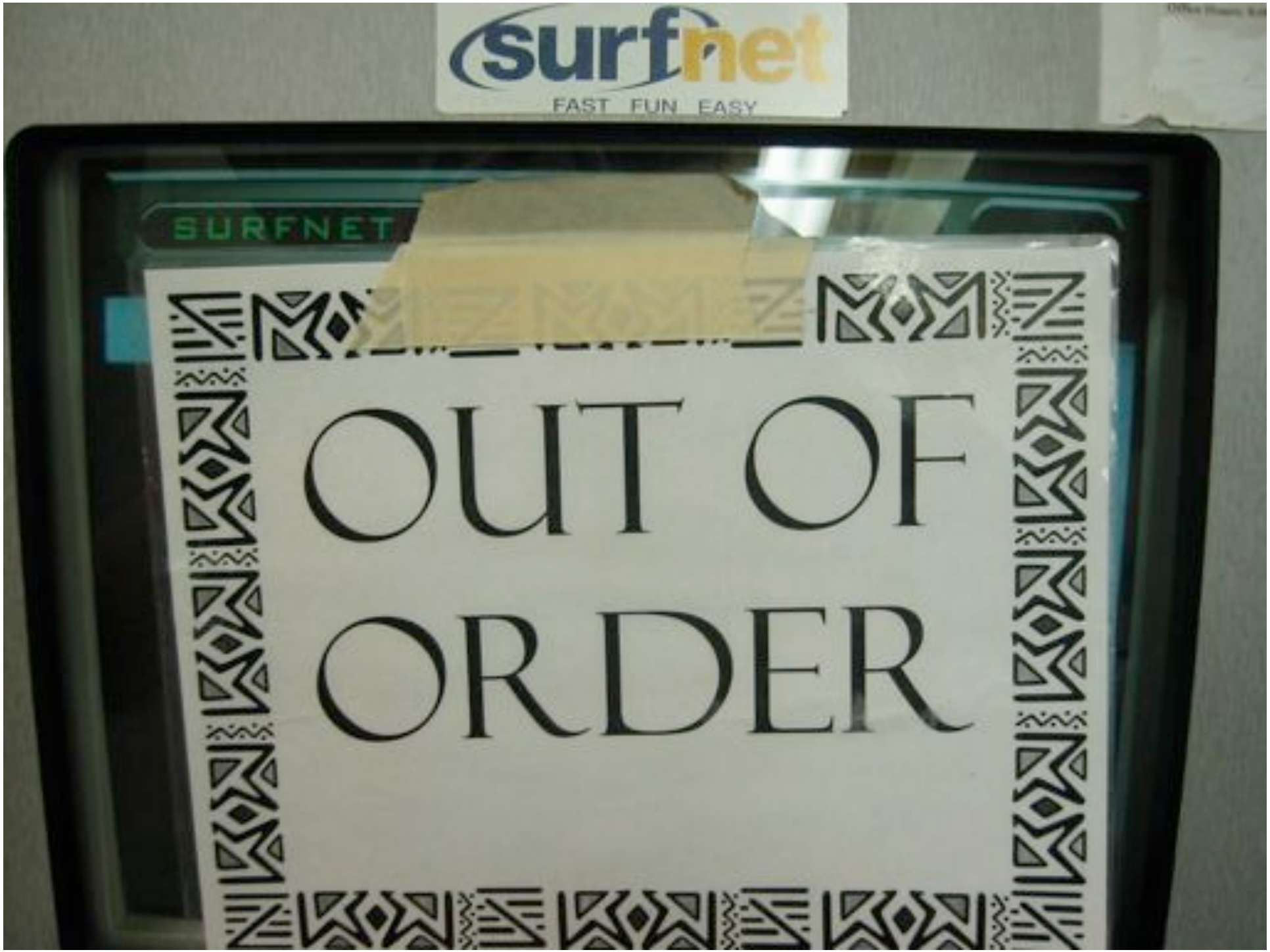
# Risks

what have we today

what to avoid


surfnet
FAST. FUN. EASY
SURFNET PREMIERE
HELP
net
Three Easy Steps :
Click the START button
Insert money...
$0.25 per minute...
Example :
$1 = 4 minutes
$5 = 20 minutes
No change is provided!
Surf the web!


surfnet
FAST FUN EASY
SURFNET PREMIERE
HELP
019
2
ROYAL MAIL
FRANKING
SERVICE
2nd
Paid
Check your
email
here!
Click the Start Button to begin


surfnet
FAST FUN EASY
SURFNET
OUT OF
ORDER

# Conclusions

- We try to go for fast (subsecond) Lambda setup and teardown, that is different from most other initiatives
- We need to work on GMPLS, SOA, webservices, RDF, supporting tools to make this happen
- We need to stress the current control loops and procedures to get there
- Workflow systems and/or applications need to become network aware.

# Questions ?


Credits: some slides from Paola Grosso or Henri Bal

sara
Computing & Networking Services

Advanced School for Computing and Imaging

SURFnet