

# CineGrid Networking

CG-2008

Cees de Laat

University of Amsterdam



# Contents

1. Use cases CineGrid & Networks
2. Formats - Numbers - Bits
3. Global Lambda Integrated Facility
4. A LightPath
5. Transport Protocol issues
6. End System Issues
7. Network Storage
8. Q/A



# CineGrid Mission

To build an interdisciplinary **community** that is focused on the **research, development, and demonstration** of **networked** collaborative tools to enable the production, **use** and **exchange** of very-high-quality digital media over **photonic networks**.

<http://www.cinegrid.org/>



# Keio/Calit2 Collaboration: Trans-Pacific 4K Teleconference

Like High-Def? Here Comes the Next Level

By **JOHN MARKOFF**  
Published: September 26, 2005

**The New York Times**  
ON THE WEB

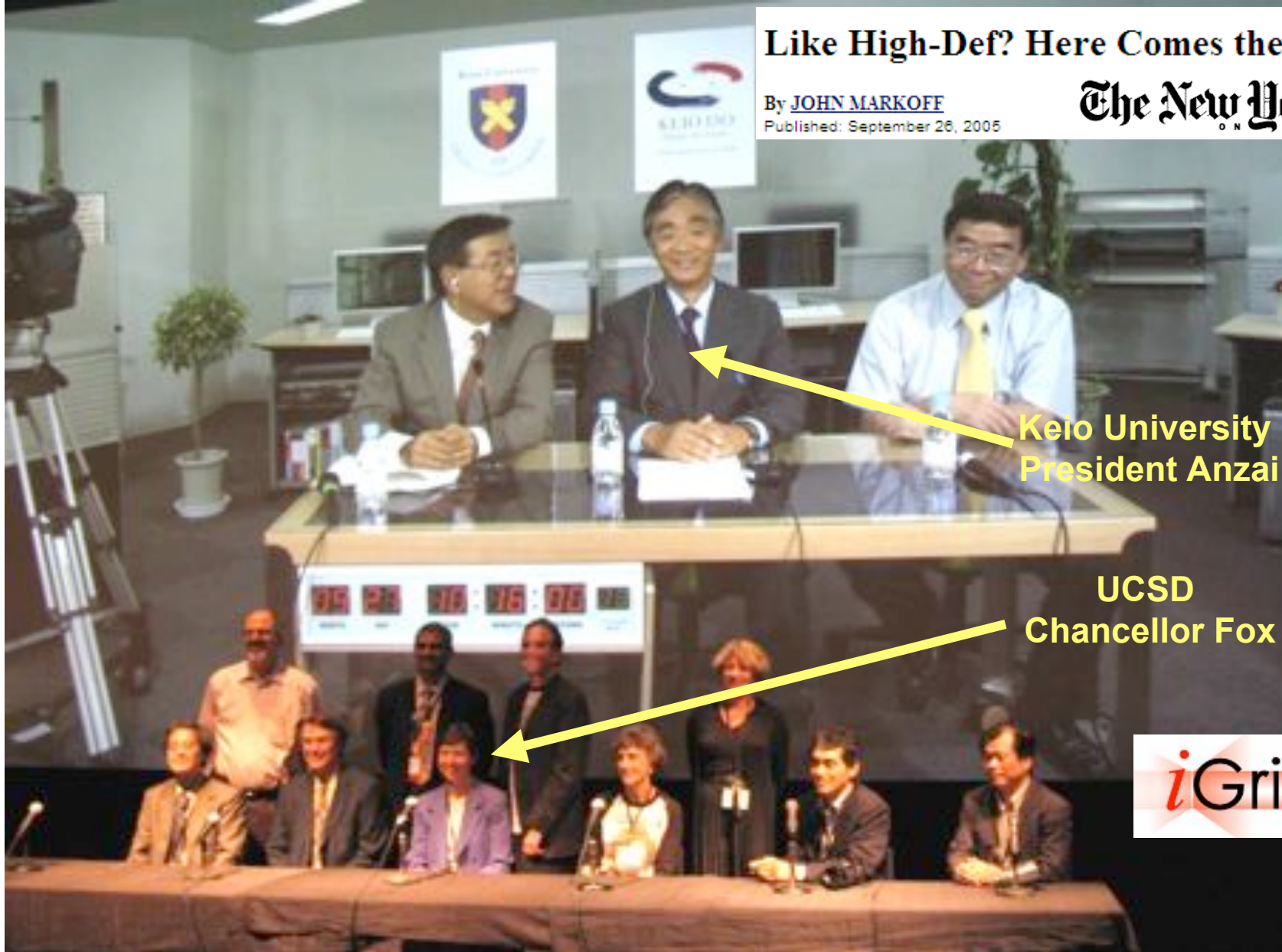
Used  
1Gbps  
Dedicated

Sony  
NTT  
SGI

Keio University  
President Anzai

UCSD  
Chancellor Fox

iGrid 2005



# CineGrid@SARA



# First Remote Interactive High Definition Video Exploration of Deep Sea Vents



## VISIONS 2005

EXPEDITION TO THE UNDERWATER VOLCANOES OF THE NORTHEAST PACIFIC



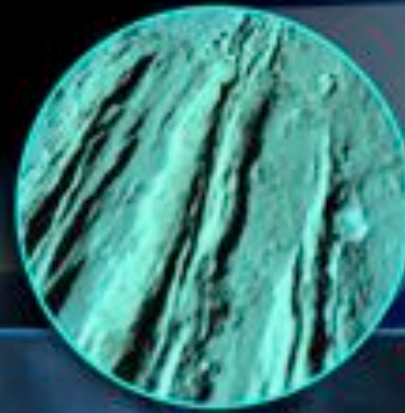
Galaxy XR Satellite

Ku-Band



RV Thompson

Collaboration

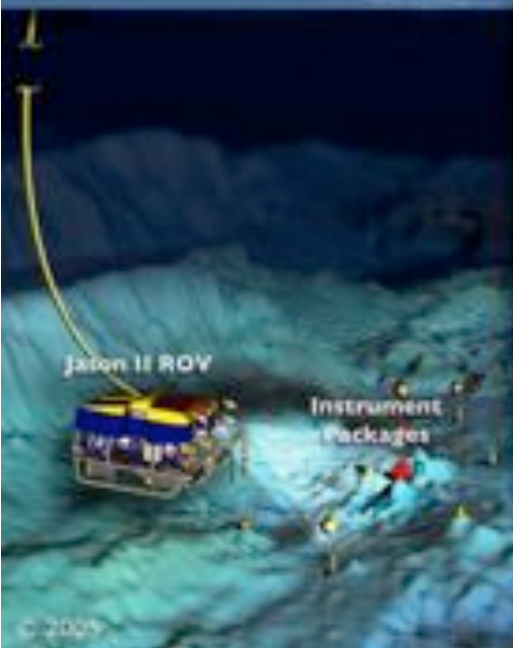


Endeavour Vent Fields

UW Research Channel

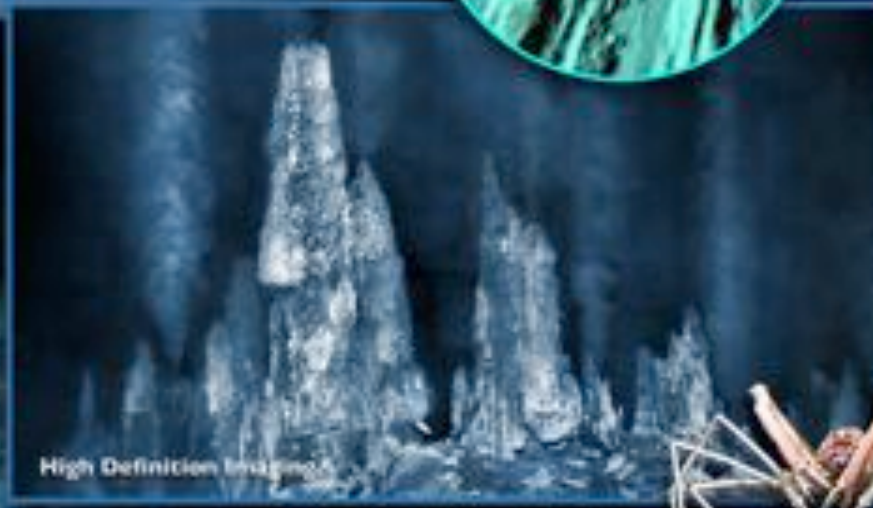
Cal-TT

I-Grid 2005



Jason II ROV

Instrument Packages



High Definition Imaging



Real-time Broadcasts:

Dates: September 28th & 29th Time: 2 to 3 pm (Pacific)

[WWW.VISIONS05.WASHINGTON.EDU](http://WWW.VISIONS05.WASHINGTON.EDU)

© 2005



Source John Delaney & Deborah Kelley, UWash

# US and International OptIPortal Sites



SIO



NCMIR



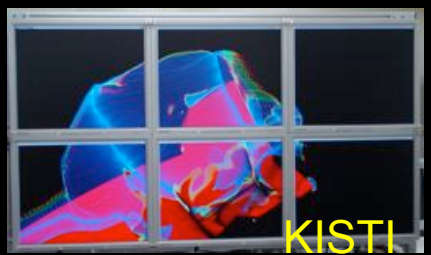
USGS EDC



NCSA &  
TRECC



SARA



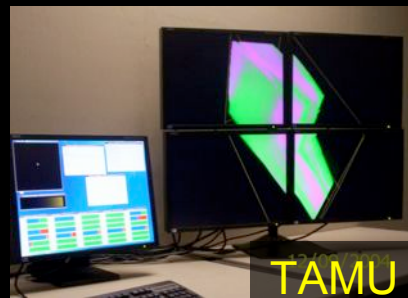
KISTI



AIST



RINCON & Nortel



TAMU



UCI



UIC



CALIT2



# The “Dead Cat” demo

SC2004 & iGrid2005

SC2004,  
Pittsburgh,  
Nov. 6 to 12, 2004  
iGrid2005,  
San Diego,  
sept. 2005

Produced by:  
Michael Scarpa  
Robert Belleman  
Peter Slood

Many thanks to:  
AMC  
SARA  
GigaPort  
UvA/AIR  
Silicon Graphics,  
Inc.  
Zoölogisch Museum

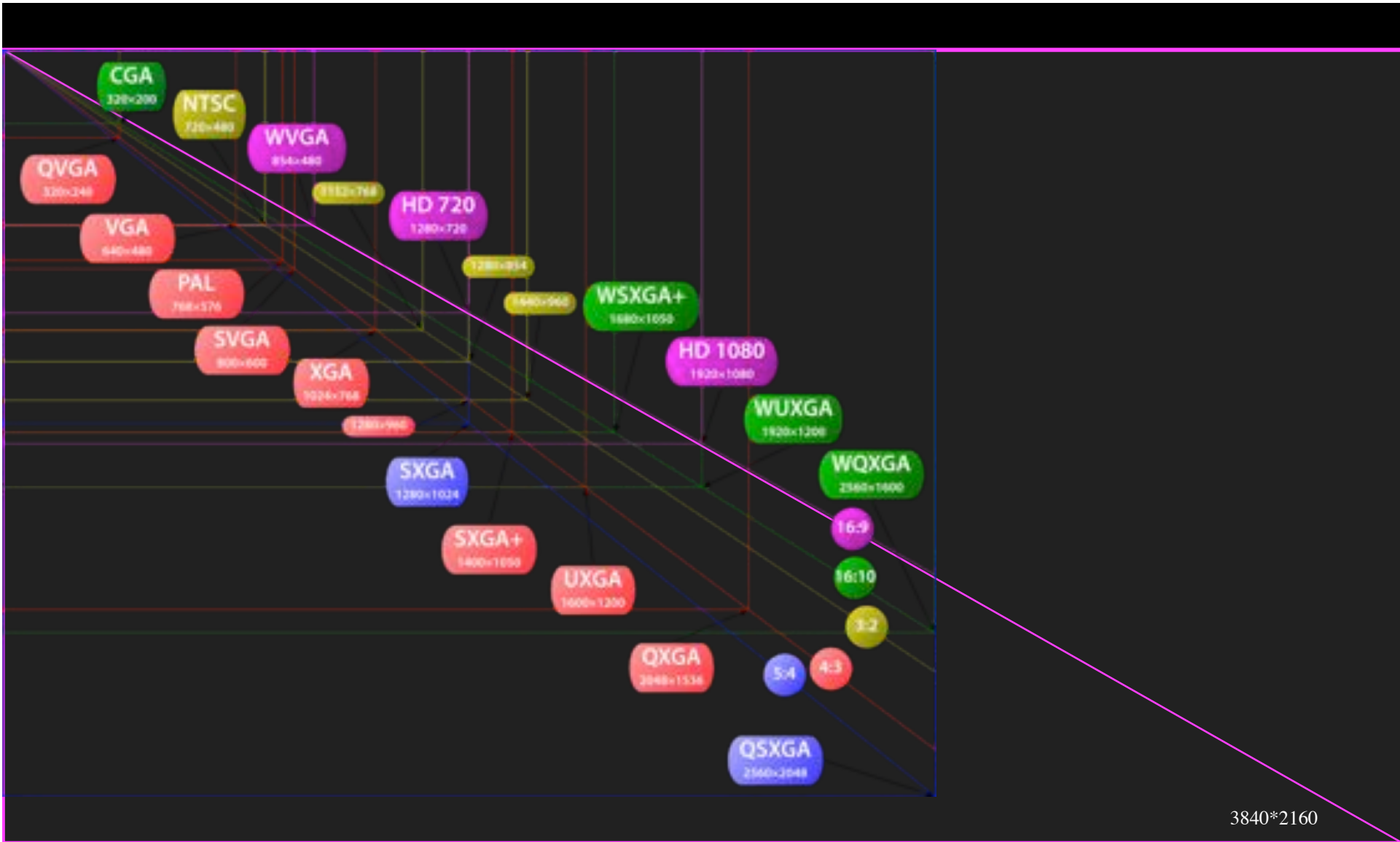




# Contents

1. Use cases CineGrid & Networks
2. **Formats - Numbers - Bits**
3. Global Lambda Integrated Facility
4. A LightPath
5. Transport Protocol issues
6. End System Issues
7. Network Storage
8. Q/A





# Formats - Numbers - Bits



# Format - Numbers - Bits (examples!)

Format	X	Y	Rate /s	Color bits/pix	Frame pix	Frame MByte	Flow MByt/s	Stream Gbit/s
720p HD	1280	720	60	24	921600	2.8	170	1.3
1080p HD	1920	1080	30	24	2073600	6.2	190	1.5
2k	2048	1080	24 48	36	2211840	10	240 480	1.2 2.4
SHD	3840	2160	30	24	8294400	25	750	6.0
4k	4096	2160	24	36	8847360	40	960	7.6

Note: this is excluding sound!

Note: these are raw uncompressed data rates ex overhead!



# Formats - Numbers - Bits

- Formats:

- uncompressed unreadable (UMF)

3/4 GBytesec

- compressed unreadable (jpeg2000)

300 - 700 Mbit/s

- uncompressed readable (eg TIFF)

1.2 GB/s, 4.3 TB/h

- compressed readable (eg DXT)

300 - 800 Mbit/s

- Do not compress away the science!

- Storage

- Holland festival taking uncompressed about 12 TByte



# Number, numbers and more numbers!

- **Digital Motion Picture for Audio Post-Production**
  - 1 TV Episode Dubbing Reference 1 GB
  - 1 Theatrical 5.1 Final Mix 8 GB
  - 1 Theatrical Feature Dubbing reference 30 GB
- **Digital Motion Picture Acquisition**
  - 6:1 up to 20:1 shooting ratios
  - 4k @ 24 FPS @ 10bit/color: ~48MB/Frame uncompressed
  - ~8TB for Finished 2 Hr Feature
- **Digital Dailies**
  - HD compressed MPEG-2 @ 25Mb/s
  - Data Size: ~22GB for 2 Hours
- **Digital Post-production and Visual Effects**
  - Terabytes, Gigabytes, Megabytes To Select Sites Depending on Project
- **Digital Motion Picture Distribution**
  - Film Printing in Regions
    - Features ~8TB
    - Trailers ~200GB
  - Digital Cinema to Theatres
    - Features ~200 - 300GB DCP
    - Trailers ~2 - 4GB DCP
- **Online Download**
  - Features ~1.3GB
  - TV Shows ~600MB



# Contents

1. Use cases CineGrid & Networks
2. Formats - Numbers - Bits
3. **Global Lambda Integrated Facility**
4. A LightPath
5. Transport Protocol issues
6. End System Issues
7. Network Storage
8. Q/A



# GLIF Mission Statement

- GLIF is a world-scale Lambda-based Laboratory for **application** and **middleware development** on emerging LambdaGrids, where applications rely on dynamically configured networks based on optical wavelengths
- GLIF is an environment (networking infrastructure, network engineering, system integration, middleware, applications) to accomplish **real work**





GLIF 2008

Visualization courtesy of Bob Patterson, NCSA  
Data collection by Maxine Brown.





# Calit2 is Partnering with CENIC to Connect California Industries and Researchers Into CineGrid

Partnering with SFSU's Institute for Next Generation Internet

Calit2's CineGrid Team is Working with Cinema Industry in LA and SF

In addition, 1Gb and 10Gb Connections to:

- Seattle then to Asia, Australia, Canada
- Chicago, Amsterdam, Europe, Russia, Asia
- Tijuana, Rosarita Beach, Ensenada

Extending SoCal OptIPuter to USC School of Cinema-Television

Laurin Herr,  
Pacific Interface  
Project Leader

Digital Archive of Films

Prototype of CineGrid

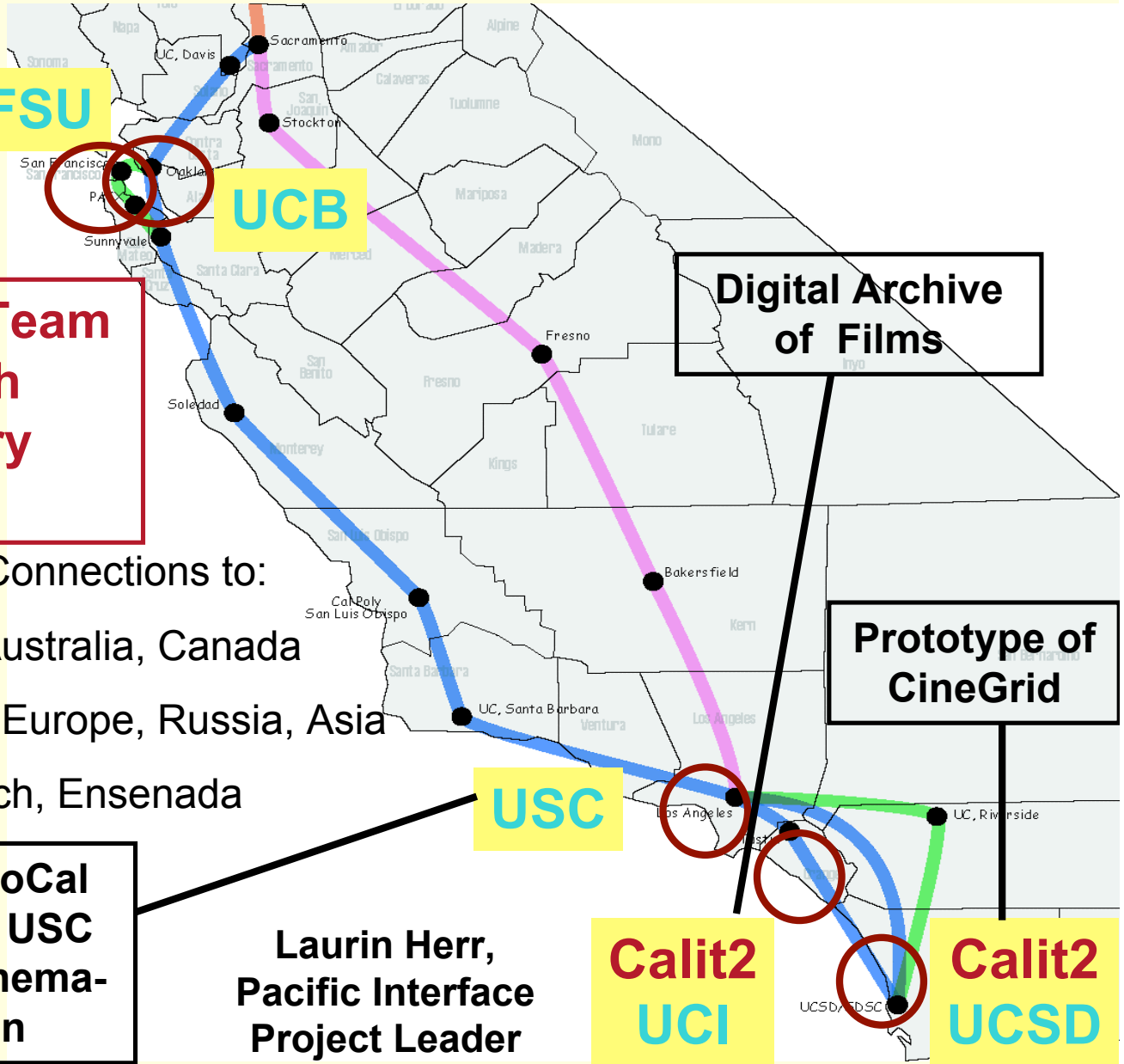
Calit2  
UCI

Calit2  
UCSD

SFSU

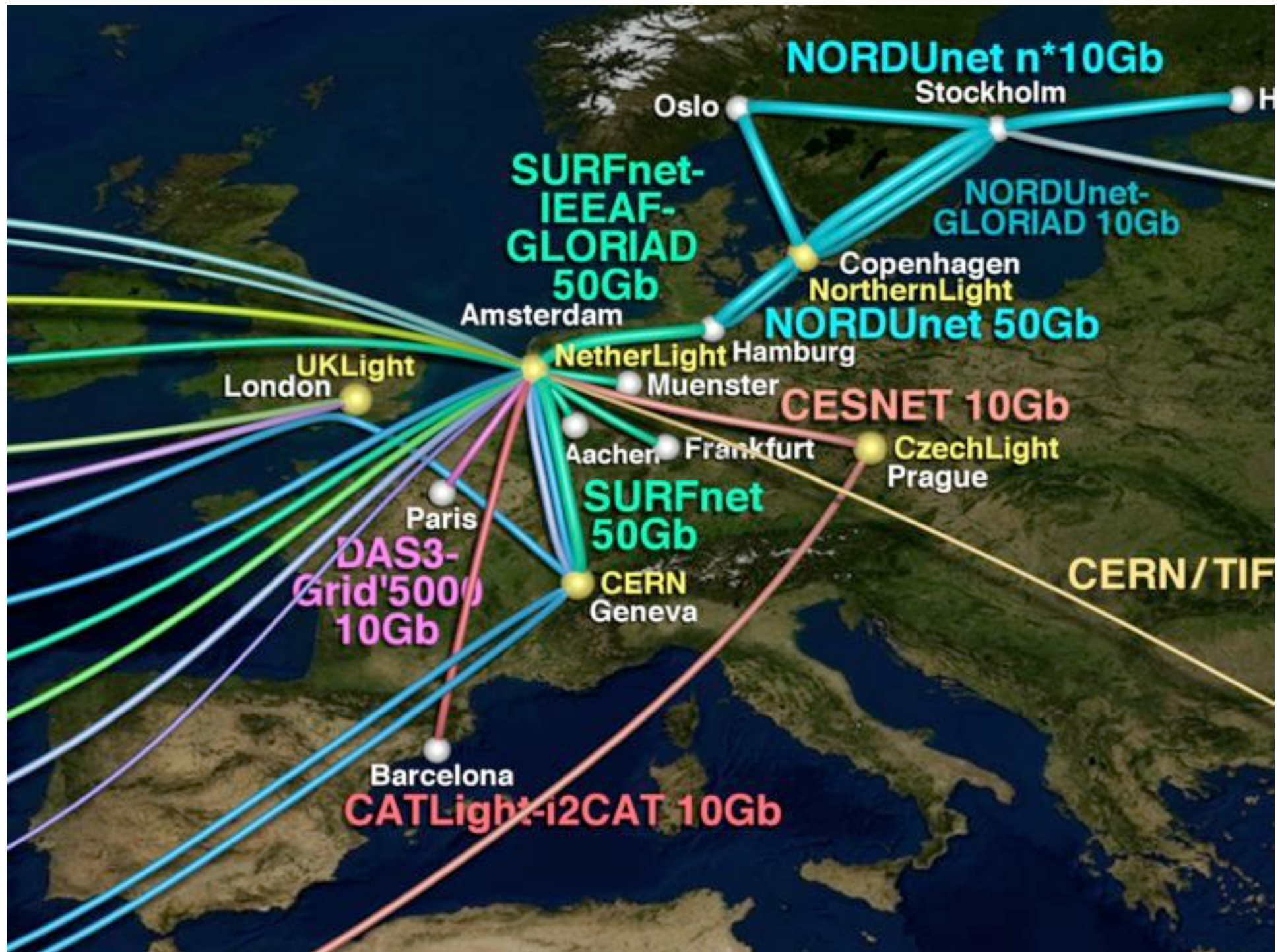
UCB

USC



# CENIC Connects to 10Gb Research and Education Networks Nationwide and Worldwide





VIZ

DATA

NetherLight

GRID

SUPER

DataExploration

RemoteControl

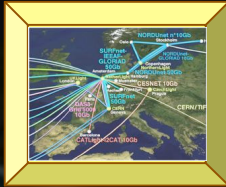
Management

Backup

TV

Medical

CineGrid



Gaming

Mining

Web2.0



Media

Visualisation

Conference

Meta

Security

Workflow

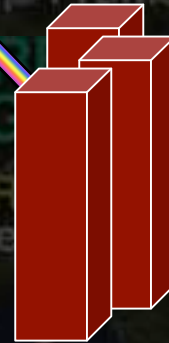
Clouds



Distributed

EventProcessing

NetherLight



Simulations

StreamProcessing

Predictions

Barcelona CATLight-12CAT 10Gb

NORUnet n\*10Gb Stockholm

SUNET IEEAF-GLORIAN 50Gb

NORUnet-GLORIAD 10Gb Copenhagen NorthernLight

Amsterdam 50Gb Muenster CESNET 10Gb

London Aachen Frankfurt CzechLight

Geneva CERN 7 T/E

# Contents

1. Use cases CineGrid & Networks
2. Formats - Numbers - Bits
3. Global Lambda Integrated Facility
4. **A LightPath**
5. Transport Protocol issues
6. End System Issues
7. Network Storage
8. Q/A



# What is a LightPath

- A LightPath is a circuit like connection that connects end systems to each other. This uses usually the same infrastructure as the Internet, but a LightPath gets dedicated resources next to Internet.
- A LightPath can be a combination of:
  - A color in a fiber (Lambda)
  - Sonet/sdh circuit in a sonet infrastructure
  - Vlans and dedicated ports in an ethernet switch
  - Etc.
- Aim is to get predictable and knowable connection characteristics
- Let us look at examples setups used recently!



New:  Overview  Throughput  Scroll line 
  
 Keyval  Load  Ping  UDP  Plot

Ping All [ms] from / to node125.das3.hacs.nl (LIACS-125)

Skipped tests: UvA-236-M, UvA-239-M

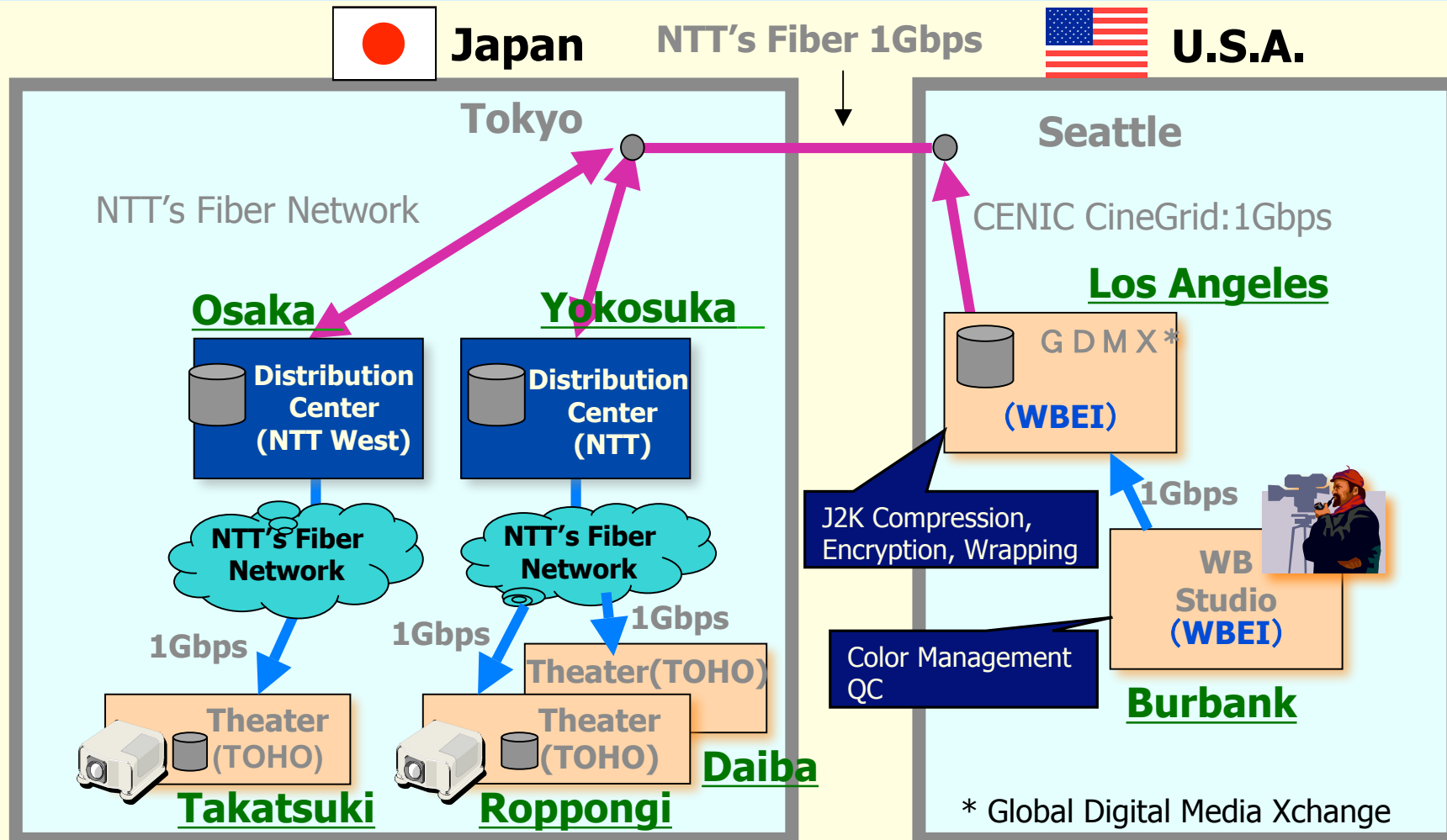
Date	Time	⇒ YU-083	⇐ YU-083	⇒ YU-085	⇐ YU-085	⇒ LIACS-127	⇐ LIACS-127	⇒ UvA-236	⇐ UvA-236	⇒ UvA-239	⇐ UvA-239
31/05/2007	12:30:01			1.380 / 1.382 / 1.410	1.380 / 1.383 / 1.420						
31/05/2007	12:00:01			1.380 / 1.383 / 1.410	1.380 / 1.384 / 1.450						
31/05/2007	11:30:01			1.380 / 1.383 / 1.410	1.380 / 1.382 / 1.390						
31/05/2007	11:00:02			1.380 / 1.382 / 1.410	1.380 / 1.382 / 1.400						
31/05/2007	10:30:01			1.380 / 1.383 / 1.390	1.380 / 1.382 / 1.390						
31/05/2007	10:00:01			1.380 / 1.382 / 1.410	1.380 / 1.383 / 1.410						
31/05/2007	09:30:01			1.380 / 1.384 / 1.410	1.380 / 1.382 / 1.400						
31/05/2007	09:00:01			1.380 / 1.382 / 1.410	1.380 / 1.383 / 1.400						
31/05/2007	08:30:02			1.380 / 1.383 / 1.410	1.380 / 1.382 / 1.400						
31/05/2007	08:00:01			1.380 / 1.383 / 1.410	1.380 / 1.383 / 1.410						
31/05/2007	07:30:02			1.380 / 1.382 / 1.390	1.380 / 1.381 / 1.390						
31/05/2007	07:00:01			1.380 / 1.382 / 1.410	1.380 / 1.383 / 1.400						
31/05/2007	06:30:01			1.380 / 1.383 / 1.410	1.380 / 1.382 / 1.390						
31/05/2007	06:00:01			1.380 / 1.382 / 1.410	1.380 / 1.382 / 1.420						
31/05/2007	05:30:01			1.380 / 1.382 / 1.400	1.380 / 1.382 / 1.410						
31/05/2007	05:00:01			1.380 / 1.382 / 1.410	1.380 / 1.382 / 1.390						
31/05/2007	04:30:01			1.380 / 1.381 / 1.390	1.380 / 1.380 / 1.390						
31/05/2007	04:00:01			1.380 / 1.382 / 1.410	1.380 / 1.384 / 1.410						
31/05/2007	03:30:02			1.380 / 1.384 / 1.410	1.380 / 1.382 / 1.400						
31/05/2007	03:00:02			1.380 / 1.382 / 1.410	1.380 / 1.382 / 1.400						
31/05/2007	02:30:01			1.380 / 1.382 / 1.400	1.380 / 1.382 / 1.400						
31/05/2007	02:00:01			1.380 / 1.383 / 1.410	1.380 / 1.384 / 1.410						
31/05/2007	01:30:01			1.380 / 1.382 / 1.410	1.380 / 1.382 / 1.390						
31/05/2007	01:00:01			1.380 / 1.382 / 1.410	1.380 / 1.383 / 1.400						

Very constant and predictable!

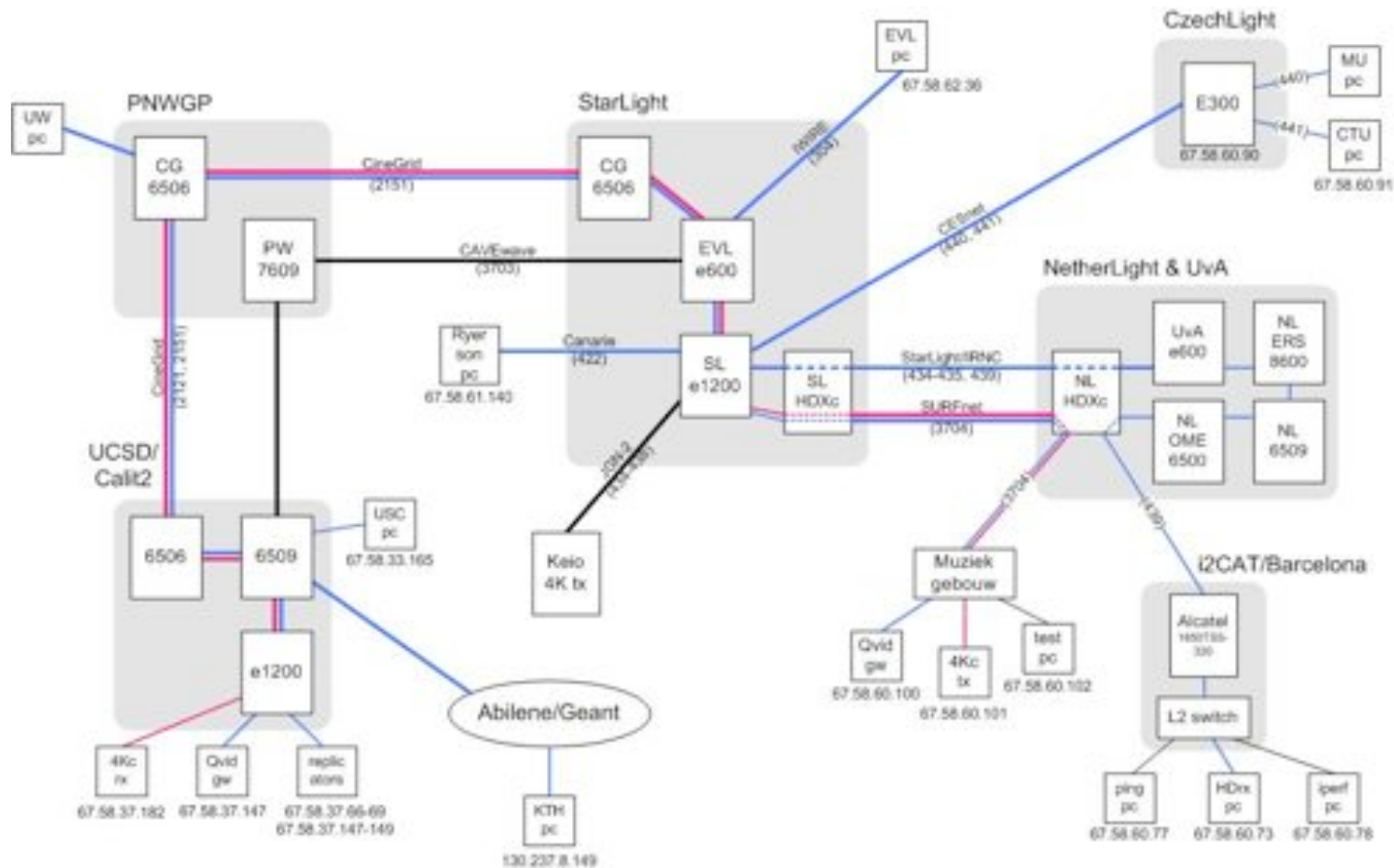


# Network for “4K Pure Cinema” Trial

DCP is directly transferred from GDMX in LA to distribution centers in Japan via fiber network. Within Japan, DCP is distributed from the distribution centers to TOHO theaters. Key is distributed from Osaka center, based on the contract between WB Japan and TOHO cinemas.





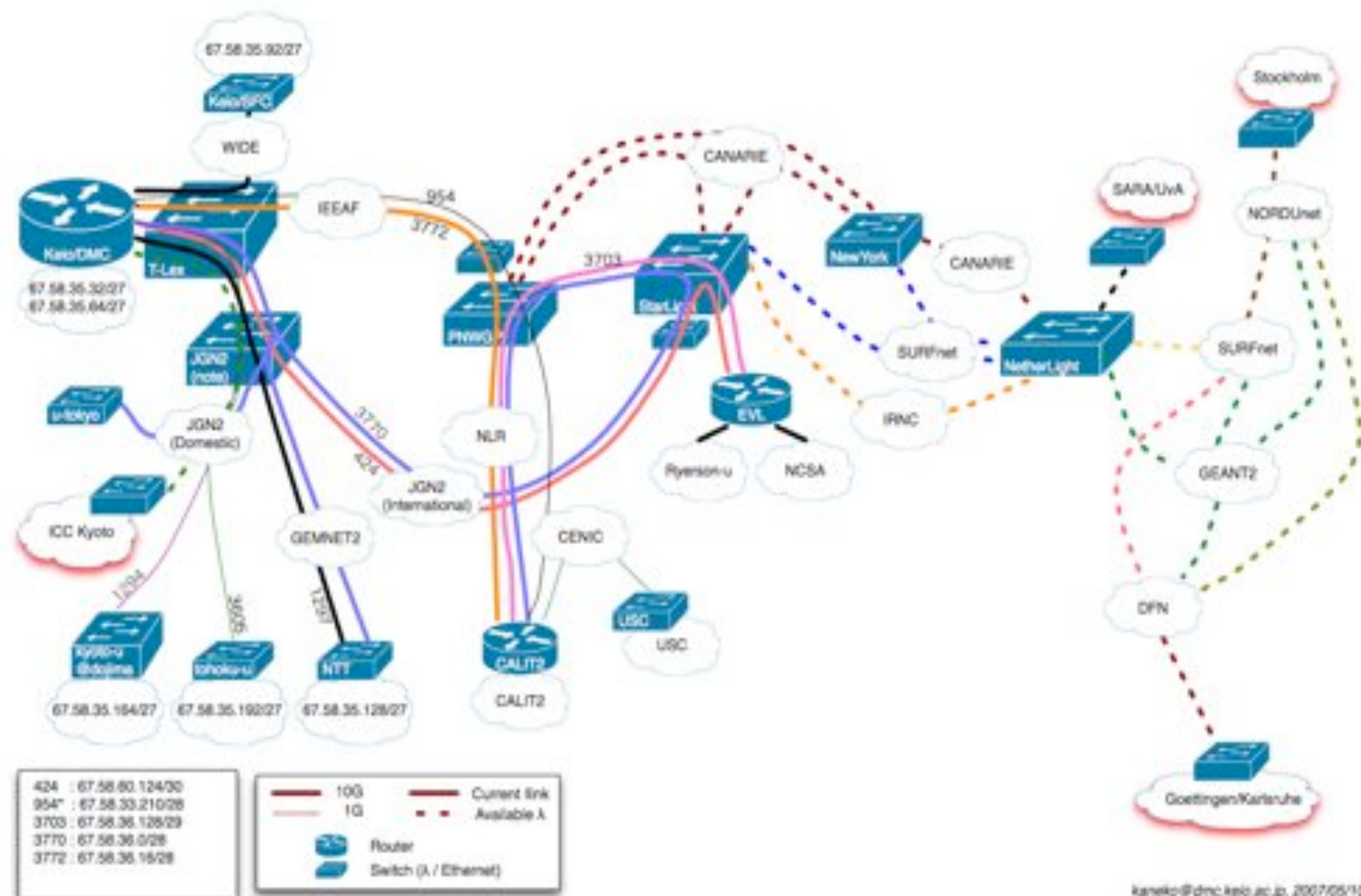


— AMS -> Calit2  
 4K compressed  
— AMS -> Calit2 -> sites  
 HD compressed

## Holland Festival CineGrid 2007

19-21 June 2007  
 Drawing by Alan Verlo, et al.

## Current Links & Available Links for Kyoto Prize Events



# Contents

1. Use cases CineGrid & Networks
2. Formats - Numbers - Bits
3. Global Lambda Integrated Facility
4. A LightPath
5. **Transport Protocol issues**
6. End System Issues
7. Network Storage
8. Q/A



# Internet Transport Protocols

- **IP = Internet Protocol**
  - Connectionless packet transport service
  - Datagrams of max 64 kByte
  - Can be fragmented down the way
  - Packets can get lost, duplicated or out of order!
- **TCP/IP = Transmission Control Protocol**
  - Reliable byte-stream over potentially unreliable packet service
  - Connection oriented, exactly once and in order, end to end duplex
- **UDP = User Datagram Protocol**
  - Packet service up to 64 kByte
  - Connectionless, unidirectional, L2 switches may start flooding
  - Unreliable delivery, can get out of order, duplicated, lost



# Flow control vs Congestion control

- Flow control
  - To prevent a fast sender overflowing a slow receiver
  - Receiver signals sender so it can adapt
- Congestion control
  - Traffic jams in the Internet: packets may get lost
  - For TCP protocol control loops via ack's and ICMP packets
  - TCP is friendly protocol, can adapt but performance usually takes severe hit
  - RTT is reaction and recovery time



# Windows and buffering for reliable protocols

- Round Trip Time (rtt) is time it takes to send a shortest message and get the answer back (unix tool ping)
- That is the shortest time the sender can know that traffic arrived at the other end
- Sender can only discard old data after receiving ack's
- Lightspeed in fiber = 200000 km/s
- 100 km = 200 km round trip = 1/1000 sec = 1 ms rtt
  - Amsterdam - Geneve  $\approx$  20 ms
  - Amsterdam - Chicago  $\approx$  90 ms
  - Amsterdam - San Diego  $\approx$  160 ms
  - Amsterdam - Tokyo  $\approx$  250 ms
  - Amsterdam - Sydney  $\approx$  300 ms



# Buffer space

$$\text{Window} = \text{RTT} * \text{BW}$$

RTT	100 Mbit/s	1 Gbit/s	10 Gbit/s
1	12.5 kB	125 kB	1.25 MB
2	25 kB	250 kB	2.5 MB
5	62.5 kB	615 kB	6.15 MB
10	125 kB	1.25 MB	12.5 MB
20	250 kB	2.5 MB	25 MB
50	625 kB	6.25 MB	62.5 MB
100	1.25 MB	12.5 MB	125 MB
200	2.5 MB	25 MB	250 MB
500	6.25 MB	62.5 MB	625 MB
1000	12.5 MB	125 MB	1250 MB



# TCP Tuning (if not auto-tuning)

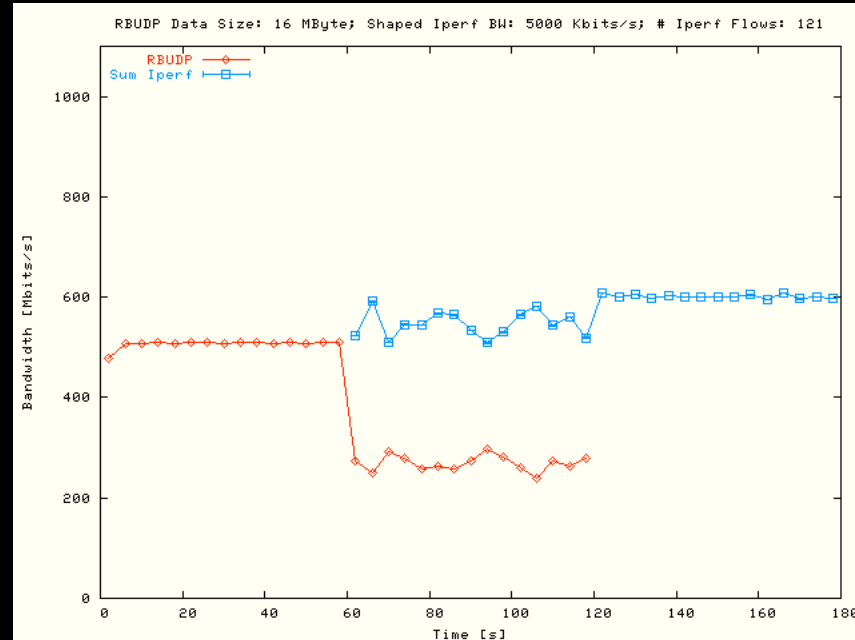
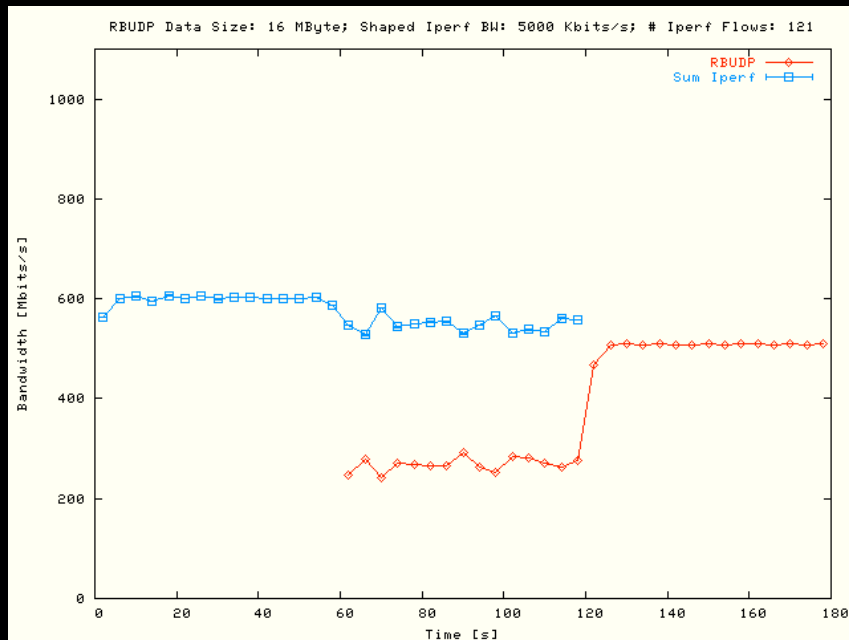
- 1 Gbit/s on 160 ms RTT (= Amsterdam - San Diego) :
  - `sysctl -w kern.ipc.maxsockbuf=50000000`
  - `sysctl -w net.inet.tcp.sendspace=21000000`
  - `sysctl -w net.inet.tcp.recvspace=21000000`
  - `sysctl -w net.inet.udp.maxdgram=57344`
  - `sysctl -w net.inet.udp.recvspace=74848`
  - `sysctl -w net.local.stream.sendspace=32768`
  - `sysctl -w net.local.stream.recvspace=32768`
  - `sysctl -w kern.ipc.somaxconn=512`
  - `sysctl -w net.inet.tcp.mssdflt=1460`
  - `sysctl -w net.inet.tcp.delayed_ack=2`
  - `sysctl -w net.inet.tcp.rfc1323=1`
  - `sysctl -w net.inet.tcp.rfc1644=1`
  - `sysctl -w net.inet.tcp.newreno=1`





# Other issues & protocols

- When using UDP watch for bottleneck!
- About 10 other non standard protocols
- FAST TCP
  - Modified receiver algorithms
- RBUDP
  - Runs on top of UDP, simple back-off and retransmission scheme



# Contents

1. Use cases CineGrid & Networks
2. Formats - Numbers - Bits
3. Global Lambda Integrated Facility
4. A LightPath
5. Transport Protocol issues
6. End System Issues
7. Network Storage
8. Q/A



# End System Issues

- Ethernet card interface to computer bus system
  - PCI-X
    - 32/64 bit 66/133/266 MHZ -> about 8 Gbit/s max in 133 MHZ mode
  - PCI-Express
    - 2.5 Gbit/s per lane, 4, 8, 16 lanes
- Memory organization
- CPU cache
  - Effect when things go out of cache (small windows, etc.)
- CPU core
  - Takes 1 core to handle network (affinity may help)
- Disk raid subsystem
  - raid0 twice as fast as raid5
  - One disk does typically 40 MB/s write, 60 MB/s read



# Contents

1. Use cases CineGrid & Networks
2. Formats - Numbers - Bits
3. Global Lambda Integrated Facility
4. A LightPath
5. Transport Protocol issues
6. End System Issues
7. Network Storage
8. Q/A



# Amsterdam CineGrid S/F node

## “COCE”

DAS-3 @ UvA

DP AMD processor nodes

comp node

⋮ 77x

comp node

head node

bridge node

bridge node

bridge node

bridge node

bridge node

bridge node

bridge node

bridge node

storage node

100 TByte

M  
Y  
R  
I  
N  
E  
T

NetherLight, StarPlane  
the cp testbeds  
and beyond

Rembrandt Cluster  
total 22 TByte disk space  
@ LightHouse

Opteron 64 bit nodes

head node

comp node

comp node

comp node

comp node

comp node

comp node

comp node

comp node

comp node

streaming node

8 TByte

**GlimmerGlass  
photonic switch**

10 Gbit/s

10 Gbit/s

NORTEL  
8600  
L2/3 switch

F10  
L2/3 switch

10 Gbit/s

suitcees &  
briefcees

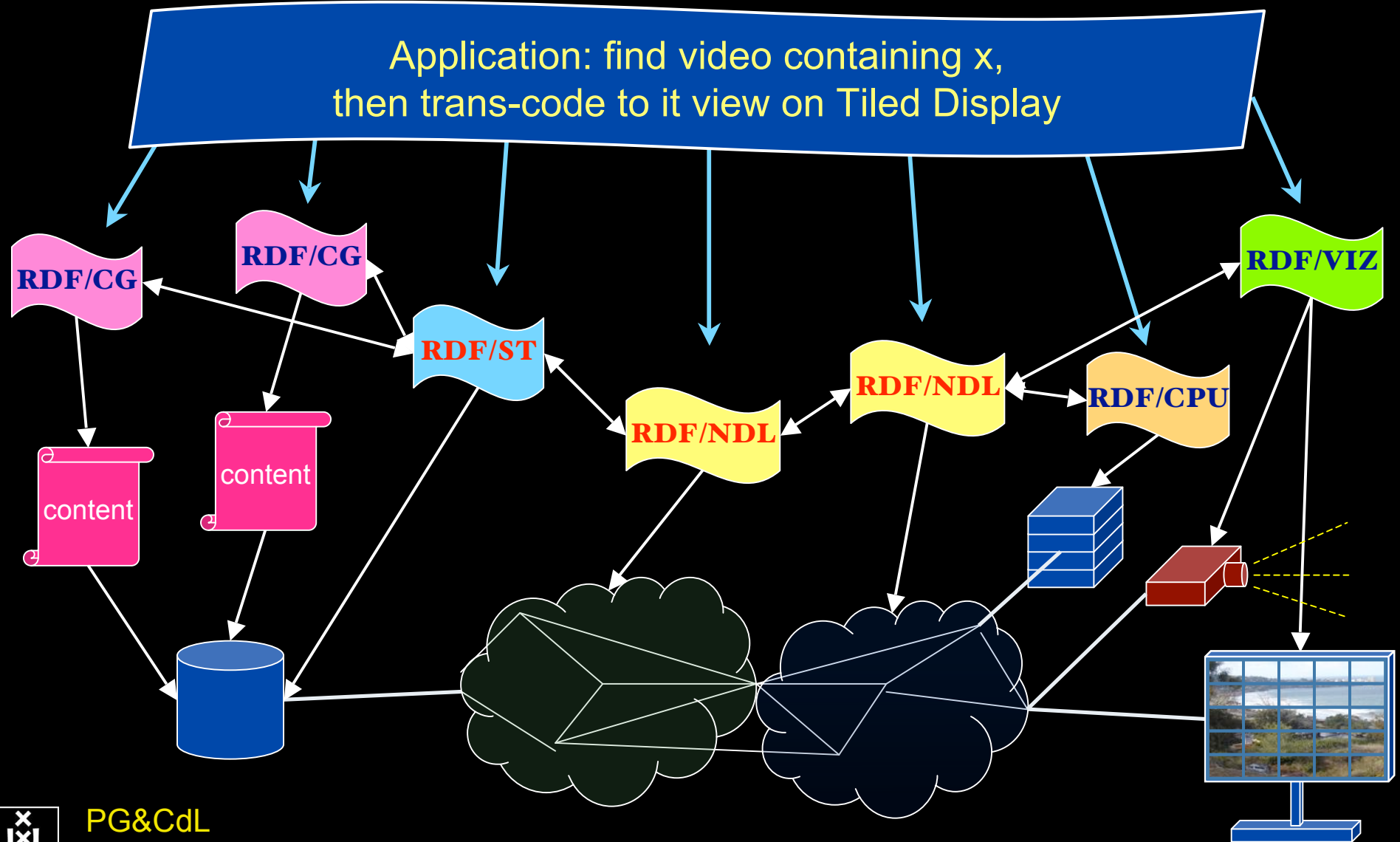
SURF  
NET

Node 41

sara



# RDF describing Infrastructure



# Contents

1. Use cases CineGrid & Networks
2. Formats - Numbers - Bits
3. Global Lambda Integrated Facility
4. A LightPath
5. Transport Protocol issues
6. End System Issues
7. Network Storage
8. Q/A





# Questions ?

[www.cinegrid.org](http://www.cinegrid.org)

[www.cinegrid.nl](http://www.cinegrid.nl)

[www.supertube.org](http://www.supertube.org)

[www.science.uva.nl/~delaat](http://www.science.uva.nl/~delaat)

