Internet Innovation to support Science & Education.

Cees de Laat

EU COMMIT

UvA

NWO

PID/EFRO

SURFnet

NLESC

TNO

NWO/ncf
… more users!

… more data!

… more realtime!

Internet developments
Internet developments

... more data!

Speed

Volume

Deterministic

Real-time

Scalable

Secure

... more users!
GPU cards are disruptive!

- Fastest supercomputer in the world
- Nr. 500 supercomputer in the world
- 1 single Graphics Processing Unit

Top 500

$20,000,000$

$500$

7 year

$100$ piflops

$10$ piflops

$1$ piflops

$100$ tflops

$10$ tflops

$1$ tflops

$100$ gflops

$10$ gflops

$1$ gflops

$100$ mflops

$10$ mflops

$1$ mflops

1993

1995

1997

1999

2001

2003

2005

2007

2009

2011

2013

2015

2017

2019
Data storage: doubling every 1.5 year!
Multiple colors / Fiber

Per fiber: ~ 80-100 colors * 50 GHz
Per color: 10 – 40 – 100 Gbit/s
BW * Distance ~ 2*10^{17} bm/s

Wavelength Selective Switch

New: Hollow Fiber!
⇒ less RTT!
Next Generation Wireless LAN Technology 802.11ac 1 Gbps throughput with

WiFi is one of the most preferred communication protocol LAN due to the easy comparison and convenience in the digital home. While consumer PC products has just started to migrate to a much higher bandwidth of 802.11n wireless LAN now working on next-generation standard definition is already in progress.
Wireless Networks

It is a bit freaky with this wireless technology.

protocol LAN due to the easy comparison and convenience in the digital home. While consumer PC products has just started to migrate to a much higher bandwidth of 802.11n wireless LAN now working on next-generation standard definition is already in progress.
<table>
<thead>
<tr>
<th></th>
<th>Ijkdijk/Urban Flood</th>
<th>LifeWatch/ENVR</th>
<th>CosmoGrid/eVLBI</th>
<th>CineGrid</th>
<th>SURFnet/GLIF/Cloud</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green-IT</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Privacy/Trust</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Authorization/policy</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Programmable networks</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40-100Gig/TCP/WF/QoS</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Topology/Architecture</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Optical Photonic</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>
ATLAS detector @ CERN Geneve
LHC Data Grid Hierarchy
CMS as example, Atlas is similar

CMS detector: 15m X 15m X 22m
12,500 tons, $700M.

Tier 1
Italian Regional Center
German Regional Center
NIKHEF Dutch Regional Center
FermiLab, USA Regional Center

Tier 0 + 1

Tier 2
Tier2 Center
Center
Center
Center

Tier 3
Institute
Institute
Institute

Tier 4
Physics data cache
Workstations

~PByte/sec
100000 flops/byte
10 Pflops/s

Online System

~100 MBytes/sec
event reconstruction

~2.5 Gbits/sec

~0.6-2.5 Gbps

Status 2002!

CERN/CMS data goes to 6-8 Tier 1 regional centers, and from each of these to 6-10 Tier 2 centers.

Physicists work on analysis “channels” at 135 institutes. Each institute has ~10 physicists working on one or more channels.

2000 physicists in 31 countries are involved in this 20-year experiment in which DOE is a major player.

Courtesy Harvey Newman, CalTech and CERN
A. Lightweight users, browsing, mailing, home use
   Need full Internet routing, one to all

B. Business/grid applications, multicast, streaming, VO’s, mostly LAN
   Need VPN services and full Internet routing, several to several + uplink to all

C. E-Science applications, distributed data processing, all sorts of grids
   Need very fat pipes, limited multiple Virtual Organizations, P2P, few to few

For the Netherlands 2011
\[ \Sigma A = \Sigma B = \Sigma C \approx 1 \text{ Tb/s} \]

However:
A -> all connects
B -> on several
C -> just a few (SP, LHC, LOFAR)

Ref: Cees de Laat, Erik Radius, Steven Wallace, "The Rationale of the Current Optical Networking Initiatives"
Towards Hybrid Networking!

- Costs of photonic equipment 10% of switching 10% of full routing
  - for same throughput!
  - Photonic vs Optical (optical used for SONET, etc, 10-50 k$/port)
  - DWDM lasers for long reach expensive, 10-50 k$

- Bottom line: look for a hybrid architecture which serves all classes in a cost effective way
  - map A -> L3, B -> L2, C -> L1 and L2

- Give each packet in the network the service it needs, but no more!

$L1 \approx 2-3 \text{ k$}/\text{port}$  
$L2 \approx 5-8 \text{ k$}/\text{port}$  
$L3 \approx 75+ \text{ k$}/\text{port}$
How low can you go?
The GLIF – lightpaths around the world
The GLIF – lightpaths around the world
In The Netherlands SURFnet connects between 180:
- universities;
- academic hospitals;
- most polytechnics;
- research centers.
with an indirect ~750K user base

~ 8860 km scale comparable to railway system
Alien light
From idea to realisation!

40Gb/s alien wavelength transmission via a multi-vendor 10Gb/s DWDM infrastructure

Alien wavelength advantages
- Direct connection of customer equipment\(^1\) ➔ cost savings
- Avoid OEO regeneration ➔ power savings
- Faster time to service\(^2\) ➔ time savings
- Support of different modulation formats\(^3\) ➔ extend network lifetime

Alien wavelength challenges
- Complex end-to-end optical path engineering in terms of linear (i.e. OSNR, dispersion) and non-linear (FWM, SPM, XPM, Raman) transmission effects for different modulation formats.
- Complex interoperability testing.
- End-to-end monitoring, fault isolation and resolution.
- End-to-end service activation.

In this demonstration we will investigate the performance of a 40Gb/s PM-QPSK alien wavelength installed on a 10Gb/s DWDM infrastructure.

New method to present fiber link quality, FoM (Figure of Merit)
In order to quantify optical link grade, we propose a new method of representing system quality: the FOM (Figure of Merit) for concatenated fiber spans.

\[
\text{FOM} = \sum_{i=1}^{N} 10 \log_{10} L_i
\]

Where:
- \( L_i \) span losses in dB
- \( N \) number of spans

Transmission system setup
JOINT SURFnet-NORDUnet 40Gb/s PM-QPSK alien wavelength DEMONSTRATION.

Test results
Error-free transmission for 23 hours, 17 minutes ➔ BER < 3.0 \times 10^{-15}

Conclusions
- We have investigated experimentally the all-optical transmission of a 40Gb/s PM-QPSK alien wavelength via a concatenated native and third party DWDM system that both were carrying live 10Gb/s wavelengths.
- The end-to-end transmission system consisted of 1056 km of TWRS (TrueWave Reduced Slope) transmission fiber.
- We demonstrated error-free transmission (i.e. BER below 10-15) during a 23 hour period.
- More detailed system performance analysis will be presented in an upcoming paper.

Alien wavelength advantages
- Direct connection of customer equipment\(^1\) ➔ cost savings
- Avoid OEO regeneration ➔ power savings
- Faster time to service\(^2\) ➔ time savings
- Support of different modulation formats\(^3\) ➔ extend network lifetime

Test results
Error-free transmission for 23 hours, 17 minutes ➔ BER < 3.0 \times 10^{-15}

Conclusions
- We have investigated experimentally the all-optical transmission of a 40Gb/s PM-QPSK alien wavelength via a concatenated native and third party DWDM system that both were carrying live 10Gb/s wavelengths.
- The end-to-end transmission system consisted of 1056 km of TWRS (TrueWave Reduced Slope) transmission fiber.
- We demonstrated error-free transmission (i.e. BER below 10-15) during a 23 hour period.
- More detailed system performance analysis will be presented in an upcoming paper.

References
\(^1\) “OPTICAL SOLUTIONS FOR 40-GBIT/S NETWORKS,” O. GERSTE et al., NOC'03 WB \(^2\) “A BENCHMARK STUDY OF OEO REGENERATION,” AUDRANCE, GRICT, OPCI \(^3\) “A BENCHMARK OF AL UNB REGEN WW,” ABDUL A. AND AHMAD, SBARRA, ATB \(N\) NETWORKS

Acknowledgements
\(N\) ORTEL AND \(N\) ORTEL FOR PROVIDING US WITH BASELINE ON TWRS BI DONT FOR THE EXPERIMENT AND ALSO FOR THEIR SUPPORT AND ASSISTANCE DURING THE EXPERIMENTS. WE ALSO ACKNOWLEDGE TELEFIS AND NORTEL FOR THEIR INTEGRATION WORK AND SIMULATION SUPPORT.
40Gb/s alien wavelength transmission via a multi-vendor 10Gb/s DWDM infrastructure

Alien light
From idea to realisation!

Alien wavelength advantages
- Direct connection of customer equipment → cost savings
- Avoid OEO regeneration → power savings
- Faster time to service → time savings
- Support of different modulation formats → extend network lifetime

Alien wavelength challenges
- Complex end-to-end optical path engineering in terms of linear (i.e. OSNR, dispersion) and non-linear (FWM, SPM, XPM, Raman) transmission effects for different modulation formats.
- Complex interoperability testing.
- End-to-end monitoring, fault isolation and resolution.
- End-to-end service activation.

In this demonstration we will investigate the performance of a 40Gb/s PM-QPSK alien wavelength installed on a 10Gb/s DWDM infrastructure.

New method to present fiber link quality, FoM (Figure of Merit)
In order to quantify optical link grade, we propose a new method of representing system quality: the FOM (Figure of Merit) for concatenated fiber spans.

Transmission system setup
JOINT SURFnet-NORDUnet 40Gb/s PM-QPSK alien wavelength DEMONSTRATION.

Test results
Error-free transmission for 23 hours, 17 minutes → BER < 3.0 × 10⁻¹⁰

Conclusions
- We have investigated experimentally the all-optical transmission of a 40Gb/s PM-QPSK alien wavelength via a concatenated native and third party DWDM system that both were carrying live 10Gb/s wavelengths.
- The end-to-end transmission system consisted of 1056 km of TWRS (TrueWave Reduced Slope) transmission fiber.
- We demonstrated error-free transmission (i.e. BER below 10⁻¹⁵) during a 23 hour period.
- More detailed system performance analysis will be presented in an upcoming paper.
ClearStream @ TNC2011

Setup codename: FlightCees

UvA
- 2* dual 2.8 GHz Q-core
- I7 3.2 GHz Q-core
- Mellanox
- 40G E

Copenhagen
- 2* dual 2.8 GHz Q-core
- Mellanox

CIENA OME 6500

CIENA DWDM

CERN

Amsterdam – Geneva (CERN) – Copenhagen – 4400 km (2700 km alien light)
http://tnc11.delaat.net

Amsterdam (UvA) Live RX Traffic

Copenhagen POP RX Traffic

27.99 Gbps to Amsterdam <--> 31.45 Gbps to Copenhagen

Total Throughput 59.44 Gbps RTT 44.010 ms
Results (rtt = 17 ms)

- Single flow iPerf 1 core -> 21 Gbps
- Single flow iPerf 1 core <> -> 15+15 Gbps
- Multi flow iPerf 2 cores -> 25 Gbps
- Multi flow iPerf 2 cores <> -> 23+23 Gbps
- DiViNe <> -> 11 Gbps
- Multi flow iPerf + DiVine -> 35 Gbps
- Multi flow iPerf + DiVine <> -> 35 + 35 Gbps
Performance Explained

- Mellanox 40GE card is PCI-E 2.0 8x (5GT/s)
- 40Gbit/s raw throughput but ....
- PCI-E is a network-like protocol
  - 8/10 bit encoding -> 25% overhead -> 32Gbit/s maximum data throughput
  - Routing information
- Extra overhead from IP/Ethernet framing
- Server architecture matters!
  - 4P system performed worse in multithreaded iperf
Server Architecture

DELL R815
4 x AMD Opteron 6100

Supermicro X8DTT-HIBQF
2 x Intel Xeon
We used `numactl` to bind `iperf` to cores.
Where when will it happen?
IJKDIJK

Sensors: 15000km* 800 bps/m -> 12 Gbit/s to cover all Dutch dikes
Sensor grid: instrument the dikes

First controlled breach occurred on sept 27th ‘08:

Many small flows -> 12 Gb/s

Many Pflops/s
User Programmable Virtualized Networks.

The network is virtualized as a collection of resources.

UPVNs enable network resources to be programmed as part of the application.

Mathematica interacts with virtualized networks using UPVNs and optimize network + computation.

ref: Robert J. Meijer, Rudolf J. Strijkers, Leon Gommans, Cees de Laat, User Programmable Virtualized Networks, accepted for publication to the IEEE e-Science 2006 conference Amsterdam.
In the Intercloud virtual servers and networks become software

- Virtual Internets adapt to the environment, grow to demand, iterate to specific designs
- Network support for application specific interconnections are merely optimizations: Openflow, active networks, cisco distributed switch
- But how to control the control loop?

![Diagram of control loop](image)
Interactive Networks

Rudolf Strijkers\textsuperscript{1,2}
Marc X. Makkes\textsuperscript{1,2}
Mihai Christea\textsuperscript{1}
Laurence Muller\textsuperscript{1}
Robert Belleman\textsuperscript{1}
Gees de Laat\textsuperscript{1}
Robert Meijer\textsuperscript{1,2}

\textsuperscript{1} University of Amsterdam, Amsterdam The Netherlands
\textsuperscript{2} TNO Information and Communication Technology, Groningen, The Netherlands
We investigate: for complex networks!
LinkedIN for Infrastructure

- From semantic Web / Resource Description Framework.
- The RDF uses XML as an interchange syntax.
- Data is described by triplets (Friend of a Friend):

```
Subject  Predicate  Object

Location
  name
  connectedTo

Device
  description
  capacity

Interface
  locatedAt
  encodingType

Link
  hasInterface
  encodingLabel
```
NetherLight in RDF

<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:ndl="http://www.science.uva.nl/research/air/ndl#"/>
<!-- Description of Netherlight -->
<ndl:Location rdf:about="#Netherlight">
  <ndl:name>Netherlight Optical Exchange</ndl:name>
</ndl:Location>
<!-- TDM3.amsterdam1.netherlight.net -->
<ndl:Device rdf:about="#tdm3.amsterdam1.netherlight.net">
  <ndl:name>tdm3.amsterdam1.netherlight.net</ndl:name>
  <ndl:locatedAt rdf:resource="#amsterdam1.netherlight.net"/>
  <ndl:hasInterface rdf:resource="#tdm3.amsterdam1.netherlight.net:501/1"/>
  <ndl:hasInterface rdf:resource="#tdm3.amsterdam1.netherlight.net:501/3"/>
  <ndl:hasInterface rdf:resource="#tdm3.amsterdam1.netherlight.net:501/4"/>
  <ndl:hasInterface rdf:resource="#tdm3.amsterdam1.netherlight.net:503/1"/>
  <ndl:hasInterface rdf:resource="#tdm3.amsterdam1.netherlight.net:503/2"/>
  <ndl:hasInterface rdf:resource="#tdm3.amsterdam1.netherlight.net:503/3"/>
  <ndl:hasInterface rdf:resource="#tdm3.amsterdam1.netherlight.net:503/4"/>
  <ndl:hasInterface rdf:resource="#tdm3.amsterdam1.netherlight.net:501/2"/>
</ndl:Device>
<!-- all the interfaces of TDM3.amsterdam1.netherlight.net -->
<ndl:Interface rdf:about="#tdm3.amsterdam1.netherlight.net:501/1">
  <ndl:name>tdm3.amsterdam1.netherlight.net:POS501/1</ndl:name>
  <ndl:connectedTo rdf:resource="#tdm4.amsterdam1.netherlight.net:5/1"/>
</ndl:Interface>
<ndl:Interface rdf:about="#tdm3.amsterdam1.netherlight.net:501/2">
  <ndl:name>tdm3.amsterdam1.netherlight.net:POS501/2</ndl:name>
  <ndl:connectedTo rdf:resource="#tdm1.amsterdam1.netherlight.net:12/1"/>
</ndl:Interface>
Multi-layer descriptions in NDL

- IP layer
- Ethernet layer
- STS layer
- UTP layer
- SONET switch with Ethernet intf.
- Ethernet & SONET switch
- SONET switch
- SONET switch with Ethernet intf.
- End host

Université du Quebec
CA Net Canada
StarLight Chicago
MAN LAN New York
NetherLight Amsterdam
Universiteit van Amsterdam
Multi-layer Network PathFinding

Path between interfaces A1 and E1:
A1-A2-B1-B4-D4-D2-C3-C4-C1-C2-B2-B3-D3-D1-E2-E1

Scaling: Combinatorial problem
Define a common information model for **infrastructures** and **services**. Base it on Semantic Web.

J. van der Ham, F. Dijkstra, P. Grosso, R. van der Pol, A. Toonk, C. de Laat
*A distributed topology information system for optical networks based on the semantic web*,
In: Elsevier Journal on Optical Switching and Networking, Volume 5, Issues 2-3, June 2008, Pages 85-93

R. Koning, P. Grosso and C. de Laat
*Using ontologies for resource description in the CineGrid Exchange*
Why?

I want to:

“Show Big Bug Bunny in 4K on my Tiled Display using green Infrastructure”

- Big Bugs Bunny can be on multiple servers on the Internet.
- Movie may need processing / recoding to get to 4K for Tiled Display.
- Needs deterministic Green infrastructure for Quality of Experience.
- Consumer / Scientist does not want to know the underlying details.
  ➞ His refrigerator also just works.
RDF describing Infrastructure
“I want”

Application: find video containing x, then trans-code to it view on Tiled Display
The Ten Problems with the Internet

1. **Energy Efficient Communication**
2. Separation of Identity and Address
3. Location Awareness
4. **Explicit Support for Client-Server Traffic and Distributed Services**
5. Person-to-Person Communication
6. Security
7. **Control, Management, and Data Plane separation**
8. Isolation
9. Symmetric/Asymmetric Protocols
10. **Quality of Service**

*Nice to have:*
- Global Routing with Local Control of Naming and Addressing
- **Real Time Services**
- **Cross-Layer Communication**
- Manycast
- Receiver Control
- Support for Data Aggregation and Transformation
- **Support for Streaming Data**
- Virtualization

TimeLine

Cogni3ve Nets and clouds

Sustainable Internet

Machine Learning +

“I Want” Internet 3.0

Virtualized Internet

Good Old Trucking

Programmable Networks

NetApp’s

CineGrid

SF for CineGrid

NDL SF for complex nets

CineGrid

Virtualized Internet

OpenFlow

PBT/PLSB

(G)MPLS

TCP

LightPaths - GLIF

Hybrid Nets

TBN

Policy

NM

OCCI

NSI

“greenIT&Nets”

MDM, SCTCP, …

AAA

TBN

Policy

TimeLine

2005

2011

2020
Challenges

• Data – Data – Data
  – Archiving, publication, searchable, transport, self-describing, DB innovations needed, multi disciplinary use

• Virtualisation
  – Another layer of indeterminism

• Greening the Infrastructure
  – e.g. Department Of Less Energy: http://www.ecrinitiative.org/pdfs/ECR_3_0_1.pdf

• Disruptive developments
  – BufferBloath, Revisiting TCP, influence of SSD’s & GPU’s
  – Multi layer Glif Open Exchange model
  – Invariants in LightPaths (been there done that 😊)
    • X25, ATM, SONET/SDH, Lambda’s, MPLS-TE, VLAN’s, PBT, OpenFlow, ….
  – Authorization & Trust & Security and Privacy
The Way Forward!

• Nowadays scientific computing and data is dwarfed by commercial & cloud, there is also no scientific water, scientific power.
  • Understand how to work with elastic clouds
  • Trust & Policy & Firewalling on VM/Cloud level

• Technology cycles are 3 – 5 year
  • Do not try to unify but prepare for diversity
  • Hybrid computing & networking
  • Compete on implementation & agree on interfaces and protocols

• Limitation on natural resources and disruptive events
  • Energy becomes big issue
  • Follow the sun
  • Avoid single points of failure (aka Amazon, Blackberry, …)
  • Better very loosely coupled than totally unified integrated…
ECO-Scheduling

What type of route should be planned?

- Fastest route
- Eco route
- Shortest route
- Avoid motorways
- Walking route
Q & A

I did not talk about:
- CineGrid, digital Cinema on CI
- Knowledge complexity
- Security & privacy
- AAA
- ...

http://ext.delaat.net/

Slides thanks to:
- Paola Grosso
- Sponsors see slide 1. 😊
- SNE Team & friends, see below