Lambda-Grid developments
History - Present - Future

Cees de Laat

SURFnet
EU
BSIK
NWO
University of Amsterdam
Contents

1. The need for hybrid networking
2. StarPlane; a grid controlled photonic network
3. RDF/Network Description Language
4. Tera-networking
5. Programmable networks
LHC Data Grid Hierarchy
CMS as example, Atlas is similar

Tier 0 +1

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

Tier 2
Tier 2 Center

Tier 3
Institute

Tier 4
Physics data cache
Workstations

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.
Data intensive scientific computation through global networks

- Nuclear experiments
- Belle Experiments
- Nobeyama Radio Observatories (VLBI)
- X-ray astronomy Satellite ASUKA
- SUBARU Telescope
- Digital Sky Survey
- GRAPE6

Very High-speed Network

Data Reservoir

Data Reservoir

Local Accesses

Distributed Shared files

Data analysis at University of Tokyo
CineGrid@SARA
Sensor Grids

~ 40 Tbit/s

www.lofar.org

eVLBI

Longer term VLBI is easily capable of generating 8 Gbit/s or more. The sensitivity of the VLBI array scales with bandwidth (=data-rate) and there is a strong push to move to higher bandwidths. Rates of 8 Gbit/s or more are entirely feasible. The VLBI correlator is under development. It is expected that parallelized VLBI correlator will remain the most efficient approach to solving the VLBI problem: multi-gigabit correlator and parallelized correlator and multiplication factor.

Westerbork Synthesis Radio Telescope - Netherlands
The SCARIe project

**SCARIe**: a research project to create a Software Correlator for e-VLBI.

**VLBI Correlation**: signal processing technique to get high precision image from spatially distributed radio-telescope.

To equal the hardware correlator we need:

- 16 streams of 1Gbps
- $16 \times 1\text{Gbps}$ of data
- 2 Tflops CPU power
- $2 \text{Tflop} / 16 \text{Gbps} = 1000 \text{flops}/\text{byte}$

**THIS IS A DATA FLOW PROBLEM !!!**
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 2007
\[ \Sigma A = \Sigma B = \Sigma C \approx 250 \text{ Gb/s} \]
However:
- A -> all connects
- B -> on several
- C -> just a few (SP, LHC, LOFAR)
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 ≈ 2-3 k$/port  
L2 ≈ 5-8 k$/port  
L3 ≈ 75+ k$/port
Trends

• We have made baby-steps on the path to optical networking
  – Still many mails and phone calls
• See several trends:
  – lambda’s get fatter and cheaper
  – photonic technology cheap per bandwidth
  – embedded computation capacity increasing
  – latency and high bandwidth congestion avoidance conflict
  – ethernet is getting circuit properties (PBT)
  – applications need more and more predictable behaviour
How low can you go?

Application Endpoint A

Router

Ethernet

SONET

DWDM

Fiber

Local Ethernet

POS

MEMS

Regional dark fiber

15454 6500 HDXc

Trans-Oceanic

Application Endpoint B

NetherLight

UKLight

GLIF

StarLight

NetherLight
The playfield => GLIF
In The Netherlands SURFnet connects between 180:
- universities;
- academic hospitals;
- most polytechnics;
- research centers.
with an indirect ~750K user base

Red crosses = StarPlane

~ 6000 km scale comparable to railway system
Common Photonic Layer (CPL) in SURFnet6 supports up to 72 Lambda’s of 10 G each 40 G soon.
Contents

1. The need for hybrid networking
2. StarPlane; a grid controlled photonic network
3. RDF/Network Description Language
4. Tera-networking
5. Programmable networks
StarPlane
DWDM
backplane

SURFnet

university

WS+AAA

CdL
QOS in a non destructive way!

- **Destructive QOS:**
  - have a link or $\lambda$
  - set part of it aside for a lucky few under higher priority
  - rest gets less service

- **Constructive QOS:**
  - have a $\lambda$
  - add other $\lambda$’s as needed on separate colors
  - move the lucky ones over there
  - rest gets also a bit happier!
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
Dispersion compensating modem: eDCO from NORTEL
(Try to Google eDCO :-)

Solution in 5 easy steps for dummy’s:
1. try to figure out $T(f)$ by trial and error
2. invert $T(f) \rightarrow T^{-1}(f)$
3. computationally multiply $T^{-1}(f)$ with Fourier transform of bit pattern to send
4. inverse Fourier transform the result from frequency to time space
5. modulate laser with resulting $h'(t) = F^{-1}(F(h(t)).T^{-1}(f))$

(ps. due to power ~ square $E$ the signal to send looks like uncompensated received but is not)
The StarPlane vision is to give flexibility directly to the applications by allowing them to choose the logical topology in real time, ultimately with sub-second lambda switching times on part of the SURFnet6 infrastructure.
### Overview Net Tests between DAS-3 Hosts

- Authorize here to store the current table settings in your cookies file.
- See the getting started introduction or the user guide for a description of the table below.
- See also the hosts documentation.
- Some observations about the package and the required bandwidth.

Select ping value: min, avg, max, all, lost.
Select UDP value: ntr, lost.

#### DAS-3 Net Test Results

**Date:** 31/05/2007  
**Time:** 12:30:01

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>VU-083</td>
<td>0</td>
<td>0</td>
<td>0.087</td>
<td>0</td>
<td>0.013</td>
<td>0.017</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>VU-085</td>
<td>0</td>
<td>0</td>
<td></td>
<td>0.01</td>
<td>0.013</td>
<td>0.017</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>LIACS-125</td>
<td>0.087</td>
<td>0</td>
<td>0</td>
<td>0.013</td>
<td>0.017</td>
<td>0.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIACS-127</td>
<td>0</td>
<td>0</td>
<td>0.013</td>
<td>0.017</td>
<td>0.15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UvA-236</td>
<td>0</td>
<td>0</td>
<td>0.013</td>
<td>0.017</td>
<td>0.15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UvA-239</td>
<td>0</td>
<td>0</td>
<td>0.017</td>
<td>0.15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UvA-236-M</td>
<td>0</td>
<td>0</td>
<td>0.15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UvA-239-M</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Ping Min (ms)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>VU-083</td>
<td></td>
<td></td>
<td></td>
<td>1.380</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VU-085</td>
<td></td>
<td></td>
<td></td>
<td>1.380</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIACS-125</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.380</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIACS-127</td>
<td></td>
<td></td>
<td></td>
<td>1.380</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UvA-236</td>
<td>1.380</td>
<td></td>
<td>1.380</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UvA-239</td>
<td></td>
<td></td>
<td>1.380</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UvA-236-M</td>
<td></td>
<td></td>
<td>1.380</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UvA-239-M</td>
<td></td>
<td></td>
<td>1.380</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Throughput [Mbps]

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>VU-083</td>
<td></td>
<td></td>
<td></td>
<td>4884.22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VU-085</td>
<td></td>
<td></td>
<td></td>
<td>4884.22</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Very constant and predictable!
Contents

1. The need for hybrid networking
2. StarPlane; a grid controlled photonic network
3. RDF/Network Description Language
4. Tera-networking
5. Programmable networks
Architecture SC06
Network Description Language

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

  Subject ➔ Predicate ➔ Object

  - Location: name, connectedTo
  - Device: description, capacity
  - Interface: locatedAt, encodingType, encodingLabel
  - Link: hasInterface
The Modelling Process

Network Elements → Functional Elements → Syntax
<?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:504/1"/>
    <ndl:hasInterface rdf:resource="#tdm3.amsterdam1.netherlight.net:504/2"/>
    <ndl:hasInterface rdf:resource="#tdm3.amsterdam1.netherlight.net:504/3"/>
    <ndl:hasInterface rdf:resource="#tdm3.amsterdam1.netherlight.net:504/4"/>
    <!-- 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>
  </ndl:Device>
</rdf:RDF>
NDL Generator and Validator

[Image of NDL Generator and Validator interface]

**NDL for the GLIF - NDL Validator**

NDL - Network Description Language - is an ontology for description of (hybrid) networks, and provisioning. The GLIF collaboration makes use of NDL to describe each individual domain, maps.

This page will provide you with tools to validate an NDL file. We provide here two types of validation:

- Syntax validation
- Content validation

**Syntax validation**

We can validate that the NDL file you generated is written following the latest NDL schema. You will get back feedback on its validity.

Please paste your NDL file below:

```
<xml version="1.0" encoding="UTF-8" http://www.w3.org/1999/xhtml” xmlns="http://www.w3.org/2000/09/xmllang"
    xmlns:ndl="http://www.w3.org/2000/09/ndl"
    xmlns:geo="http://www.w3.org/2003/01/geo/wgs84_pos"
    xmlns:sif="http://www.w3.org/2000/09/sif"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  <ndl:ndlLocation rdf:about="#foo">
    <ndl:name rdf:resource="#name1"/>
    <ndl:location rdf:resource="#location1"/>
  </ndl:ndlLocation>
  <ndl:ndlDevice rdf:about="#dev1">
    <ndl:name rdf:resource="#name2"/>
    <ndl:device rdf:resource="#dev2"/>
  </ndl:ndlDevice>
</xml>
```

**Submit**

**Content validation**

Often NDL files reference information contained in other files managed by others. Such as for example when an interface on a local device connects to an interface to a remote device. The content validator performs a few basic checks to see that the information contained in cross-referencing NDL files is consistent.

Please enter the URL of the NDL file to be validated

```
Submit
```

See [http://trafficlight.uva.netherlight.nl/NDL-demo/](http://trafficlight.uva.netherlight.nl/NDL-demo/)
NDL SN6 Visualisation
Multi-layer extensions to NDL

End host

SONET switch
with Ethernet intf.

Ethernet &
SONET switch

SONET switch

SONET switch
with Ethernet intf.

End host

IP layer

Ethernet layer

STS layer

UTP layer

OC-192 layer

fiber layer
A weird example

- Université du Quebec
  - Gigabit Ethernet
  - can adapt GE in STS-24c

- StarLight Chicago
  - can adapt GE in STS-24c or STS-3c-7v
  - OC-192 (22 free)

- CA★Net Canada
  - 2x OC-192 (87 free)

- MAN LAN New York
  - OC-192 (38 free)
  - 2x OC-192 (63 free)

- Universiteit van Amsterdam
  - GE
  - can adapt GE in STS-3c-7v

- NetherLight Amsterdam
  - free
  - (free)
  - (free)
The result :-)
MultiDomain MultiLayer pathfinding in action
MultiDomain MultiLayer pathfinding in action
OGF NML-WG
Open Grid Forum - Network Markup Language workgroup

Chairs:
Paola Grosso – Universiteit van Amsterdam
Martin Swany – University of Delaware

Purpose:
To describe network topologies, so that the outcome is a standardized network description ontology and schema, facilitating interoperability between different projects.

https://forge.gridforum.org/sf/projects/nml-wg
RDF describing Infrastructure

Application: find video containing x, then trans-code to it view on Tiled Display
Contents

1. The need for hybrid networking
2. StarPlane; a grid controlled photonic network
3. RDF/Network Description Language
4. Tera-networking
5. Programmable networks
What constitutes a Tb/s network?

CALIT2 has 8000 Gigabit drops ?->? Terabit Lan?

look at 80 core Intel processor
  – cut it in two, left and right communicate 8 TB/s

think back to teraflop computing!
  – MPI makes it a teraflop machine

massive parallel channels in hosts, NIC’s

TeraApps programming model supported by
  – TFlops  ->  MPI / Globus
  – TBytes  ->  OGSA/DAIS
  – TPixels ->  SAGE
  – TSensors ->  LOFAR, LHC, LOOKING, CineGrid, ...
  – Tbit/s  ->  ?

ref Larry Smarr & CdL
Need for discrete parallelism

- it takes a core to receive 1 or 10 Gbit/s in a computer
- it takes one or two cores to deal with 10 Gbit/s storage
- same for Gigapixels
- same for 100’s of Gflops
- Capacity of every part in a system seems of same scale
- look at 80 core Intel processor
  - cut it in two, left and right communicate 8 TB/s
- massive parallel channels in hosts, NIC’s
- Therefore we need to go massively parallel allocating complete parts for the problem at hand!
User Programmable Virtualized Networks allows the results of decades of computer science to handle the complexities of application specific networking.

- The network is virtualized as a collection of resources
- UPVNs enable network resources to be programmed as part of the application
- Mathematica, a powerful mathematical software system, can interact with real networks using UPVNs
Mathematica enables advanced graph queries, visualizations and real-time network manipulations on UPVNs

Topology matters can be dealt with algorithmically
Results can be persisted using a transaction service built in UPVN

Initialization and BFS discovery of NEs

```mathematica
Needs["WebServices""]
<<DiscreteMath`Combinatorica`
<<DiscreteMath`GraphPlot

InitNetworkTopologyService["edge.ict.tno.nl"]

Available methods:
{DiscoverNetworkElements, GetLinkBandwidth, GetAllIpLinks, Remote, NetworkTokenTransaction}

Global`upvnverbose = True;

AbsoluteTiming[nes = BFSDiscover["139.63.145.94"];
result = BFSDiscoverLinks["139.63.145.94", nes];]

Getting neighbours of: 139.63.145.94
Internal links: {192.168.0.1, 139.63.145.94}

(...)

Global`upvnverbose = True;

AbsoluteTiming[nes = BFSDiscover["139.63.145.94"];
result = BFSDiscoverLinks["139.63.145.94", nes];]

Getting neighbours of: 192.168.2.3
Internal links: {192.168.2.3}

Transaction on shortest path with tokens

nodePath = ConvertIndicesToNodes[
  ShortestPath[ g, 
    Node2Index[nids, "192.168.3.4"],
    Node2Index[nids, "139.63.77.49"], nids]]

Path: {192.168.3.4, 192.168.3.1, 139.63.77.30, 139.63.77.49}
Committed
```

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.
Functional building blocks

Network layers

Application layers

Use Interface

Control Interface (protocols API’s)

Network Service

Application
Power is a big issue

• UvA cluster uses (max) 30 kWh
• 1 kWh ~ 0.1 €
• per year -> 26 k€/y
• add cooling 50% -> 39 k€/y
• Emergency power system -> 50 k€/y
• per rack 10 kWh is now normal
• **YOU BURN ABOUT HALF THE CLUSTER OVER ITS LIFETIME!**

• Terminating a 10 Gb/s wave costs about 200 W
• Entire loaded fiber -> 16 kW
• Wavelength Selective Switch : few W!
Questions?

I did not talk about StarPlane...