GNARP 2009:
Optical Networks for e-Science

Cees de Laat
GLIF.is founding member

SURFnet
EU
BSIK
NWO
University of Amsterdam
LOFAR as a Sensor Network

- LOFAR is a large distributed research infrastructure:
  - Astronomy:
    - >100 phased array stations
    - Combined in aperture synthesis array
    - 13,000 small “LF” antennas
    - 13,000 small “HF” tiles
  - Geophysics:
    - 18 vibration sensors per station
    - Infrasound detector per station
  - >20 Tbit/s generated digitally
  - >40 Tflop/s supercomputer
  - innovative software systems
    - new calibration approaches
    - full distributed control
    - VO and Grid integration
    - datamining and visualisation
The SCARle project

SCARle: 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 * 1Gbps of data
- 2 Tflops CPU power
- 2 TFlop / 16 Gbps = 1000 flops/byte

THIS IS A DATA FLOW PROBLEM !!!
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 Sloot

Many thanks to:
AMC
SARA
GigaPort
UvA/AIR
Silicon Graphics, Inc.
Zoölogisch Museum
Keio/Calit2 Collaboration: Trans-Pacific 4K Teleconference

Like High-Def? Here Comes the Next Level

By JOHN MARKOFF
Published: September 26, 2005

Keio University
President Anzai

UCSD
Chancellor Fox

Used
1Gbps Dedicated

Sony
NTT
SGI

Keio University
President Anzai

UCSD
Chancellor Fox

iGrid 2005
Formats - Numbers - Bits
<table>
<thead>
<tr>
<th>Format</th>
<th>X</th>
<th>Y</th>
<th>Rate/s</th>
<th>Color bits/pix</th>
<th>Frame pix</th>
<th>Frame MByte</th>
<th>Flow MByte/s</th>
<th>Stream Gbit/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>720p HD</td>
<td>1280</td>
<td>720</td>
<td>60</td>
<td>24</td>
<td>921600</td>
<td>2.8</td>
<td>170</td>
<td>1.3</td>
</tr>
<tr>
<td>1080p HD</td>
<td>1920</td>
<td>1080</td>
<td>30</td>
<td>24</td>
<td>2073600</td>
<td>6.2</td>
<td>190</td>
<td>1.5</td>
</tr>
<tr>
<td>2k</td>
<td>2048</td>
<td>1080</td>
<td>24</td>
<td>36</td>
<td>2211840</td>
<td>10</td>
<td>240</td>
<td>1.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>48</td>
<td></td>
<td></td>
<td></td>
<td>480</td>
<td>2.4</td>
</tr>
<tr>
<td>SHD</td>
<td>3840</td>
<td>2160</td>
<td>30</td>
<td>24</td>
<td>8294400</td>
<td>25</td>
<td>750</td>
<td>6.0</td>
</tr>
<tr>
<td>4k</td>
<td>4096</td>
<td>2160</td>
<td>24</td>
<td>36</td>
<td>8847360</td>
<td>40</td>
<td>960</td>
<td>7.6</td>
</tr>
</tbody>
</table>

Note: this is excluding sound!
Note: these are raw uncompressed data rates ex overhead!
## Buffer space

Window = RTT * BW

<table>
<thead>
<tr>
<th>RTT</th>
<th>100 Mbit/s</th>
<th>1 Gbit/s</th>
<th>10 Gbit/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12.5 kB</td>
<td>125 kB</td>
<td>1.25 MB</td>
</tr>
<tr>
<td>2</td>
<td>25 kB</td>
<td>250 kB</td>
<td>2.5 MB</td>
</tr>
<tr>
<td>5</td>
<td>62.5 kB</td>
<td>615 kB</td>
<td>6.15 MB</td>
</tr>
<tr>
<td>10</td>
<td>125 kB</td>
<td>1.25 MB</td>
<td>12.5 MB</td>
</tr>
<tr>
<td>20</td>
<td>250 kB</td>
<td>2.5 MB</td>
<td>25 MB</td>
</tr>
<tr>
<td>50</td>
<td>625 kB</td>
<td>6.25 MB</td>
<td>62.5 MB</td>
</tr>
<tr>
<td>100</td>
<td>1.25 MB</td>
<td>12.5 MB</td>
<td>125 MB</td>
</tr>
<tr>
<td>200</td>
<td>2.5 MB</td>
<td>25 MB</td>
<td>250 MB</td>
</tr>
<tr>
<td>500</td>
<td>6.25 MB</td>
<td>62.5 MB</td>
<td>625 MB</td>
</tr>
<tr>
<td>1000</td>
<td>12.5 MB</td>
<td>125 MB</td>
<td>1250 MB</td>
</tr>
</tbody>
</table>
Welcome to the Amsterdam CineGrid distribution node. Below are the latest additions of super-high-quality video to our node.

For more information about CineGrid and our efforts look at the about section.

Latest Additions

Wypke

Available formats:
- 4k dot (4.6 KB)

Duration: 1 hour and 8 minutes
Created: 1 week, 2 days ago
Author: Wypke
Categories:

Prague Train

Available formats:
- 4k dot (3.9 KB)

Duration: 27 hours and 46 minutes
Created: 1 week, 2 days ago
Author: CineGrid
Categories: delta prague train

VLC: Big Buck Bunny

Available formats:
- 1080p MPEG4 (1.1 GB)

Duration: 1 hour and 9 minutes
Created: 1 month, 5 weeks ago
Author: Blender Foundation
Categories: animation Blender bunny CGI
Amsterdam CineGrid S/F node

“COCE”

Rembrandt Cluster
total 22 TByte disk space
@ LightHouse

DP AMD processor nodes

Comp node

Head node

Bridge node

Bridge node

Bridge node

Bridge node

Bridge node

GlimmerGlass

Photonic switch

NetherLight, StarPlane
the cp testbeds
and beyond

10 Gbit/s

Opteron 64 bit nodes

Head node

Comp node

Comp node

Comp node

Comp node

Comp node

Comp node

Comp node

Comp node

Comp node

NORTEL

8600

L2/3 switch

F10

L2/3 switch

Streaming node

8 TByte

Node 41

Suitcees &

Briefcees

100 TByte

Storage node

10 Gbit/s

10 Gbit/s
Sensor grid: instrument the dikes

First controlled breach occurred on sept 27th ‘08:

30000 sensors (microphones) to cover all Dutch dikes
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

A = B = C

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
  - 0.5 W/port

- L2 ≈ 5-8 k$/port
  - 10-15 W/port

- L3 ≈ 75+ k$/port
  - 250 W/port
How low can you go?

- Application Endpoint A
- Local Ethernet
- POS
- MEMS
- Regional dark fiber
- 15454 6500 HDXc
- Trans-Oceanic
- Application Endpoint B
- Router
- Ethernet
- SONET
- DWDM
- Fiber
- NetherLight
- UKLight
- GLIF
- StarLight
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
SURFnet6 supports up to 72 Lambda’s of 10 G each. 40 G soon.
StarPlane
DWDM
backplane

SURFnet

WS+AAA

university

CdL
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
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 $h'(t) = F^{-1}(F(h(t)).T^{-1}(f))$

(p.s. due to power ~ square $E$ the signal to send looks like uncompensated received but is not)
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!
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.
GLIF 2008

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

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

Choice of RDF instead of XML syntax
Grounded modeling based on G0805 description:

A weird example

Université du Québec

StarLight
Chicago

can adapt GE in
STS-24c or STS-3c-7v

Gigabit Ethernet

CA★Net
Canada

can adapt GE
in STS-24c

OC-192
(22 free)

2x OC-192
(87 free)

MAN LAN
New York

OC-192
(38 free)

2x OC-192
(63 free)

Universiteit van Amsterdam

NetherLight
Amsterdam

can adapt GE
in STS-3c-7v

GE

The result :-)
- Switching matrix

- Channels

- A device switches data based on:
  - The source interface
  - One or more labels

- Example label types:
  - Ethernet VLAN
  - SONET STS Channel
  - Wavelength ($\lambda$)

- For example, all data from channel 31 of interface 2 is forwarded to channel 28 of interface 4.
ITU-T G.805 describes functional elements (e.g. adaptation, termination functions, link connections, etc.) to describe **network connections**.

- We extended these function elements (e.g. with potential adaptation functions) to describes **networks**.
- We created a model to map actual network elements to functional elements.
- Defined a simple algebra to define correctness of a connection
- We created a NDL extension to describe these functional elements.
Multi-layer extensions to NDL
RDF describing Infrastructure

“*I want*”

Application: find video containing x, then trans-code to it view on Tiled Display
Semantic Reasoning
The Problem

I want HC and AB
Success depends on the order
Wouldn’t it be nice if I could request [HC, AB, ...]
NDL + PROLOG

Research Questions:
• order of requests
• complex requests
• Usable leftovers

• Reason about graphs
• Find sub-graphs that comply with rules
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

Topologies can be dealt with algorithmically

Results can be persisted using a transaction service built in UPVN

Initialization and BFS discovery of NEs

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

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

Available methods:

(DiscoverNetworkElements, GetLinkBandwidth, GetAllIpLinks, Remote, NetworkTokenTransaction)

Global`upvnverbose = True;

AbsoluteTiming[nes = BFSDiscover["139.63.145.94"];];[[1]]

AbsoluteTiming[result = BFSDiscoverLinks["139.63.145.94", nes];];[[1]]

Getting neighbours of: 139.63.145.94

Internal links: {192.168.0.1, 139.63.145.94}

 (...)

Getting neighbours of: 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];

Print["Path: ", nodePath];

If[NetworkTokenTransaction[nodePath, "green"] == True,
  Print["Committed"], Print["Transaction failed"];

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.
TeraThinking

- 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
TouchTable Demonstration @ SC08
Interactive programmable networks
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!
Multi Layer Service Architecture

- Network layers
  - Use Interface
  - Control Interface (protocols API’s)

- Application layers
  - Application
    - Network Service
Questions?

Accepted paper: A Declarative Approach to Multi-Layer Path Finding Based on Semantic Network Descriptions.
Not on the memory stick, so:

http://delaat.net/~delaat/papers/declarative_path_finding.pdf

Thanks: Paola Grosso & Jeroen vd Ham & Freek Dijkstra & team for several of the slides.