ClearStream

Prototyping 40 Gbps Transparent End-to-End Connectivity

Cosmin Dumitru
Ralph Koning
Cees de Laat
and many others (see posters)

University of Amsterdam
Internet developments

more data!

Speed
Volume

Deterministic
Real-time

Scalable
Secure

... more users!
<table>
<thead>
<tr>
<th>Category</th>
<th>Ijkjejik/Urban Flood</th>
<th>Medical</th>
<th>LifeWatch</th>
<th>EU-GN3/NOVI/Gyseyers</th>
<th>SURFnet/GLIF/Cloud</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green-IT</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Privacy/Trust</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Authorization/policy</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</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></td>
<td></td>
</tr>
</tbody>
</table>

- **Speed**:  
  - Volume

- **Deterministic**:  
  - Real-time

- **Scalable**:  
  - Secure
<table>
<thead>
<tr>
<th>Speed</th>
<th>Volume</th>
<th>Deterministic</th>
<th>Real-time</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Green-IT</th>
<th>Privacy/Trust</th>
<th>Authorization/policy</th>
<th>Programmable networks</th>
<th>40-100Gig/TCP/WF/QoS</th>
<th>Topology/Architecture</th>
<th>Optical Photonic</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

SNE @ UvA
<table>
<thead>
<tr>
<th>Green-IT</th>
<th>Privacy/Trust</th>
<th>Authorization/policy</th>
<th>Programmable networks</th>
<th>40-100Gig/TCP/WF/QoS</th>
<th>Topology/Architecture</th>
<th>Optical Photonic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>X X</td>
<td>X X X X</td>
<td>X X</td>
<td>X X X X</td>
<td>X X</td>
<td>X X X</td>
</tr>
</tbody>
</table>
Big and small flows don’t go together on the same wire!
Goals

• To demonstrate single stream single wave performance end to end operating without any obstruction.

• To break the 10 Gbps barrier.

• To test a real application on this infrastructure. DiVinE
Setup

UvA

iPerf
2 quad core
Mellanox
40G
Extreme
CIENA OME6500

DiViNe
48 core
DELL

CERN

iPerf
2 quad core
Mellanox
CV
Extreme
CIENA OME6500

DiViNe
48 core

Amsterdam – Geneva (CERN) – 1650KM (~1000Miles)
What we demonstrated

- Single server to single server performance memory to memory
  - Single stream single Lambda TCP
  - Multiple stream single Lambda TCP
  - UDP streaming
- 48 core system to 48 core system
  - Running DiViNe model checker
  - Many small messages
  - Cluster in a box!
## Servers

<table>
<thead>
<tr>
<th>Model</th>
<th>Supermicro X8DTT-HIBQF</th>
<th>Dell R815</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU</td>
<td>2 x Quad-Core Intel XEON E5620 2.4GHz</td>
<td>4 x Twelve-Core AMD Opteron 6172 2.1GHz</td>
</tr>
<tr>
<td>RAM</td>
<td>24GB</td>
<td>128GB (4 x 32GB)</td>
</tr>
<tr>
<td>NIC</td>
<td>Mellanox ConnectX2 40GE</td>
<td>Mellanox ConnectX2 40GE</td>
</tr>
<tr>
<td>OS</td>
<td>Linux 2.6.32</td>
<td>Linux 2.6.18</td>
</tr>
</tbody>
</table>
LAN Setup

QSFP+ Active Multimode fiber
40GBASE-SR4 – 4 x 10Gbit/s
MLD – Multi Lane Distribution
4 fibers for RX
4 fibers for TX
Synchronization is done at the optical level

Mellanox Connect X2 40GE
PCI-E 2.0 8x
WAN Setup

DWDM system
In production in the SURFnet network
Ethernet packed in OTU 3 frame

Ciena ActiveFlex (OME) 6500
40GE CFP Module
“Testcases”

• Single server to single server performance memory to memory
  – Single stream single Lambda TCP
  – Multiple stream single Lambda TCP

• 48 core system to 48 core system
  – Running the DiVinE model checker
  – Already used by VU University Amsterdam to test the 100G link to Hamburg
  – state space explosion problem
  – Many small messages (~400MBbit/core)
  – Cluster in a box!
Preliminary results

- 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
Calculated TCP window size – **40.5** MB for 19Gbit sustained throughput link -
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
CPU Topology benchmark

We used numactl to bind iperf to cores
Supercomputing 2010

Dutch – Research Consortium booth-to-booth
40GE demonstration
Demo setup codename: Flightcees

Ciena ActiveFlex (OME) 6500

Broadcom 40GE 18 port L2 Ethernet Switch

Supermicro Intel Server

Dell R815 Server
Live stats - Supercomputing 2010
Innovative 100G solutions unlock your...
Achievements

- ~19 Gbps Core to Core throughput (single flow iperf)
  - We need a faster CPU core
- ~25 Gbps CPU to CPU throughput (multi flow iperf)
  - PCI-E bottleneck
  - 22Gbps full duplex
- Broke the 10G barrier using a real world application: DiVinE: peaks of 11Gbps
- 40G “pipe” filled with just 2 servers
- Demonstrated that 40G Ethernet can be transported over long distance
Hybrid Networking <-> computing

Routers ↔ Supercomputers

Ethernet switches ↔ Grid & Cloud

Photonic transport ↔ GPU’s

What matters:

Energy consumption/multiplication
Energy consumption/bit transported
Utilizing shared expertise in advanced photonic, leading edge hardware and high-performance computing, the team created a network application testbed using the 1650 km Cross Border Fiber between NetherLight and CERNLight, lit by SURFnet, connecting servers equipped with 40 Gigabit Ethernet network interface at the University of Amsterdam to remote servers with corresponding interfaces at GLIF 2010 in Geneva.

Network Setup
The Mellanox ConnectX-2 EN 40GbE is the first network interface that allows single stream ethernet transport far exceeding the common 10Gbps boundary limit. The achieved throughput is 26Gbps from CPU to CPU which is the practical limit of the PCI-E interface.

The network infrastructure is based on Ciena’s Optical Multiservice Edge (OME) 6500 equipped with 40 GbE interfaces, which enables data speeds to be seamlessly upgraded from 10 Gbps to 40 Gbps.

Application Setup @Supercomputing 2010
Following the success of the GLIF 2010 demo, the Supercomputing 2010 setup demonstrates two high performance servers fully utilizing the 40Gbps clear channel WAN link between the Ciena Booth and the Dutch Research booth.

Going beyond 10 Gbps leads to new challenges in applications, operating system tuning and system architecture design as new bottlenecks appear.

Special attention needs to be given to the setup of multi-core machines in order to have the best I/O performance and maximize the network throughput. During the demo the PCI-E x8 2.0 interface of the network card is saturated when using UDP or TCP traffic.

High Performance Node
Using a flexible I/O architecture, the Supermicro X8DTT with two quad-core Intel E5620 CPUs, allows extreme speeds of over 25 Gbps to be reached.

GLIF 2010 Demo
During the GLIF 2010 demonstration measurements showed constant throughput between the two remote ends. Using two servers over 70Gbps of aggregated traffic was exchanged in both directions.
Questions?

It is a bit freaky with this wireless technology.

UvA
Cees de Laat
Ralph Koning
Cosmin Dumitru

SURFnet
Erik-Jan Bos
Gerben van Malenstei
Roeland Nuijts

Ciena
David Yeung
Harry Peng
Martin Bluethner

VU
Kees Verstoep
Henri Bal

Mellanox
Bill Lee
Erez Cohen