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

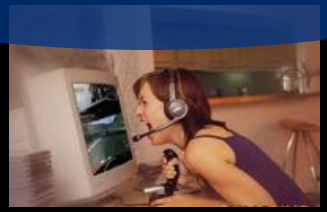
Real-time

more realtime!

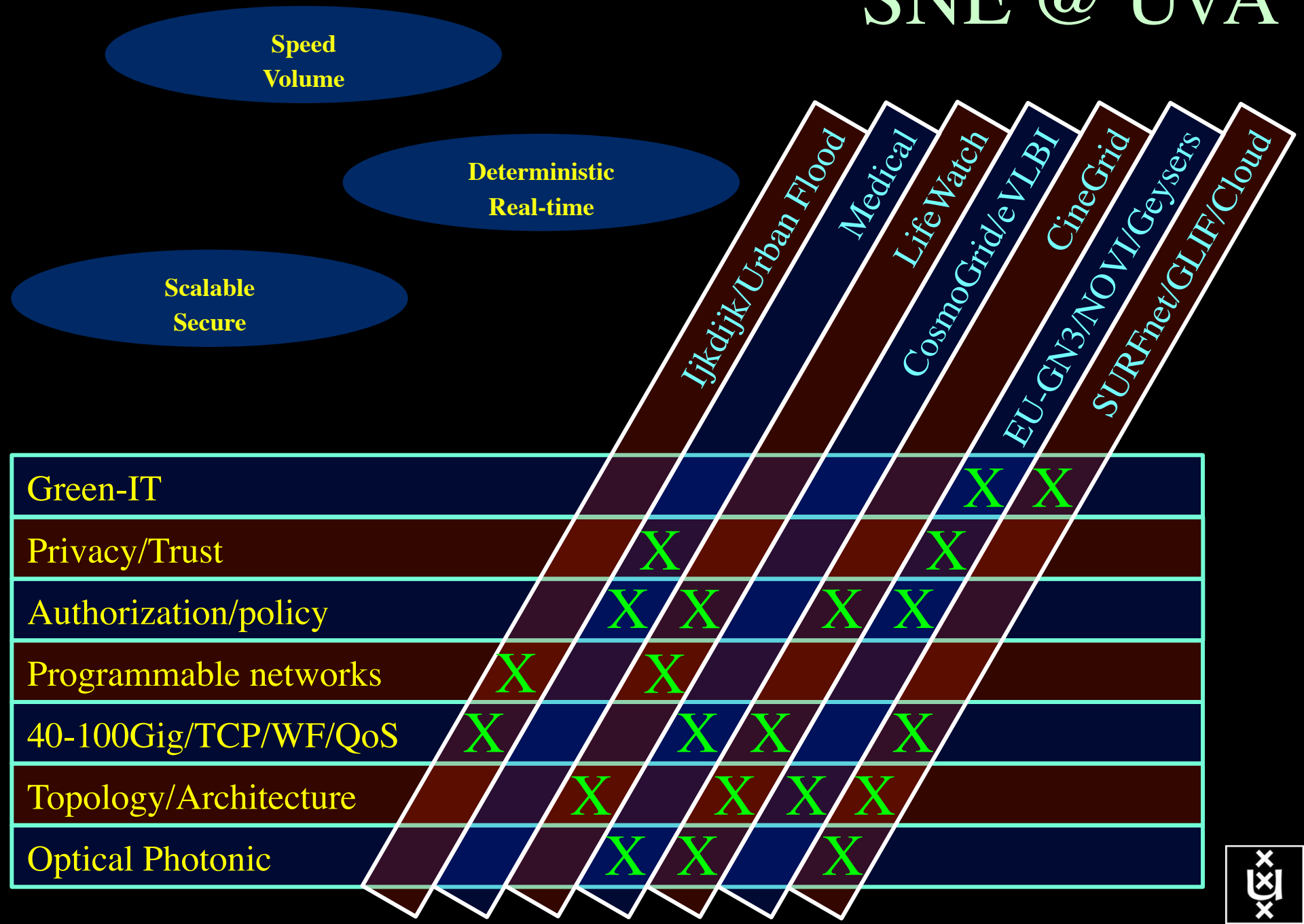
Scalable

Secure

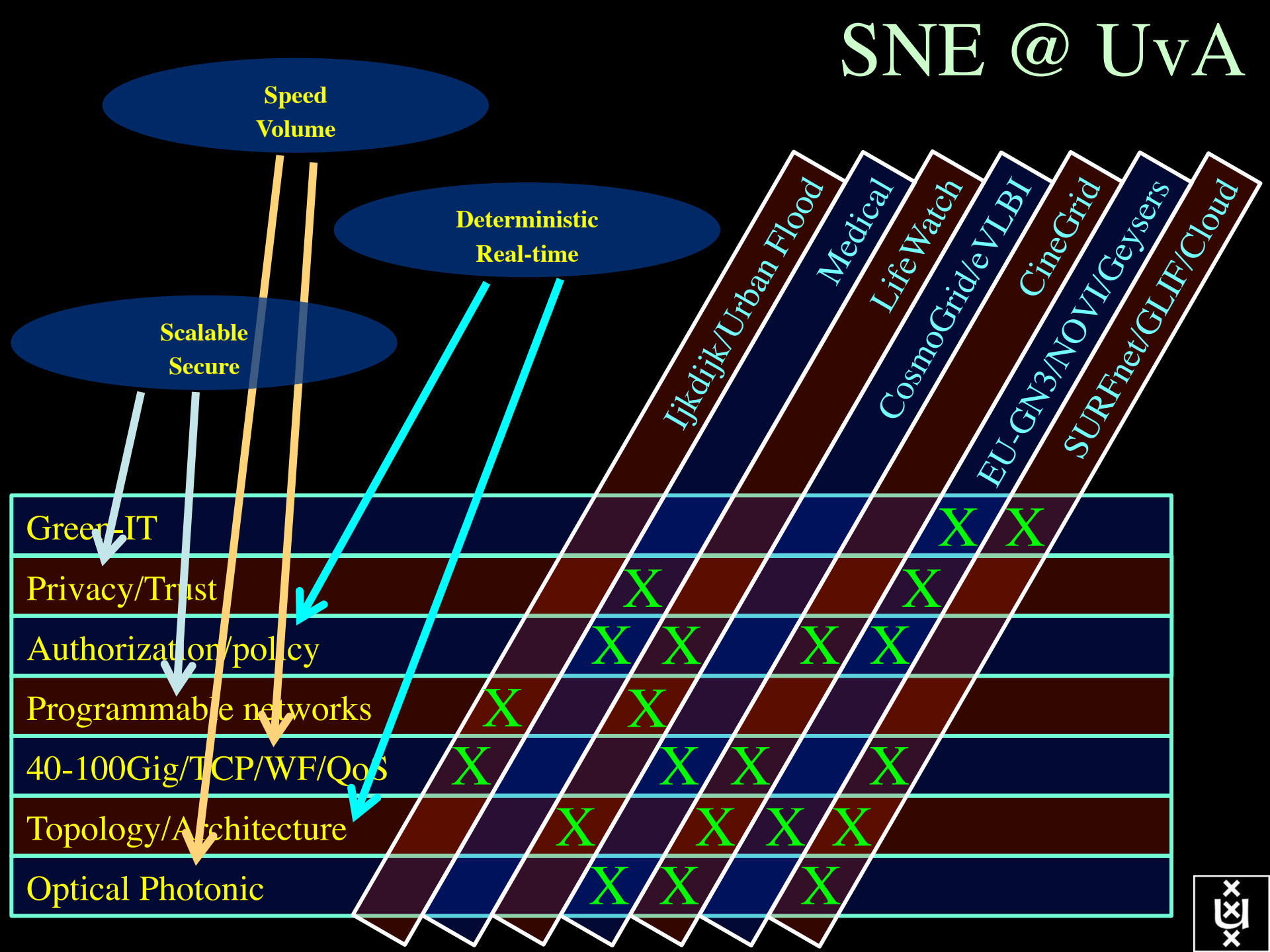
... more users!

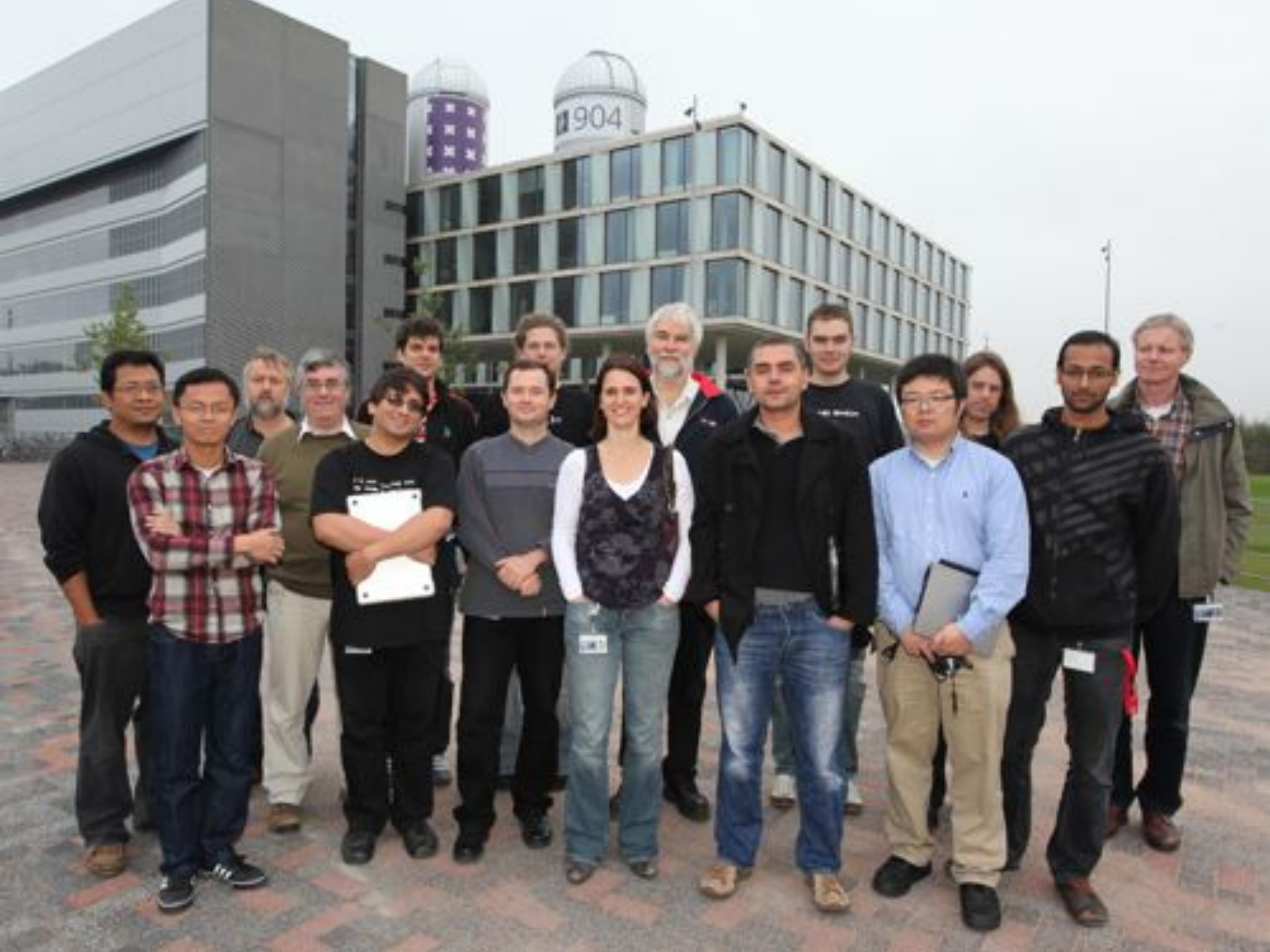


SNE @ UvA

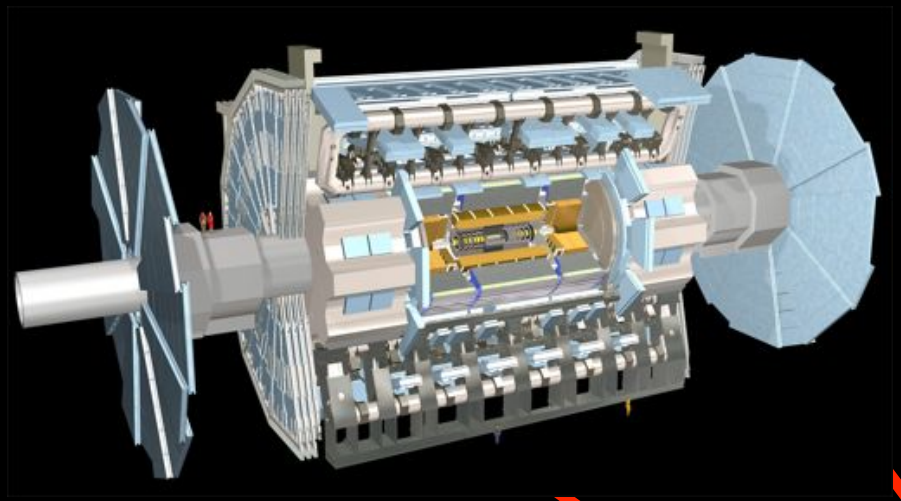

Speed
Volume
Deterministic
Real-time
Scalable
Secure
Ijkdijk/Urban Flood
Medical
LifeWatch
CosmoGrid/eVLBI
CineGrid
EU-GN3/NOVI/Geysers
SURFnet/GLIF/Cloud
Green-IT
Privacy/Trust
Authorization/policy
Programmable networks
40-100Gig/TCP/WF/QoS
Topology/Architecture
Optical Photonic
X

SNE @ UvA


Speed
Volume
Deterministic
Real-time
Scalable
Secure
Ijkdijk/Urban Flood
Medical
LifeWatch
CosmoGrid/eVLBI
CineGrid
EU-GN3/NOVI/Geysers
SURFnet/GLIF/Cloud
Green-IT
Privacy/Trust
Authorization/policy
Programmable networks
40-100Gig/TCP/WF/QoS
Topology/Architecture
Optical Photonic


904

SNE @ UvA



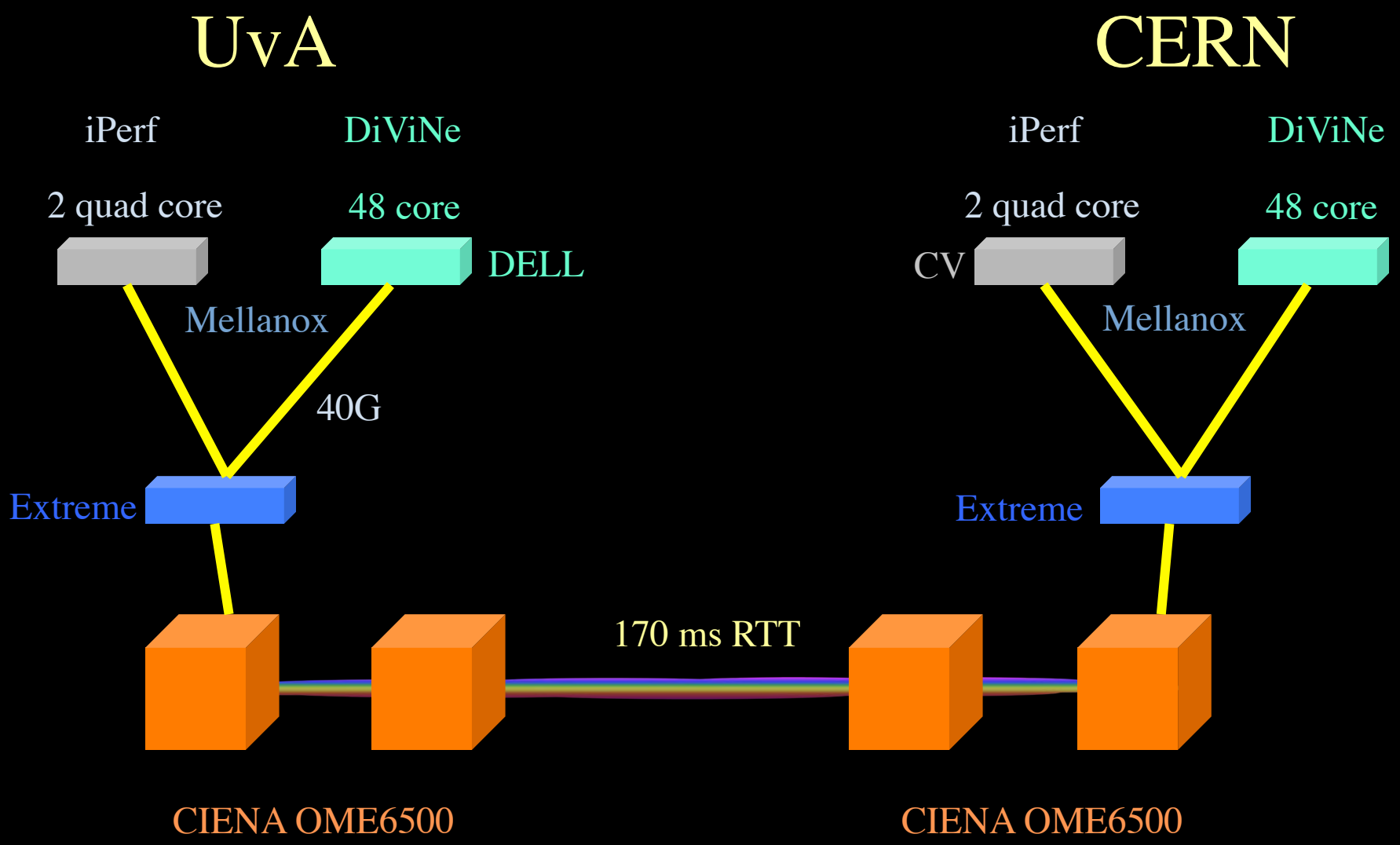
	Ijkdijk/Urban Flood	Medical	LifeWatch	CosmoGrid/eVLBI	CineGrid	EU-GN3/NOVI/Geysers	SURFnet/GLIF/Cloud
Green-IT					X	X	
Privacy/Trust		X			X		
Authorization/policy		X	X		X	X	
Programmable networks	X		X				
40-100Gig/TCP/WF/QoS	X		X	X		X	
Topology/Architecture		X		X	X	X	
Optical Photonic		X	X		X		

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



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

Model	Supermicro X8DTT-HIBQF	Dell R815
CPU	2 x Quad-Core Intel XEON E5620 2.4GHz	4 x Twelve-Core AMD Opteron 6172 2.1GHz
RAM	24GB	128GB (4 x 32GB)
NIC	Mellanox ConnectX2 40GE	Mellanox ConnectX2 40GE
OS	Linux 2.6.32	Linux 2.6.18

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

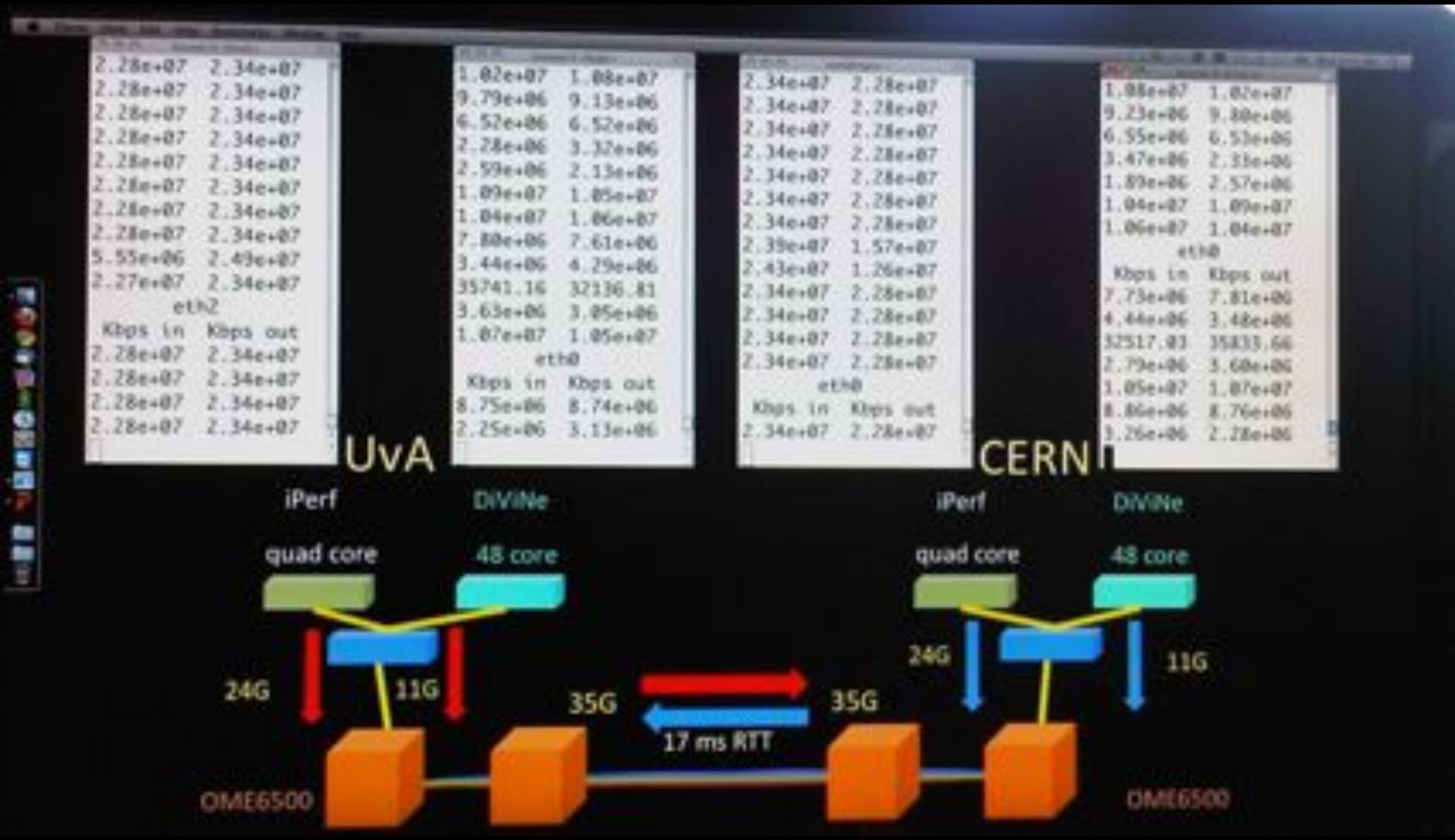
Ethernet packed in OTU 3 frame

Ciena ActiveFlex (OME) 6500

40GE CFP Module

“Testcases”

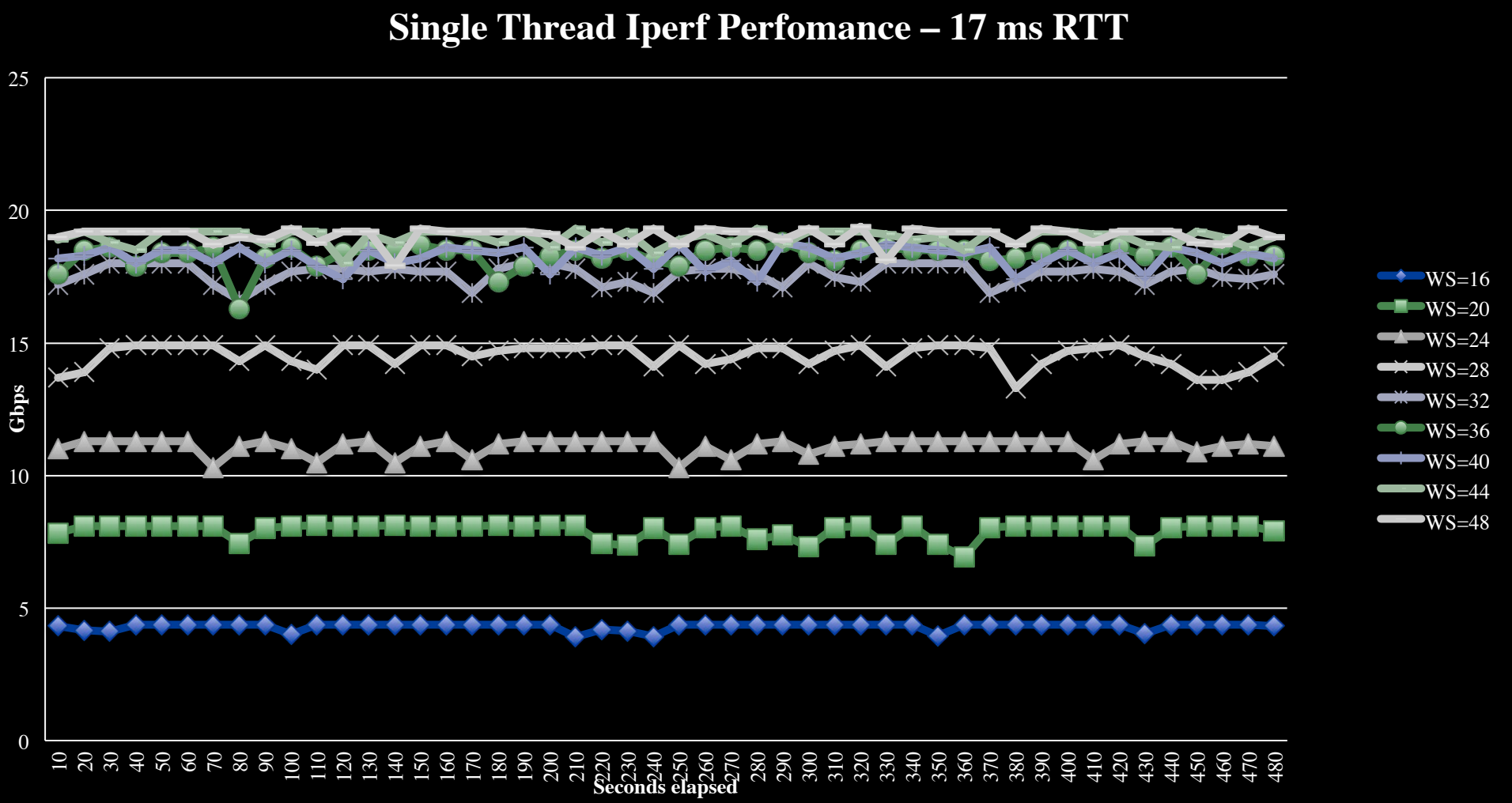
- Single server to single server performance memory to memory
  - Single stream single Lambda TCP
  - Multiple stream single Lambda TCP
- 48 core system to 48 core system
  - Running the DiVinE model checker
  - Already used by VU University Amsterdam to test the 100G link to Hamburg
  - state space explosion problem
  - Many small messages (~400MBbit/core)
  - Cluster in a box!


2.28e+07 2.34e+07
2.28e+07 2.34e+07
2.28e+07 2.34e+07
2.28e+07 2.34e+07
2.28e+07 2.34e+07
2.28e+07 2.34e+07
2.28e+07 2.34e+07
2.28e+07 2.34e+07
5.55e+06 2.49e+07
2.27e+07 2.34e+07
eth2
Kbps in Kbps out
2.28e+07 2.34e+07
2.28e+07 2.34e+07
2.28e+07 2.34e+07
2.28e+07 2.34e+07
1.02e+07 1.08e+07
9.79e+06 9.13e+06
6.52e+06 6.52e+06
2.28e+06 3.32e+06
2.59e+06 2.13e+06
1.09e+07 1.05e+07
1.04e+07 1.06e+07
7.80e+06 7.61e+06
3.44e+06 4.29e+06
35741.16 32136.81
3.63e+06 3.05e+06
1.07e+07 1.05e+07
eth0
Kbps in Kbps out
8.75e+06 8.74e+06
2.25e+06 3.13e+06
UvA
iPerf
DiViNe
quad core
48 core
24G
11G
35G
35G
17 ms RTT
OME6500
2.34e+07 2.28e+07
2.34e+07 2.28e+07
2.34e+07 2.28e+07
2.34e+07 2.28e+07
2.34e+07 2.28e+07
2.34e+07 2.28e+07
2.34e+07 2.28e+07
2.35e+07 1.57e+07
2.43e+07 1.26e+07
2.34e+07 2.28e+07
2.34e+07 2.28e+07
2.34e+07 2.28e+07
2.34e+07 2.28e+07
eth0
Kbps in Kbps out
2.34e+07 2.28e+07
1.08e+07 1.02e+07
9.23e+06 9.80e+06
6.55e+06 6.53e+06
3.47e+06 2.33e+06
1.89e+06 2.57e+06
1.04e+07 1.09e+07
1.06e+07 1.04e+07
eth0
Kbps in Kbps out
7.73e+06 7.81e+06
4.44e+06 3.48e+06
32517.03 35833.66
2.79e+06 3.60e+06
1.05e+07 1.07e+07
8.86e+06 8.76e+06
3.26e+06 2.28e+06
CERN
iPerf
DiViNe
quad core
48 core
24G
11G
OME6500

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

GLIF 2010 Measurements

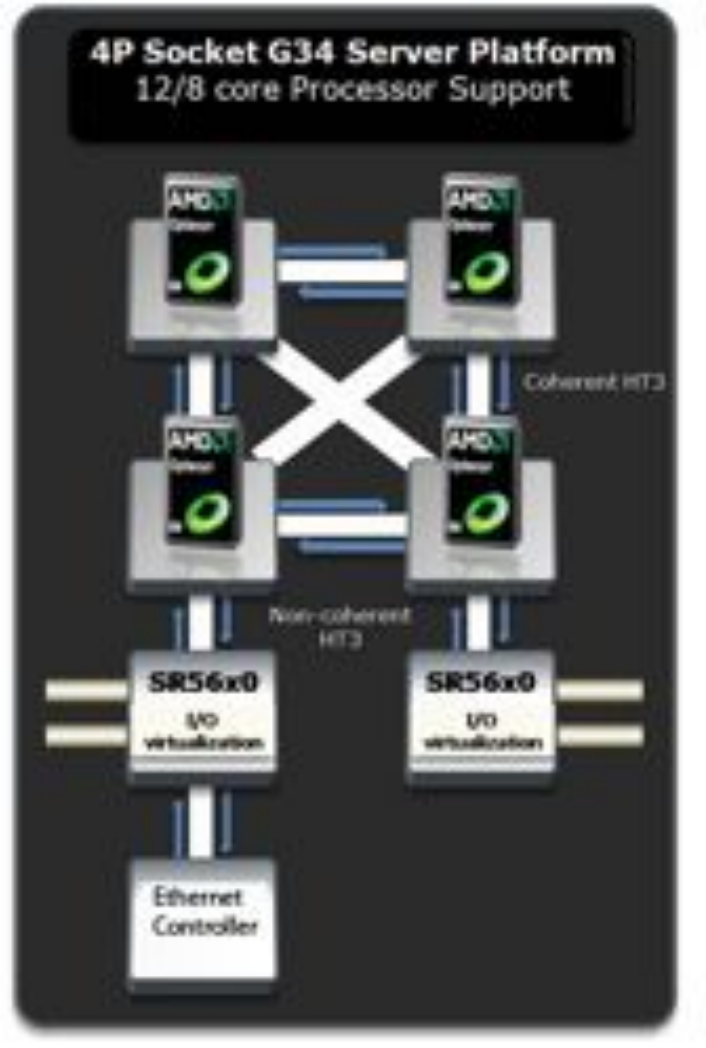


Calculated TCP window size – **40.5** MB for 19Gbit sustained throughput link -

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



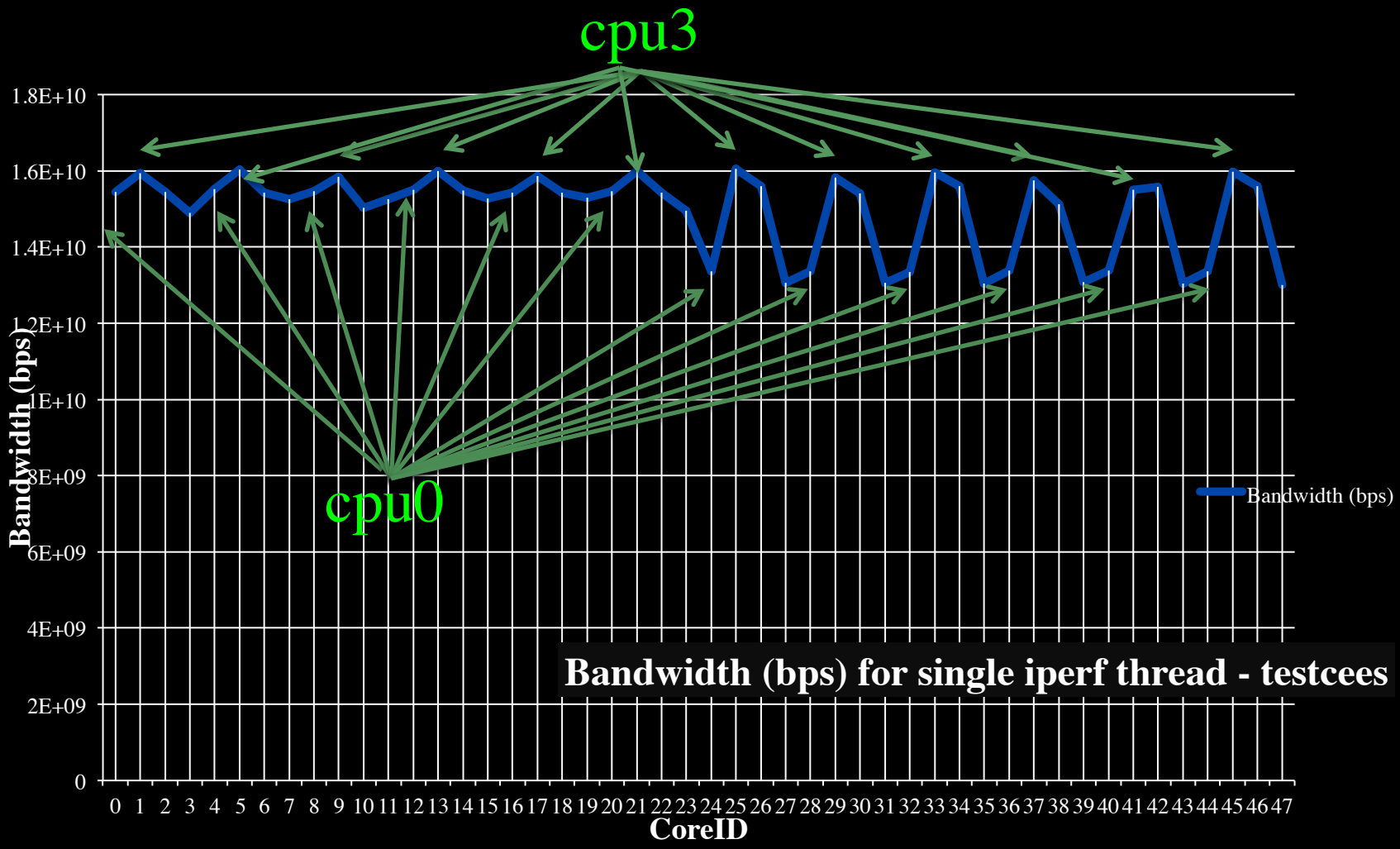
DELL R815

4 x AMD Opteron 6100

Supermicro X8DTT-HIBQF

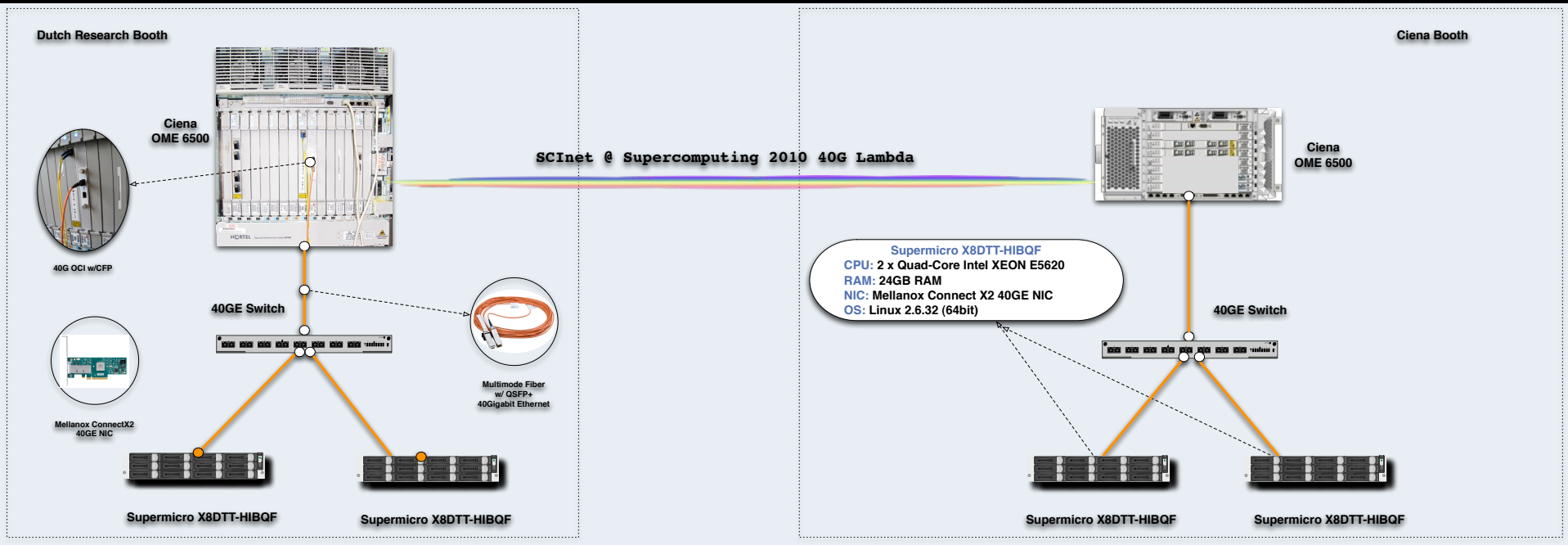
2 x Intel Xeon

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

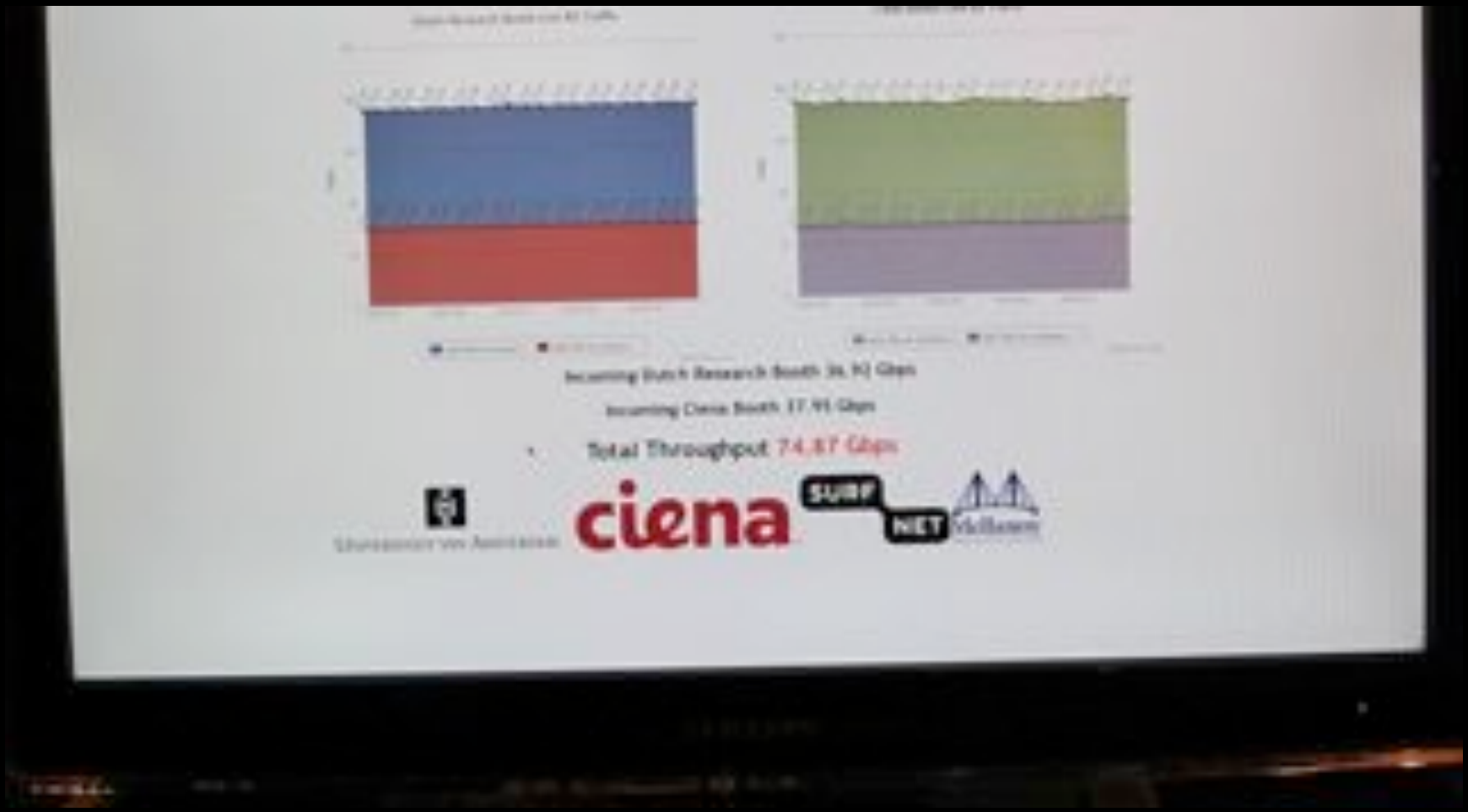
Broadcom 40GE 18 port L2 Ethernet Switch

Supermicro Intel Server

Dell R815 Server


NASA

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

delaat.net/posters

ClearStream

End-to-End Ultra Fast Transmission Over a Wide Area 40 Gbit/s Lambda

University of Amsterdam
Cosmin Dumitru
Cees de Laat
Ralph Koning
SURFnet
Erik-Jan Bos
Gerben van Malenstein
Roeland Nuijts
Ciena
David Yeung
Jan-Willem Elion
Harry Peng
Kevin McKernan
Martin Bluethner
VU University Amsterdam
Kees Verstoep
Henri Bal
Mellanox
Erez Cohen

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 throughtput 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

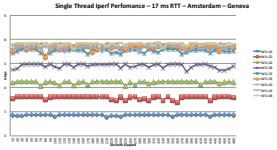
Following the succes of the GLIF 2010 demo, the Supercomputing 2010 setup demonstrates two high performance servers fully utilizing the 40Gbps clear channel WAN link between the Ciena Booth and the Dutch Research booth.

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.

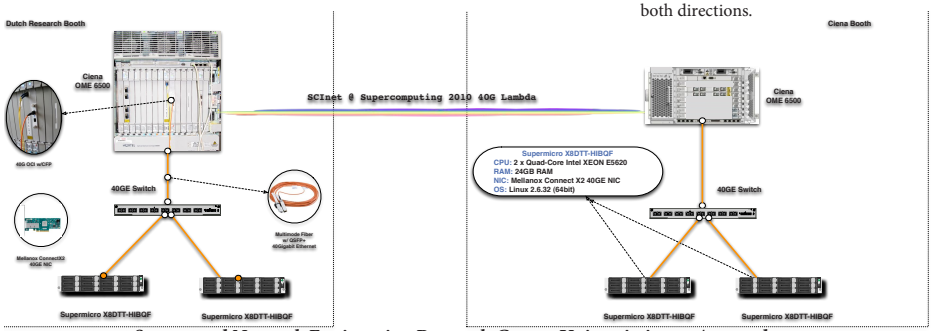
High Performance Node

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 measurments showed constant throughtput between the two remote ends. Using two servers over 70Gbps of aggregated traffic was exchanged in both directions.



System and Network Engineering Research Group, Universiteit van Amsterdam
http://science.uva.nl/research/sne

Universiteit van Amsterdam · SURFnet · ciena · Mellanox Technologies · vrije Universiteit amsterdam

ClearStream

End-to-End Ultra Fast Transmission Over a Wide Area 40 Gbit/s Lambda

University of Amsterdam
Cosmin Dumitru
Cees de Laat
Ralph Koning
SURFnet
Erik-Jan Bos
Gerben van Malenstein
Ciena
David Yeung
Jan-Willem Elion
Harry Peng
Kevin McKernan
Martin Bluethner
VU University Amsterdam
Kees Verstoep
Henri Bal
Mellanox
Erez Cohen
Bill Lee

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 throughtput 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

The DiVinE application is MPI based and in this setup uses TCP/IP as its network backend. DiVinE's runtime system is optimized to achieve good performance despite the very intensive traffic rate and high WAN latency over long distance.

We also use a server with basic UDP and TCP test tools to tune and measure capacities. 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.

DiVinE

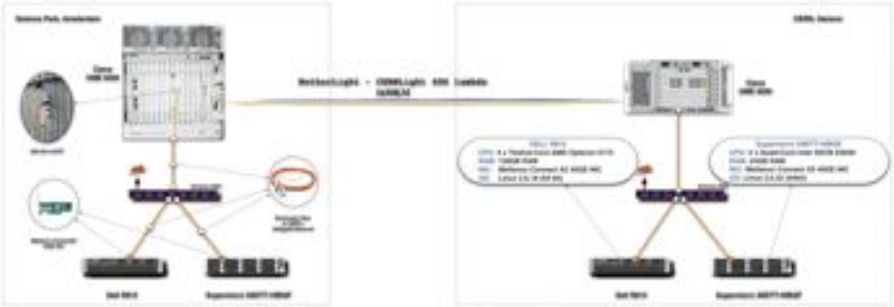
DiVinE is a tool for LTL model checking and reachability analysis of discrete distributed systems. The tool is able to efficiently exploit the aggregate computing power of multiple network-interconnected multi-cored workstations in order to deal with extremely large verification tasks.

Cluster-in-a-box

The Dell R815 is a 2U server powered by 48 AMD Opteron 6100 cores which make it as one of the densest x64 servers available on the market and is used to run the DiVinE application.

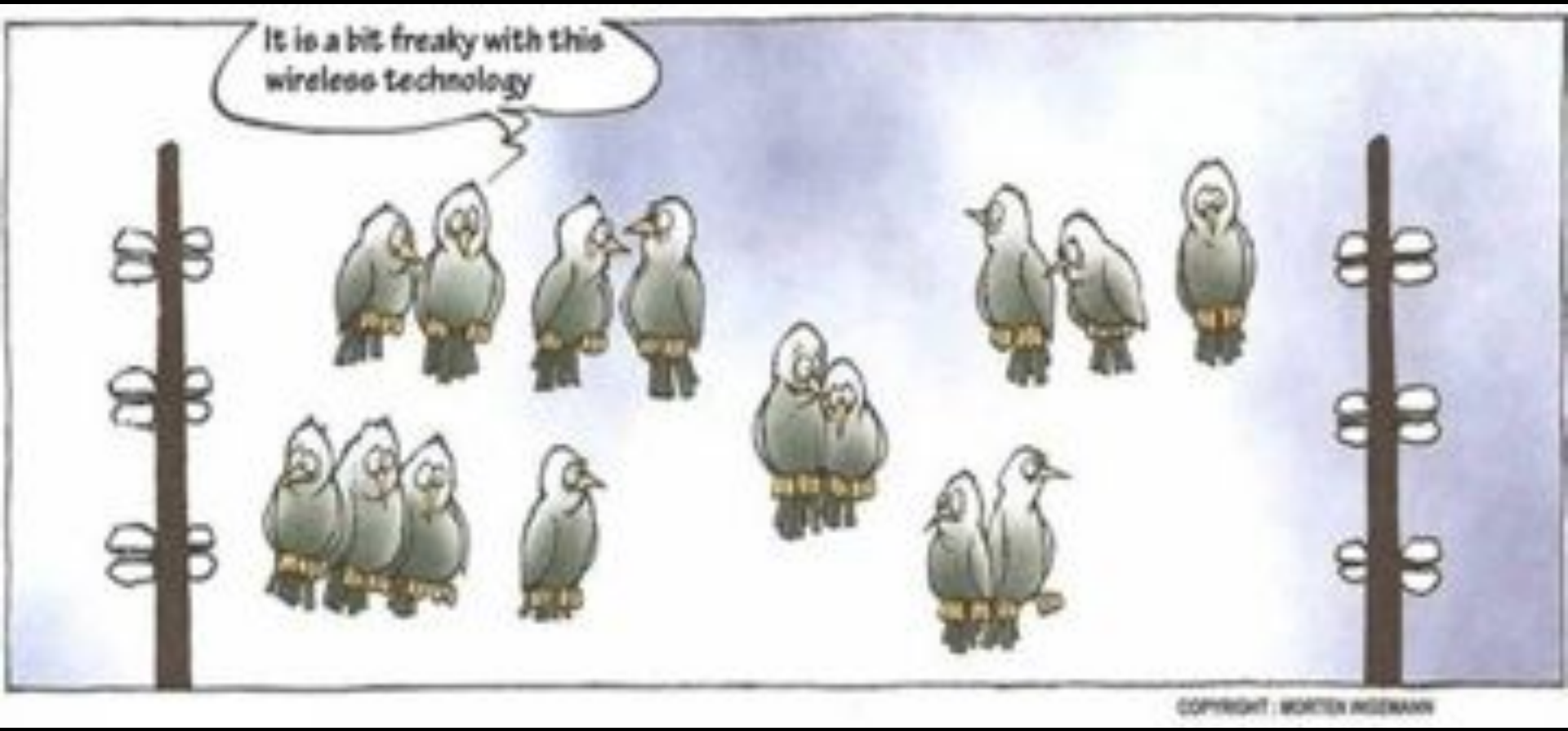
High Performance Node

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.



System and Network Engineering Research Group, Universiteit van Amsterdam
http://science.uva.nl/research/sne

Questions ?



UvA	SURFnet	Ciena	VU	Mellanox
Cees de Laat	Erik-Jan Bos	David Yeung	Kees Verstoep	Bill Lee
Ralph Koning	Gerben van Malenstein	Harry Peng	Henri Bal	Erez Cohen
Cosmin Dumitru	Roeland Nuijts	Martin Bluethner		