The Irish National e-Infrastructure

OpsCentre 10Ge Optical Networking Dec-2010

Date	**2010-12-23 22:40:56**
Task	**Tasks 3.2.3.2**
Document number	**OpsCentre/OpsCentre10GeOpticalNetworkingDec10— 27— 2010-12-23 22:40:56**
Document status	**FINAL**
Author(s)	Brian Coghlan (coghlan@cs.tcd.ie))
File	**https://grid.ie/wiki/OpsCentre/OpsCentre10GeOpticalNetworkingDec10**

This document describes the design, installation, testing and performance of the OpsCentre's 10Ge optical networking

National Development Plan 2007 - 2013

EUROPEAN UNION STRUCTURAL FUNDS

Higher Education Authority
An tÚdarás um Ard-Oideachas

AN ROINN OIDEACHAIS AGUS EOLAÍOCHTA | DEPARTMENT OF EDUCATION AND SCIENCE

CONTENTS

.	Name	Partner	Date	Signature
From	Brian Coghlan	TCD	23-DEC-2010	Brian Coghlan
Verified by	Brian Coghlan	TCD	23-DEC-2010	Brian Coghlan
Approved by	Brian Coghlan	TCD	23-DEC-2010	Brian Coghlan

Table 1: Delivery Slip

Version	Date	Summary of changes	Author(s)
0-0-DRAFT-A	19-DEC-2010	Draft version A	Brian Coghlan
1.0	23-DEC-2010	Final version 1.0	Brian Coghlan

Table 2: Document Log

Preamble File	Date	Version	Author(s)
OpsCentre/OpsCentre10GeOpticalNetworkingDec10/Preamble	19-DEC-2010	1.0	Brian Coghlan

Table 3: Preamble Log

1 EXECUTIVE SUMMARY

This document describes the design, installation, testing and performance of the OpsCentre's 10Ge optical networking.

2 INTRODUCTION

2.1 PURPOSE

This document aims at describing the design, installation, testing and performance of the OpsCentre's 10Ge optical networking. The intended audience for this documents are the Grid-Ireland OpsCentre staff and others that need to know these details.

2.2 SCOPE

The scope is restricted to the design, installation, testing and performance of the OpsCentre's 10Ge optical networking.

2.3 OVERVIEW

The document is organized as follows: first an overview and then a more detailed view.

2.4 DEFINITIONS, ACRONYMS AND ABBREVIATIONS

TCD Trinity College Dublin

OpsCentre Grid-Ireland OpsCentre

2.5 REFERENCES

1 Trinity College Dublin - Information Systems Security Policy: http://www.tcd.ie/ITSecurity/policies/infosec.php

3 BACKGROUND

The following background information is a modified extract from a document prepared by the Information Systems Services (ISS) in Trinity College Dublin (TCD), as it most clearly presents the TCD point of view.

The Grid-Ireland OpsCentre (http://grid.ie/opscentre.html) is based in the Departments of Computer Science and Statistics (SCSS) in TCD. SCSS are an Autonomously Managed Network (AMN) from an the point of view of ISS. The OpsCentre is the highest bandwidth user group in College due to the nature of its research, for example, involvement in data storage and processing for the ATLAS and LHCb experiments, and the OpsCentre is a partner in the e-INIS project to develop a national e-Infrastructure for Ireland.

Grid-Ireland has historically evolved within SCSS. Grid-Ireland OpsCentre Internet bound traffic, which comprises over 50% of all TCD Internet bound traffic, therefore traversed both SCSS ANM and ISS networks prior to arrival at the TCD network boundary ingress/egress point at TCD network firewalls. This configuration was highly inefficient and under optimised. In August 2009 ISS provided SCSS with a second link to specifically support the OpsCentre, which allocated 50% or 500Mbps of available bandwidth to SCSS/OpsCentre. However this configuration remained unfit for purpose as the OpsCentre required multi-gigabit bandwidth to/from the Internet. Internally the OpsCentre had already operated a 10Ge network for the previous 18 months. On 27th May 2010 the TCD ingress/egress point was upgraded from 1Gbps to 10Gbps (10Ge). The TCD network configuration was then as shown in Figure 1.

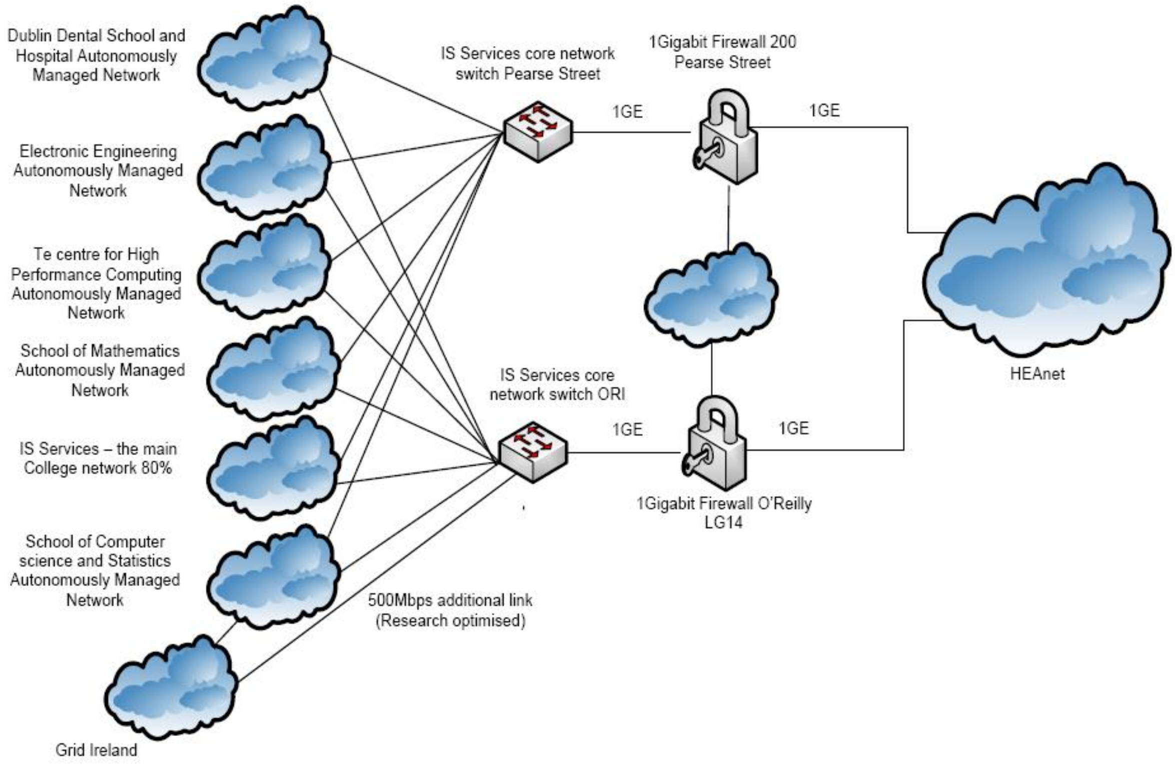

Dublin Dental School and Hospital Autonomously Managed Network
Electronic Engineering Autonomously Managed Network
Te centre for High Performance Computing Autonomously Managed Network
School of Mathematics Autonomously Managed Network
IS Services – the main College network 80%
School of Computer science and Statistics Autonomously Managed Network
Grid Ireland
500Mbps additional link (Research optimised)
IS Services core network switch Pearse Street
1GE
1Gigabit Firewall 200 Pearse Street
1GE
IS Services core network switch ORI
1GE
1GE
1Gigabit Firewall O'Reilly LG14
HEAnet

Figure 1: TCD networking circa June 2010

On 10th June 2010 ISS submitted to the TCD Information Policy Committee (IPC), on behalf of the OpsCentre and SCSS, a request for reconfiguration of the OpsCentre connectivity forward of TCD firewalls to new 10Ge routers, thereby providing optimised 10Ge connectivity between the OpsCentre and the Internet. This required IPC approval as section 2.8 of *Trinity College Dublin - Information Systems Security Policy* [1] assumes all users are located behind ISS firewalls. However the TCD IT Governance model, roles and responsibilities, IP address maps, work flow and incident response processes and procedures remain unchanged as the OpsCentre remains accountable to the SCSS AMN. ISS will maintain responsibility for the network perimeter and can enforce access control at the new 10Ge routers for Grid-Ireland OpsCentre research traffic.

This request was approved, and a period of reconfiguration and testing then took place, with substantial support from ISS, leading eventually to switchover of production grid traffic to the new 10Gbps path on 27th October 2010. The new TCD network configuration is shown in Figure 2. The connection is now via a redundant pair of failover

10Gbps optical links, one to each of the redundant pair of TCD 10Gbps front-end routers to HEAnet's Dublin ROADM ring.

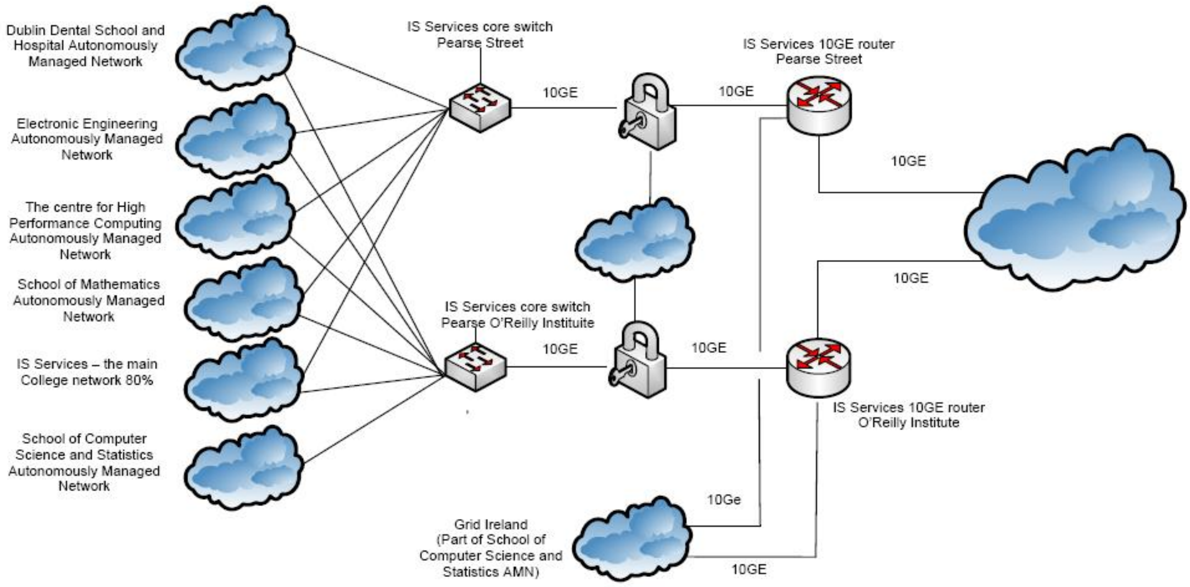

Dublin Dental School and Hospital Autonomously Managed Network
Electronic Engineering Autonomously Managed Network
The centre for High Performance Computing Autonomously Managed Network
School of Mathematics Autonomously Managed Network
IS Services – the main College network 80%
School of Computer Science and Statistics Autonomously Managed Network
IS Services core switch Pearse Street
IS Services core switch Pearse O'Reilly Instituite
IS Services 10GE router Pearse Street
IS Services 10GE router O'Reilly Institute
Grid Ireland (Part of School of Computer Science and Statistics AMN)
10GE
10GE
10GE
10GE
10GE
10GE
10Ge
10GE

Figure 2: TCD networking circa November 2010

4 GRID-IRELAND OPSCENTRE 10GE OPTICAL NETWORKING

Logically the new OpsCentre 10Gbps optical networking is a straightforward Ethernet link to HEAnet via SCSS and ISS as beforehand. Figure 3 illustrates this. The upstream link to/from HEAnet terminates at the OpsCentre firewall. The firewall has a 10Gbps downstream link to a Dell 8024 10Gbps level 3 switch, which itself has multiple 10Gbps downstream links, including channel-bonded links to a Force10 10Gbps level 2 switch that is the effective backbone of the OpsCentre's 10Ge network. This switch connects at 10Gbps to the main compute cluster and to the OpsCentre's portion of the e-INIS federated datastore.

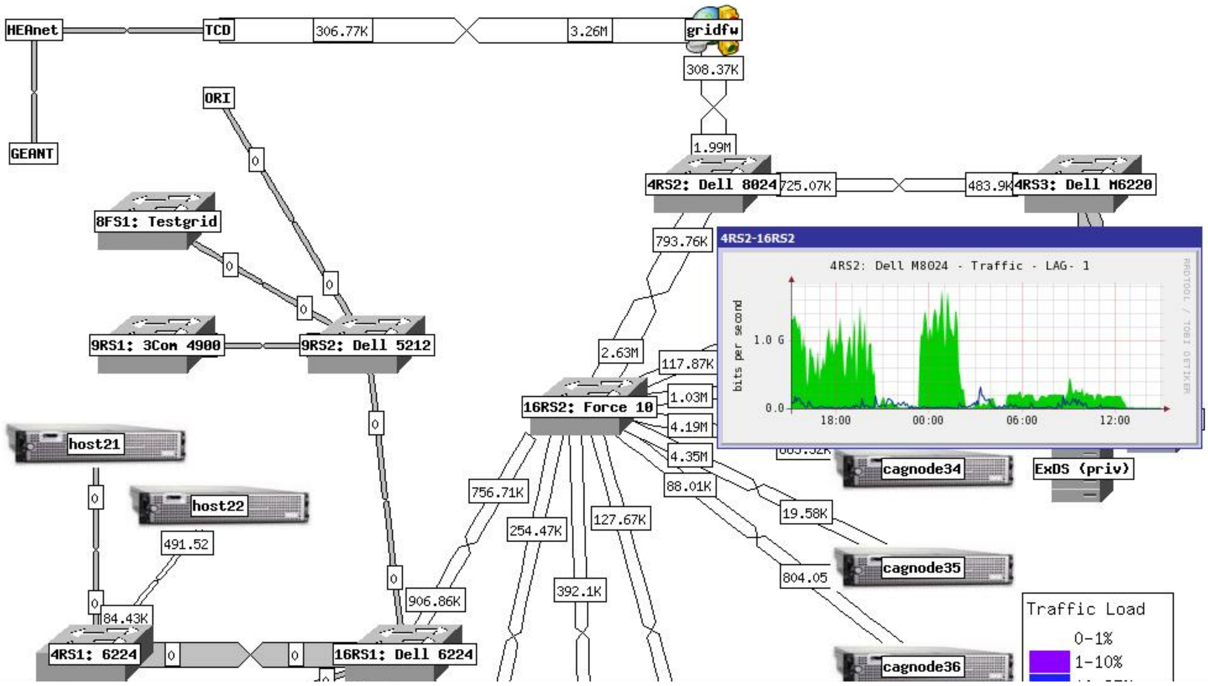


Figure 3: OpsCentre 10Ge networking

While the OpsCentre internal networking remained as was, a new 10Gbps OpsCentre firewall and two new redundant upstream 10Gbps links superceded the previous 1Gbps firewall and its single 1Gbps upstream link. Examination of available commercial 10Gbps firewalls found that the flexibility desired was only available at very great expense, and so a very fast but otherwise replicate of the previous Linux firewall was constructed from a Dell R510 dual-socket 6-core server with a solid-state disk (SSD) for the operating system plus a mirrored 2TB disk for logs and other security data, plus two Intel dual-port 10Gbps PCI-e interfaces.

5 TESTING OF OPSCENTRE 10GE OPTICAL NETWORKING

5.1 GOALS

Following initial tests the main goal was to break down more of the chain and test parts of the network path to resolve where slowness was being introduced. The next steps then were:

- Test with ACLs applied on 7606 TCD front-end router
- Test redundant link and then test failover
- Test with some traffic limiter in place

5.2 FIRST TESTS

- Test storage element (SE) *g10g-se* exchanged for another machine
  - New machine behaves in the same way, not specific to one machine.
- SFP+ module in new firewall *g10g-fw* exchanged
  - Both modules give the same performance

ISS also ordered a new module for their 7606 router and suggested that OpsCentre staff take a machine to the ISS lab to test a direct connection to the router with multimode fibre.

5.3 MULTIMODE TESTS

A Dell PE 1950 server (the server used for tests before the R510 arrived) was equipped with one single-port and one dual-port 10Ge card and configured as *g10g-fw*. This was taken to the ISS lab and connected to the 7606 router. The only problem seen was that the server assigned eth2 to the single-port card rather than the first port on the dual-port card, but this was acceptable for testing.

For basic iperf TCP tests the remote node at DIAS was unfortunately very heavily loaded and this impacted results. For 15 minutes per run the results were:

- Single stream DIAS-TCD = 209 Mbits/sec
- 20 streams DIAS-TCD = 4.80 Gbits/sec
- Single stream TCD-DIAS = 2.77 Gbits/sec
- 20 streams TCD-DIAS = 4.46 Gbits/sec
- 10 stream bi-directional = In 834 Mbits/sec, Out 5.71 Gbits/sec

For single stream UDP tests the results were:

- DIAS-TCD 3Gbps cap = 1.10 Gbits/sec
- TCD-DIAS 3Gbps cap = 1.80 Gbits/sec

5.4 SINGLEMODE CHECK

The new gbic module was installed by ISS into their 7606 router to test original hardware connecting through the new module. This was just a fairly cursory check to make sure behaviour was as before. The test machine in DIAS seemed not to be as heavily loaded any more so the results were better:

- Single stream g10g-fw-DIAS = 9.43 Gbits/sec
- Single stream DIAS-g10g-fw = 572 Mbits/sec
- 20 streams DIAS-g10g-fw = 6.78 Gbits/sec

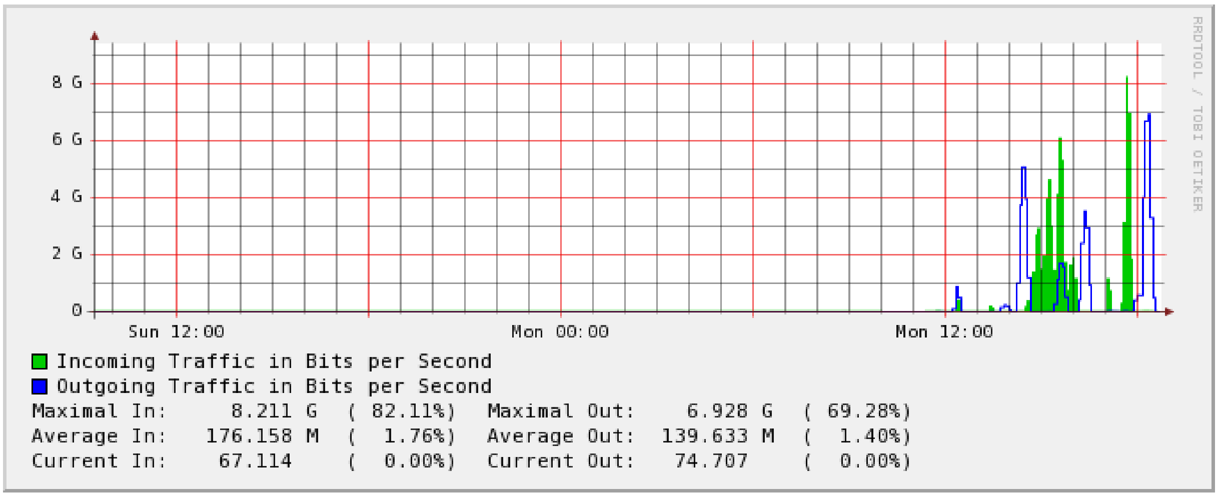


Figure 4: 10Ge traffic through single and multimode tests

5.5 GRAPH FOR 13TH SEPTEMBER

Figure 4 shows the network activity through the multimode tests and singlemode check.

5.6 SOLUTION OF MAIN PROBLEM

When routing, inbound traffic speeds were slowing to of the order of hundreds of kilobits per second (3 orders of magnitude less than outbound streams). Various combinations of connections were tried and the slowdown to Kbps was eventually isolated to something in the g10g-fw firewall:

- A duplicate g10g-fw was created and installed in the ISS lab with the difference being a multimode/SR card for eth2 (previously documented).
- Also a duplicate g10g-se was added, yielding expected behaviour, with slowdown to multi-Mbps inbound not Kbps.
- Then multimode/SR card was swapped for a single-mode/LR card. The Kbps drop-off came back. All other settings remained the same.

This was actually fixed in a new version of the driver that was released early in September 2010. New drivers were built and installed into the kernel, restoring performance to Mbps with Gbps possible using multiple streams. The remaining issue was that a per-connection limit of 100Mbps-1Gbps was observed that affected inbound traffic more than outbound.

5.7 SINGLE STREAM CHARACTERISATION 27-09-2010

The following test servers were employed:

- g10g-fw, mostly on 134.226.254.106, connected to the TCD 7606 router for these tests (also 134.226.42.1 to connect with TCD local, 134.226.42.24 to connect to g10g-se)
- g10g-se 134.226.42.4, client machine behind new firewall router
- DIAS 193.1.48.2 - test server in DIAS e-INIS site, 10GbE connection
- UCC 193.1.48.64 - test server in DIAS e-INIS site, 10GbE connection throttled to 2Gb/s
- HEAnet 193.1.228.67 - phoebus.heanet.ie, 1GbE test server
- NL 194.171.100.2 - iperf server in the Netherlands

Tests used iperf with the -r option so test was run in one direction and then the other from a single host. In general this worked but for some reason not for UCC connecting with the Netherlands - the test machine could connect out to the Netherlands but the connection back failed, possibly through firewalling. However the -d option did work in this case, causing the two connections to happen simultaneously. The extra figure gathered this way is bracketed to indicate that it was not generated in the same way as the other figures.

From - To	**FW**	**SE**	**DIAS**	**UCC**	**HEAnet**	**NL**
FW		5.00G	846M	521M	108M	945M
SE	3.63G		2.15G	1.3G	881M	1.08G
DIAS	856M	421M		1.54G	798M	885M
UCC	506M	287M	614M		895M	1.16M
HEAnet	529M	560M	916M	921M		372M
NL	289M	130M	258M	(261M)	153M	

Table 4: Single stream characterisation 27-09-2010

5.8 TESTS BEFORE AND AFTER ACL APPLIED ON 7606

A series of single stream netperf and iperf tests were run before and after the router ACLs were applied. Pings were also performed before and after - these varied around 2ms before and after.

	2pm test		**3:30pm test**	
Path	**Out**	**Back**	**Out**	**Back**
SE <–> **DIAS**	2.00Gb/s	435Mb/s	1.84Gb/s	422Mb/s
SE <–> **NL**	3.15Gb/s	29.7Mb/s	941Mb/s	25Mb/s
SE <–> **UCC**	1.04Gb/s	145Mb/s	1.25Gb/s	111Mb/s
SE <–> **HEAnet**	875Mb/s	578Mb/s	882Mb/s	588Mb/s
FW <–> **DIAS**	1.78Gb/s	815Mb/s	1.57Gb/s	667Mb/s
FW <–> **NL**	908Mb/s	158Mb/s	919Mb/s	110Mb/s
FW <–> **UCC**	1.18Gb/s	392Mb/s	1.52Gb/s	571Mb/s
FW <–> **HEAnet**	895Mb/s	903Mb/s	913Mb/s	899Mb/s

Table 5: iperf tests, both directions sequentially (-r), 120seconds per test

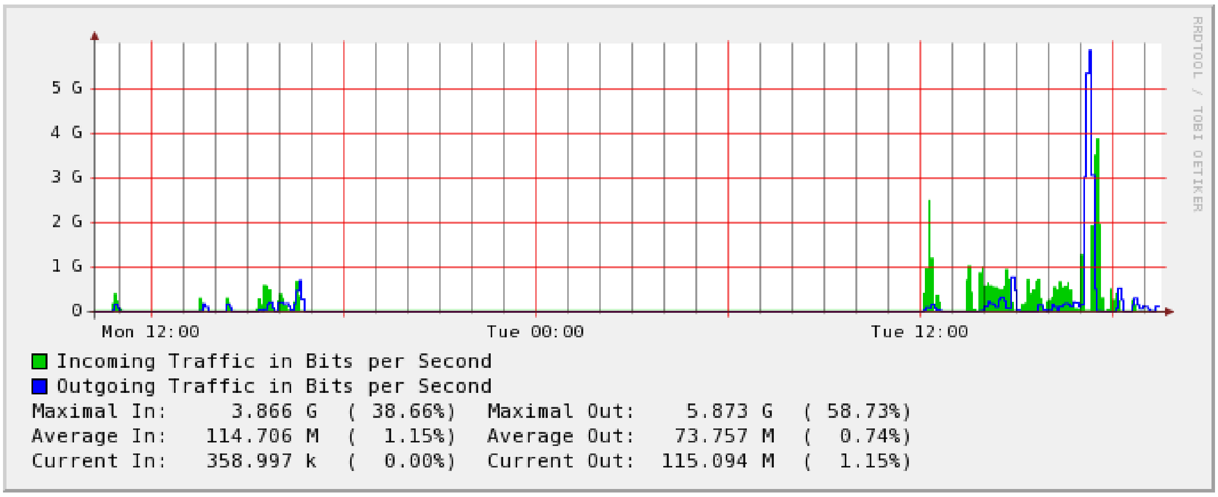
	2pm test		**3:30pm test**	
Host	**Out**	**Back**	**Out**	**Back**
SE	2336.81Mb/s	258.41Mb/s	1848.09Mb/s	289.44Mb/s
FW	1331.65Mb/s	2723.92Mb/s	1466.0Mb/s	1772.75Mb/s

Table 6: netperf TCP STREAM tests to/from DIAS

	2pm test		3:30pm test	
Host	**Out**	**Back**	**Out**	**Back**
SE	483.71/s	482.07/s	490.24/s	493.78/s
FW	495.97/s	499.27/s	505.346/s	506.07/s

Table 7: netperf TCP RR tests to/from DIAS (TCP request-response test should indicate latency)

After the ACLs were applied the link was also run fairly hard doing iperf in both directions sequentially (-r option) between g10g-se and DIAS. Figure 5 shows the traffic through the single mode link through the whole series of tests. Clearly, multi-gigabit transfers are still possible using multiple streams after the ACLs are applied.



```
🟩 Incoming Traffic in Bits per Second
🟦 Outgoing Traffic in Bits per Second
Maximal In:     3.866 G  ( 38.66%)  Maximal Out:    5.873 G  ( 58.73%)
Average In:   114.706 M  (  1.15%)  Average Out:   73.757 M  (  0.74%)
Current In:   358.997 k  (  0.00%)  Current Out:  115.094 M  (  1.15%)
```

Figure 5: 10Ge traffic through single mode link

5.9 Congestion control tests 30-09-2010

There's a suggestion that congestion control could be the cause of the slowdown in the router - one link is faster than the other and that might be expected to cause an asymmetry. The conclusion is that the congestion control algorithm makes less of a difference than the random variations in the network.

Algorithm	**Run1**	**Run2**	**Run3**	**Run4**	**Run5**	**Mean**	**Stdev**
reno	281	256	388	367	319	322.2	55.74
cubic	271	283	279	334	364	306.2	40.71
highspeed	270	301	389	310	436	341.2	68.79
htcp	303	282	322	435	360	340.4	60.17
hybla	262	248	303	417	308	307.6	66.37
lp	288	259	274	366	359	309.2	49.79
scalable	288	239	342	362	398	325.8	62.75
vegas	253	340	324	300	398	323	53.25
veno	315	307	251	360	315	309.6	38.86
westwood	301	350	332	339	303	325	21.97

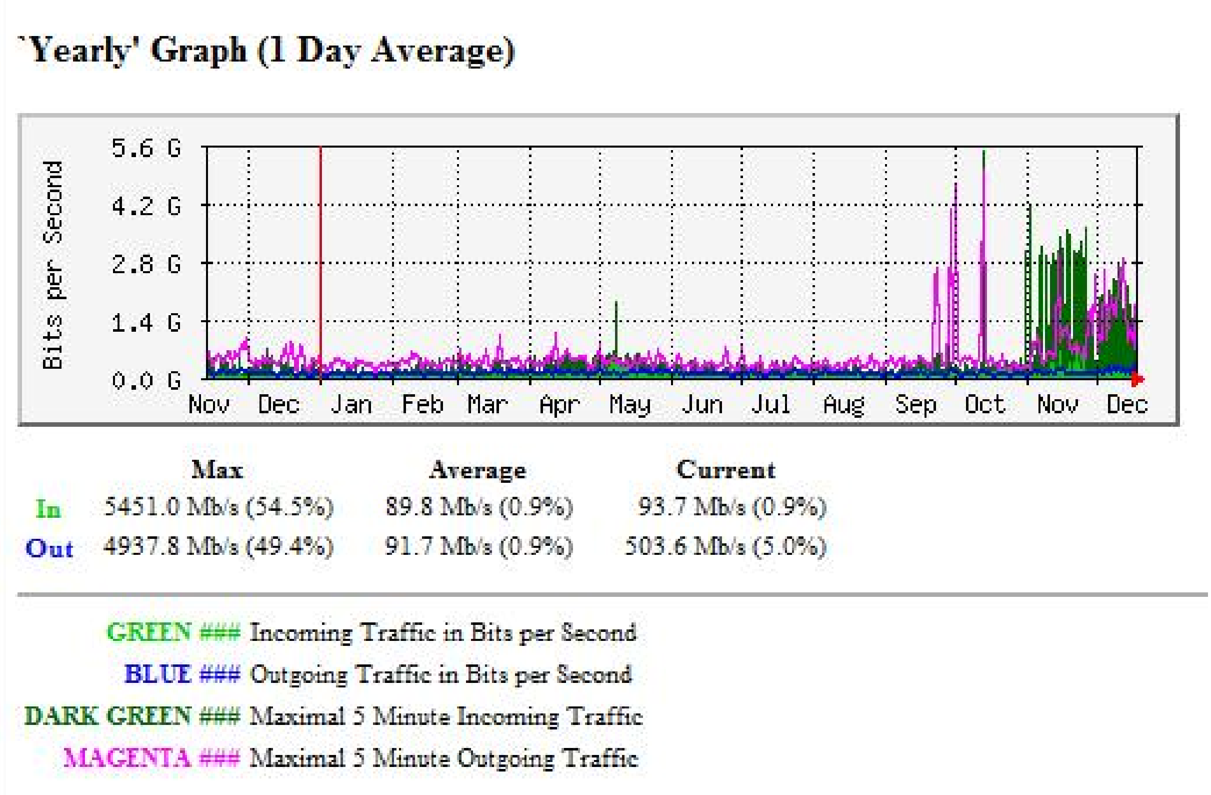
Table 8: Runs using netperf connecting from DIAS to g10g-se, each run is 30 seconds

5.10 Redundant link

The redundant link is physically connected and configured with the necessary addresses. The final step was to investigate routing priorities in Linux and to use that to achieve hot-failover.

5.11 OPSCENTRE 10GE OPTICAL NETWORKING PERFORMANCE

The introduction of 10Gbps networking has been very successful. The OpsCentre grid traffic already represented more than 50% of TCD network traffic, but there has been an order of magnitude increase since 10Gbps was enabled for production grid jobs in late Oct-2010, as shown in Figure 6.



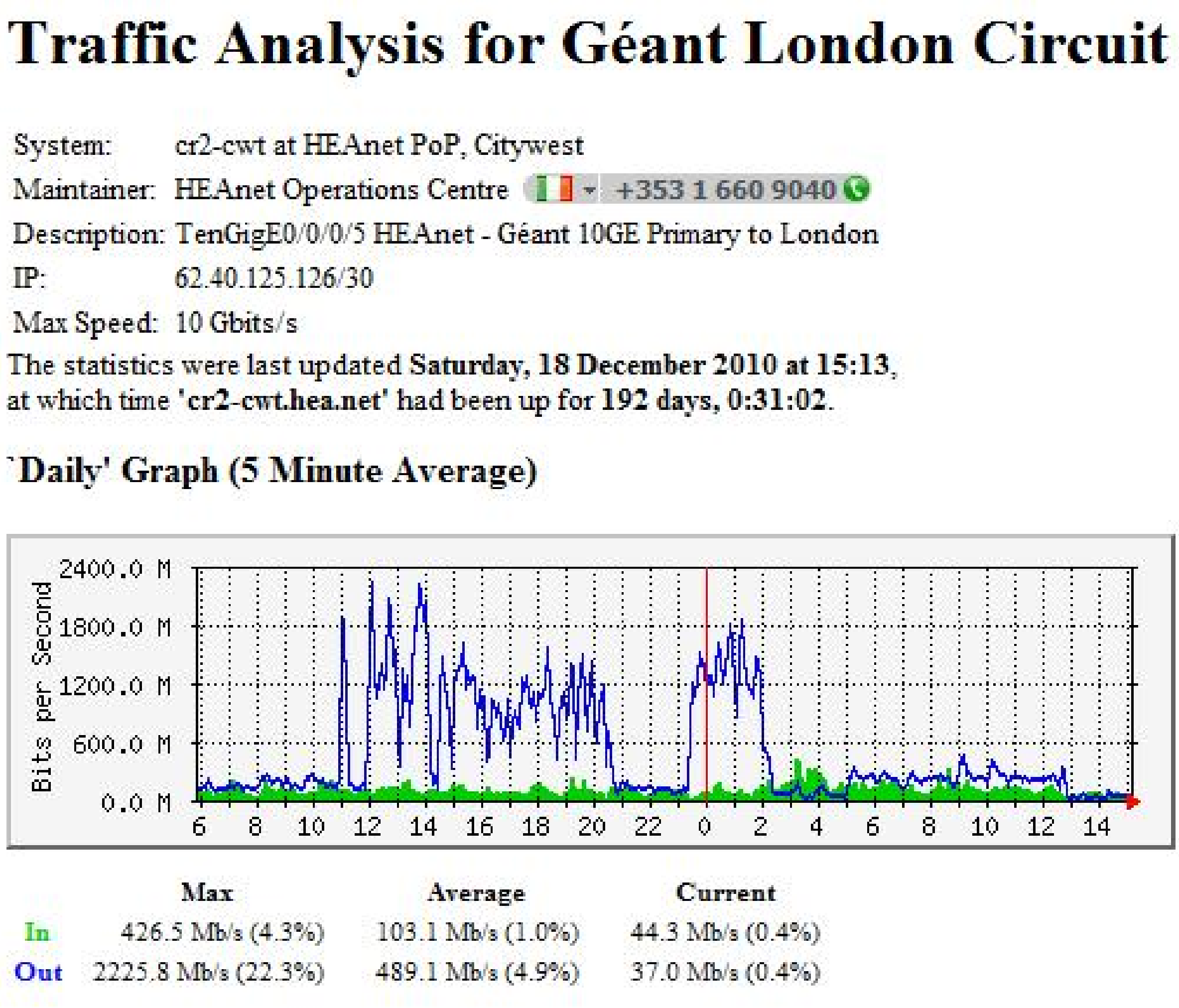
`Yearly' Graph (1 Day Average)

	Max	Average	Current
In	5451.0 Mb/s (54.5%)	89.8 Mb/s (0.9%)	93.7 Mb/s (0.9%)
Out	4937.8 Mb/s (49.4%)	91.7 Mb/s (0.9%)	503.6 Mb/s (5.0%)

GREEN ### Incoming Traffic in Bits per Second
BLUE ### Outgoing Traffic in Bits per Second
DARK GREEN ### Maximal 5 Minute Incoming Traffic
MAGENTA ### Maximal 5 Minute Outgoing Traffic

Figure 6: 10Ge external networking Yearly average 18-Dec-2010 15:14

This grid traffic also greatly improves the utilisation of HEAnet's Geant international network connections, as shown in Figure 7.



Traffic Analysis for Géant London Circuit

System:	cr2-cwt at HEAnet PoP, Citywest
Maintainer:	HEAnet Operations Centre +353 1 660 9040
Description:	TenGigE0/0/0/5 HEAnet - Géant 10GE Primary to London
IP:	62.40.125.126/30
Max Speed:	10 Gbits/s

The statistics were last updated **Saturday, 18 December 2010 at 15:13**, at which time **'cr2-cwt.hea.net'** had been up for **192 days, 0:31:02**.

`Daily' Graph (5 Minute Average)

	Max	Average	Current
In	426.5 Mb/s (4.3%)	103.1 Mb/s (1.0%)	44.3 Mb/s (0.4%)
Out	2225.8 Mb/s (22.3%)	489.1 Mb/s (4.9%)	37.0 Mb/s (0.4%)

Figure 7: 10Ge external networking 5min average 18-Dec-2010 15:14

The increased usage of the network connections, both for external connections to/from HEAnet and also internally within the OpsCentre, can be seen from Figure 8 and Figure 9. And finally, Figure 10 shows a more detailed set of traffic analysis graphs of the 10Gbps TCD connection to HEAnet.

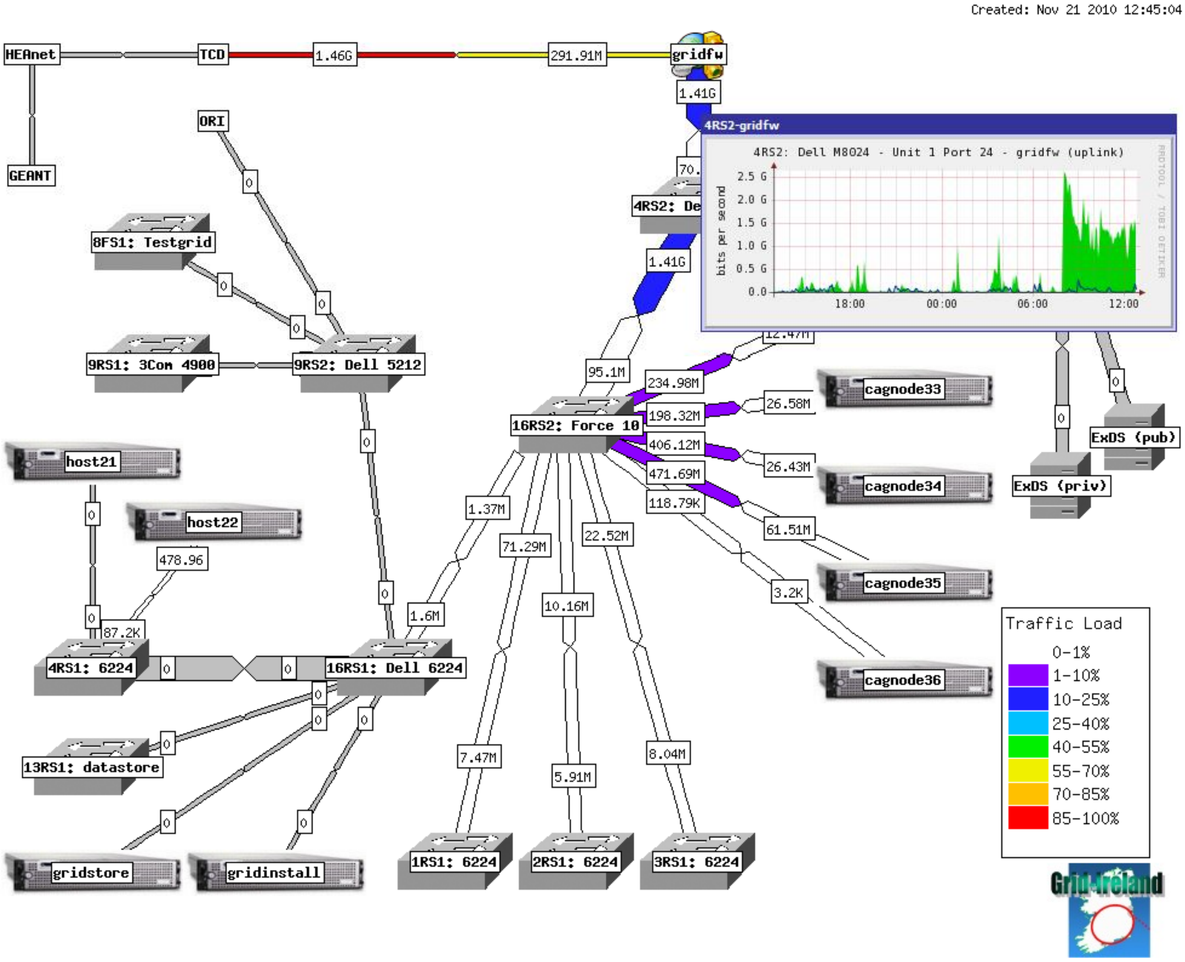


Figure 8: tcd-grid-20101121-1250-v2

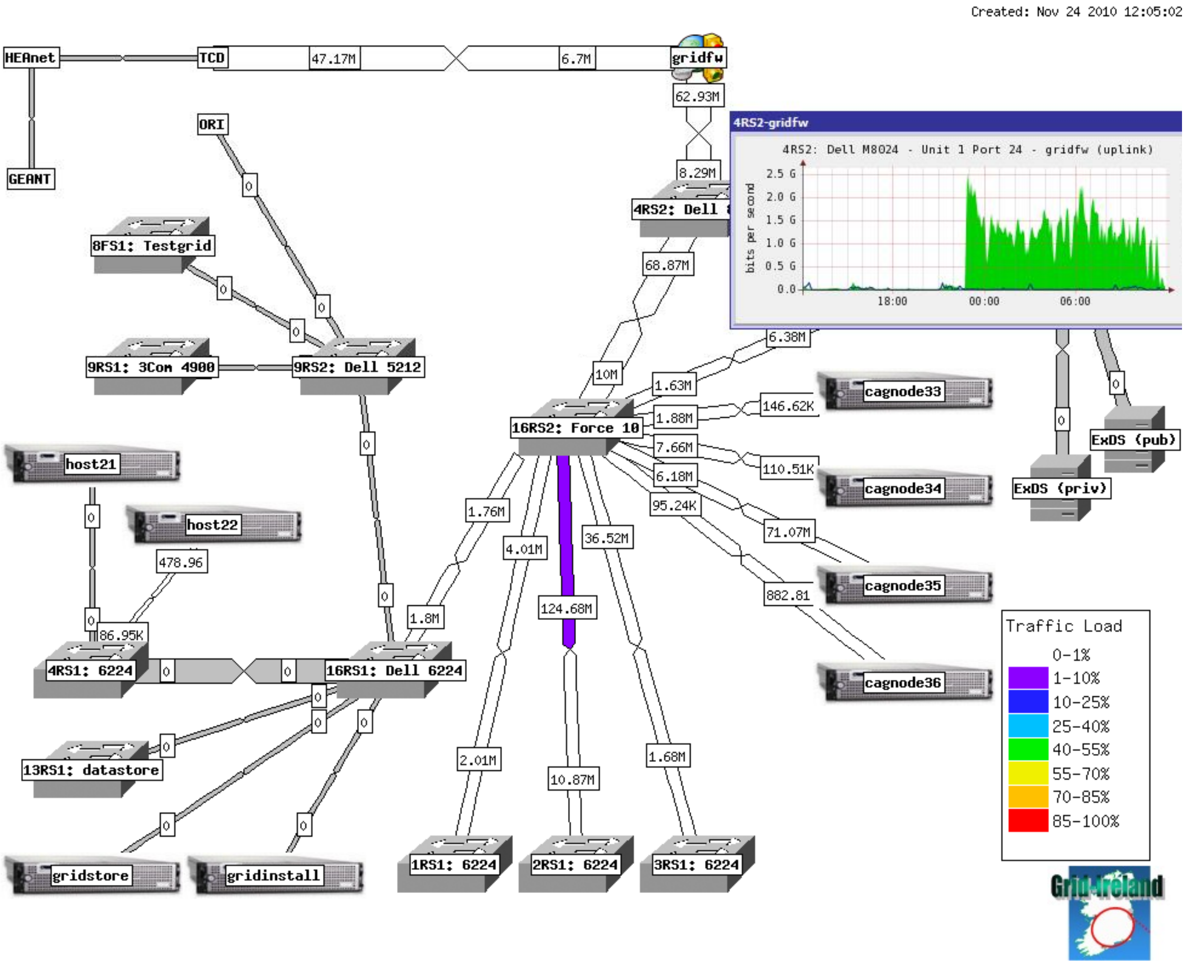
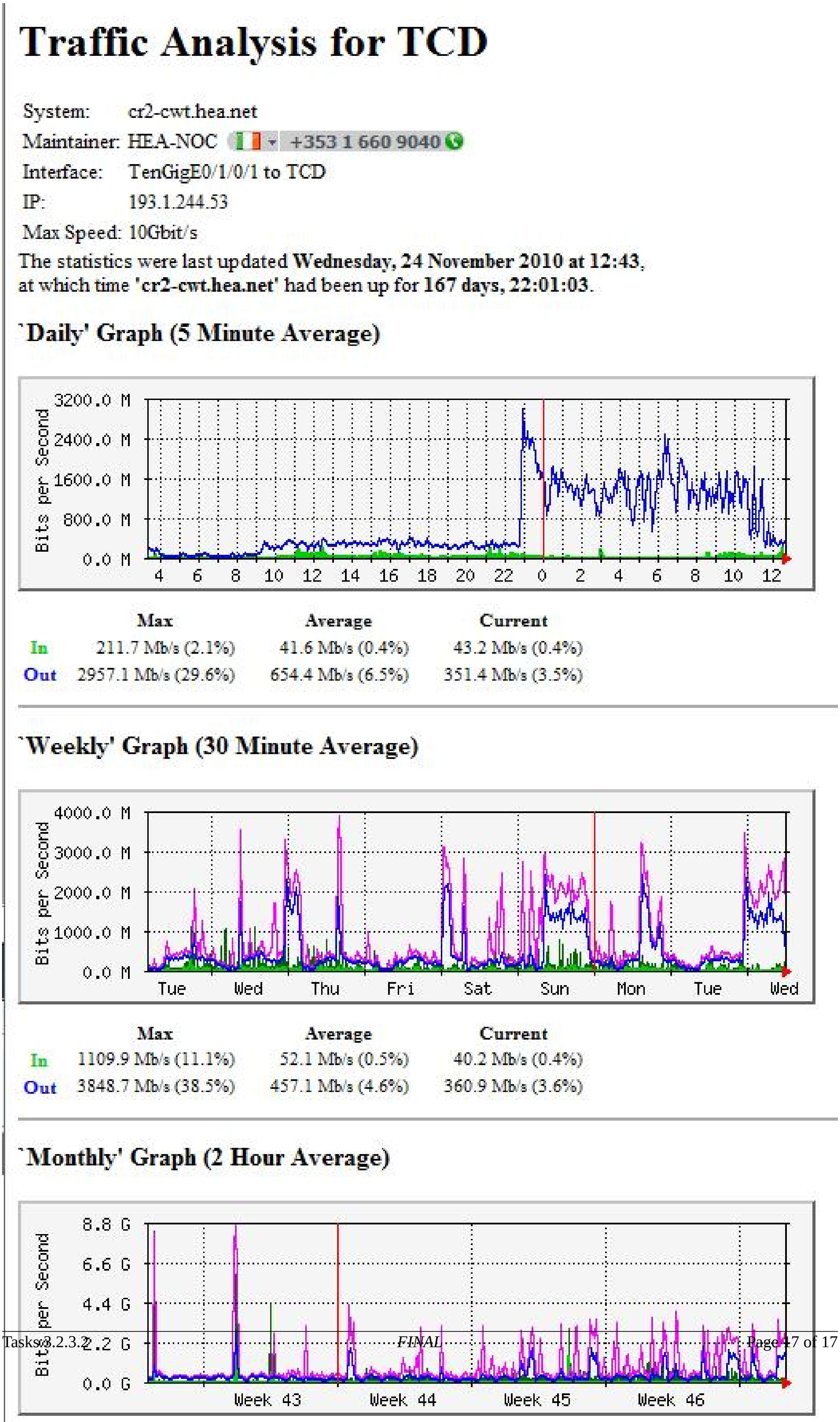


Figure 9: tcd-grid-20101124-1206-v2



Traffic Analysis for TCD

System:	cr2-cwt.hea.net
Maintainer:	HEA-NOC +353 1 660 9040
Interface:	TenGigE0/1/0/1 to TCD
IP:	193.1.244.53
Max Speed:	10Gbit/s

The statistics were last updated **Wednesday, 24 November 2010 at 12:43**,
at which time '**cr2-cwt.hea.net**' had been up for **167 days, 22:01:03**.

`Daily' Graph (5 Minute Average)

	Max	Average	Current
In	211.7 Mb/s (2.1%)	41.6 Mb/s (0.4%)	43.2 Mb/s (0.4%)
Out	2957.1 Mb/s (29.6%)	654.4 Mb/s (6.5%)	351.4 Mb/s (3.5%)

`Weekly' Graph (30 Minute Average)

	Max	Average	Current
In	1109.9 Mb/s (11.1%)	52.1 Mb/s (0.5%)	40.2 Mb/s (0.4%)
Out	3848.7 Mb/s (38.5%)	457.1 Mb/s (4.6%)	360.9 Mb/s (3.6%)

`Monthly' Graph (2 Hour Average)