AccessionIndex: TCD-SCSS-T.20121208.098
Accession Date: 8-Dec-2012
Accession By: Dr.Brian Coghlan
Object name: csTCDie Grid Site Beowulf Clusters and Datastore
Vintage: c.2009
Synopsis: Complex of clusters & storage (1500 cores/600 TB) using 1Gbps Ethernet interconnect and 10Gbps backbone, participant in DataGrid, EGEE, EGI, and CERN LHC computing. From 2013 repurposed as SCSS Cloud.

Description:
From 2009 onwards the complex of clusters and storage that comprised the csTCDie grid site were extensively upgraded, eventually approximating 1500 compute cores and 600 TB of storage, plus the Grid-Ireland site gateway, redundant OpsCentre servers and national servers.

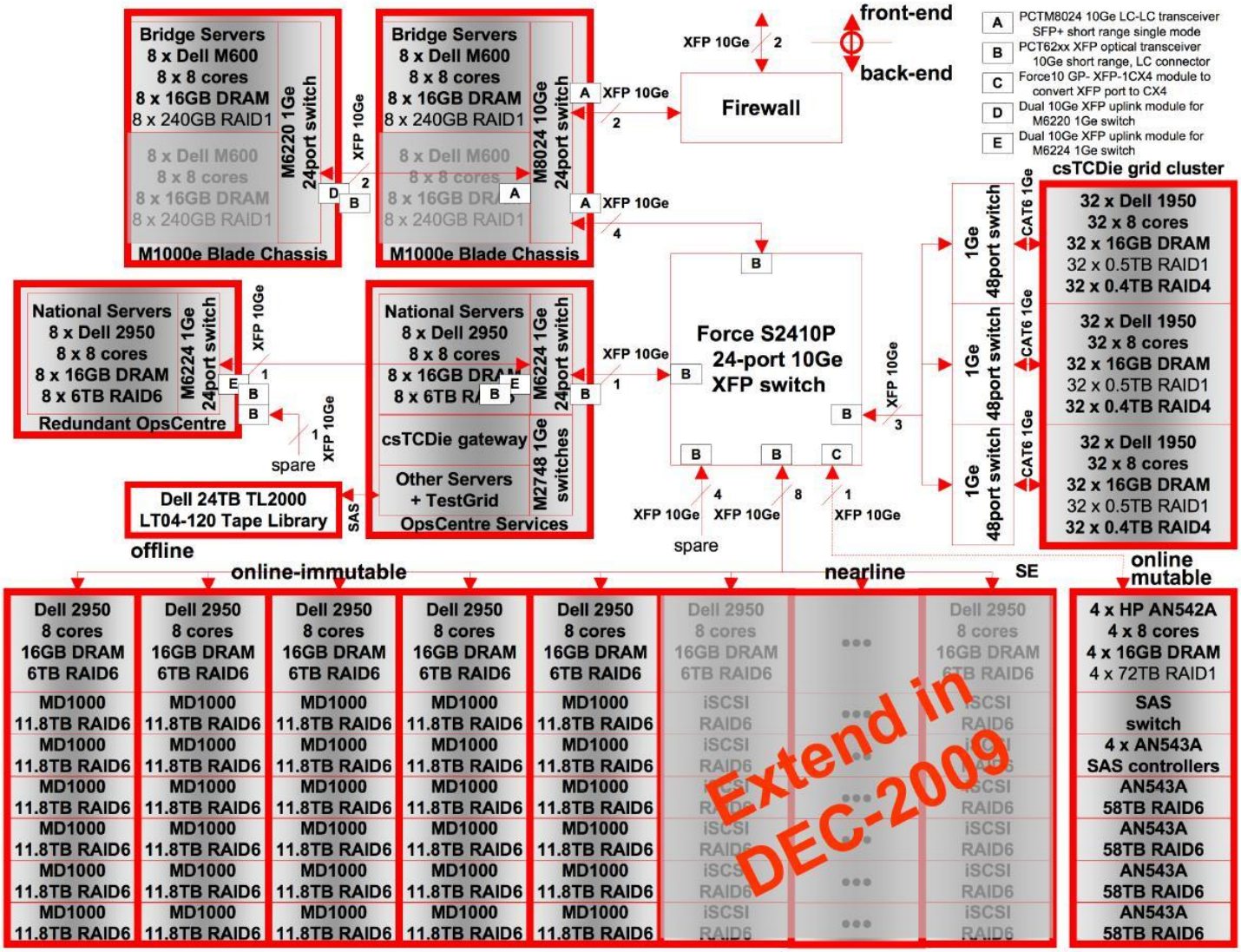

Figure 1: csTCDie Grid Site architecture

The bulk of this infrastructure remains in use, and is gradually being repurposed as the SCSS Cloud. At the front-end, dual dedicated 10Gbps links to the 40Gbps HEAnet Dublin ring, terminated on a Linux firewall with dual-socket 6-core CPUs, 48GB

memory, RAID-1 SSDs and 10Gbps adapters. Its two backend 10Gbps links fed a Dell M8024 level-3 10Gbps 24-port router. This mediated between a set of Dell blade servers able to bridge the public and private networks, see further below. Eight servers had direct 10Gbps connections, eight more for less demanding networking had 1Gbps links to a switch with dual 10Gbps paths the router. The router then connected to the private backbone via 4 x 10Gbps links to a Force10 S2410P 10Gbps switch.

The primary csTCDie grid site was based around a standard 96-node Beowulf cluster. Each node had a dual-socket 4-core CPU, 16GB of memory and 2 x 1TB of RAID-1 (mirrored) storage. The cluster interconnect used 1Gbps Ethernet, connected to the 10Gbps backbone via three 48-port 1Gbps switches with 10Gbps uplinks. An out-of-band parallel network served a central KVM keyboard/monitor/mouse.

Figure 2: csTCDie 96-node Beowulf Cluster Racks

Job execution on the primary cluster was scheduled by the site grid gateway PBS server, which was connected to the 10Gbps backbone via one of the site gateway switches. The site gateway switches also served a set of dual-redundant OpsCentre servers that hosted those services that could be successfully configured this way as well as principal national servers. These switches and servers were actually in racks either side of the primary csTCDie cluster racks.

Additional secondary clusters were connected via the site gateway switches. These included a 16-node Sony PS3 Linux cluster, a 22-node cluster of Dell R410 dual-socket 6-core servers with 48MB memory and 2 x 2TB of RAID-1 storage, a 16-node *ELgrid* cluster for adaptive eLearning, a 9-node *VRengine* with 2-d toroidal SCI interconnect for virtual reality, a 16-node GPU cluster with 64 cores and 32 GPUs, and a *TestGrid* with about 40-nodes plus porting targets (see elsewhere in this catalog

for the VRengine, GPU cluster and TestGrid). In the main these resided in a set of racks known as the OpsCentre racks. The PS3 and GPU clusters were together on large custom shelving units.

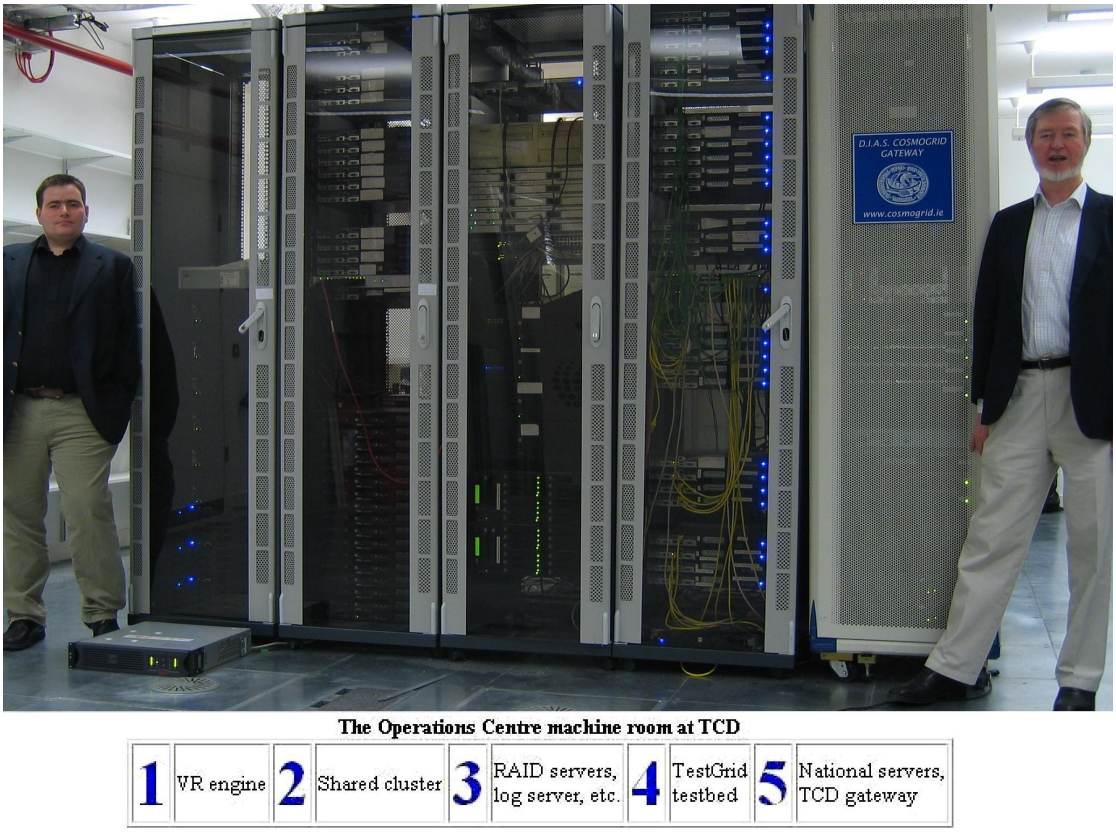

Figure 3: csTCDie OpsCentre Racks
Left: Grid-Ireland Grid Manager John Walsh, right: Dr.Brian Coghlan

The 10Gbps backbone connected to a 2-layer datastore architecture that assumed that research communities were best able to define their metadata and develop front-end interfaces, whilst common back-ends could take the best advantage of economies of scale. A set of front-end bridge servers were provided to minimize restrictions on implementations and security policies, isolating them from back-ends and each other.

The back-ends assumed communities had their own data access patterns that could be described via two properties, *mutability* and *frequency-of-access*, and so different storage technologies were provided for frequently accessed mutable and immutable data, and a third technology for infrequently-accessed data.

Frequently-accessed mutable data was handled by a HP ExDS9100 with a highly available filesystem across 328TB of raw or 232TB of usable storage, see elsewhere in this catalog. Problems with its 10Gbps virtual network switch made it unreliable.

The needs of frequently-accessed immutable data were well defined and had been dealt with for a long time by the grid community. The datastore used multiple Dell MD1000 RAID6 SAS disk arrays with 15 x 1TB disks, with a server per block of six arrays, and five blocks, yielding 354TB. Performance was kept high using large write-through caches at all levels, as the proportion of writes was very low, and no updates ever occurred, only deletes. In 2010 some of the 1TB disks were upgraded to 2TB.

Rarely or infrequently accessed data potentially allows greater storage packaging density and energy efficiency. A 70TB NexSan storage subsystem was added in late 2010 that complied with the MAID (Massive Array of Idle Disks) specification, allowing disks to change state to reflect access frequency. A Dell 24-tape high-speed tape library with an IBM Ultrium drive for 800GB LTO-4 tapes was also provided.

Figure 4: csTCDie Datastore Racks
Far left: ExDS rack, then left-to-right four MD1000 racks

Over the years the csTCDie grid site was funded from a variety of sources, such as the EU FP5, FP6 and FP7 programmes, the HEA PRTLI3 and PRTLI4 programmes, and Science Foundation Ireland, with long-standing support from the SCSS and ISS in Trinity College Dublin. It was a resource for a number of national research projects, including Cosmogrid, WebCom-G and e-INIS, and all the major EU grid projects, including DataGrid, CrossGrid, EGEE, EGEE-2, EGEE-3, Int.EU.Grid, EGI-InSPIRE and eventually the pan-EU *European Grid Infrastructure* (EGI), as well as other EU projects like StratusLab, Manychore, HELIO, SCI-BUS and ERflow.

The csTCDie grid site was also a significant contributor to the CERN LHC computing for the ATLAS and LHCb detectors, greater than for many larger countries. It was a Tier-2 site in the *Dutch ATLAS Cloud*, associated to the Netherlands Tier-1 at SARA. Typically the site received a defined proportion of the datasets, ran analysis jobs on those datasets and ran production jobs. Even when data transfers into the site from SARA were throttled to 500Mbps, transfers nevertheless peaked at 440Mbps, and transfers between datastore and cluster exceeded 7Gbps. It was necessary to update the Linux kernels with MSI-X support for the network cards otherwise the default kernel would send all the receive interrupts to a single blocking CPU core.

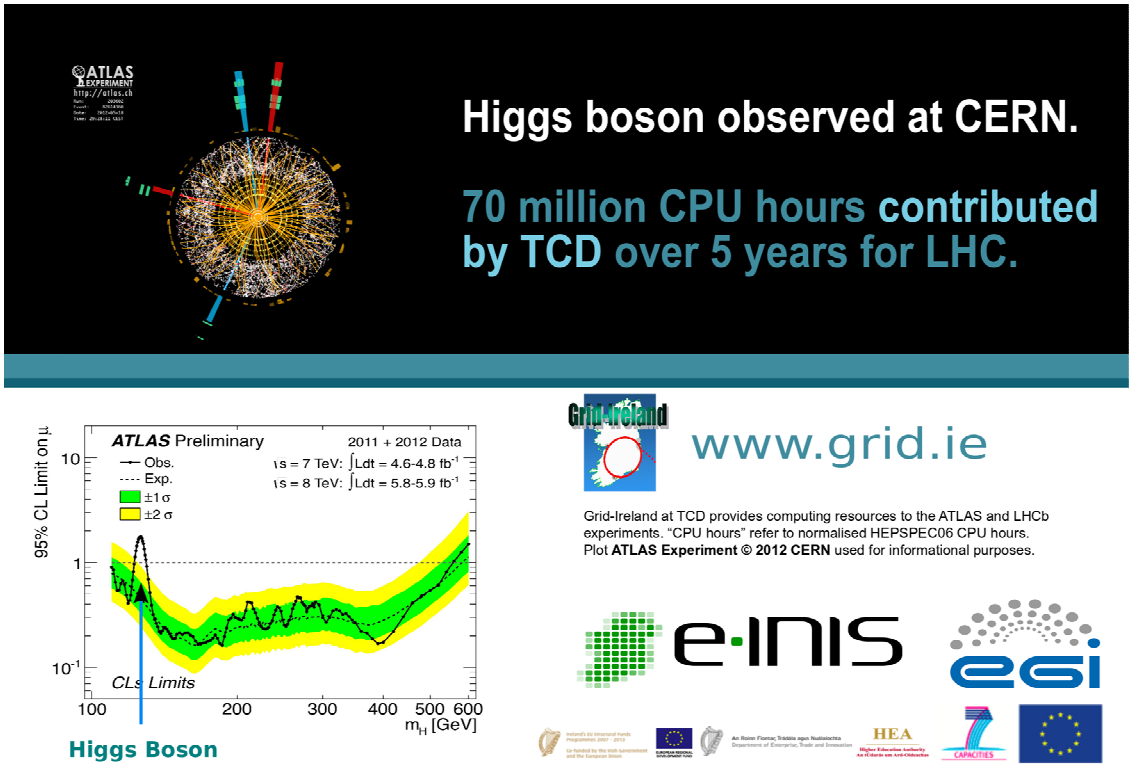

Figure 5: Grid-Ireland csTCDie site contribution to CERN LHC computing

Ultimately, when Grid-Ireland, the computational grid for Ireland, became a casualty of the economic crash of 2008, and was gracefully closed down on 31-Dec-2012, these resources were carefully repurposed in a staged manner as the SCSS Cloud for research and teaching usage.

This was the most substantial of a number of Beowulf clusters constructed by the department, some very production-oriented, others more adventurous, see elsewhere in this catalog.

The homepage for this catalog is at: https://www.scss.tcd.ie/SCSSTreasuresCatalog/
Click *'Accession Index'* (1st column listed) for related folder, or *'About'* for further guidance. Some of the items below may be more properly part of other categories of this catalog, but are listed here for convenience.

Accession Index	Object and Identification
TCD-SCSS-T.20121208.098	csTCDie Grid Site Beowulf Clusters and Datastore, Complex of clusters & storage (1500 cores/600 TB) using 1Gbps Ethernet interconnect and 10Gbps backbone, participant in DataGrid, EGEE, EGI, and CERN LHC computing. From 2013 repurposed as SCSS Cloud, c.2009.
TCD-SCSS-T.20121208.094	Experimental SCSI Cluster, 4-node prototype cluster using SCSI as interconnect, the first cluster constructed in the Department of Computer Science, Trinity College Dublin, and second cluster constructed in the Republic of Ireland, 1997.

TCD-SCSS-T.20121208.095	csTCDie Beowulf Cluster, Departmental cluster using 100Mbps Ethernet as interconnect, the second cluster constructed in the Department of Computer Science, Trinity College Dublin, 1998.
TCD-SCSS-T.20141120.003	csTCDie Grid-Ireland SCI Cluster, 16-node cluster using 400MB/s SCI switched interconnect, the third cluster constructed in the Department of Computer Science, Trinity College Dublin, c.1999.
TCD-SCSS-T.20121208.097	VRengine, 9-node virtual reality engine using 600MB/s SCI 2-d toroidal interconnect, c.2005.
TCD-SCSS-T.20121208.106	csTCDie PS3 Cluster, Ten nodes from a 16-node Sony Playstation PS3 cluster plus build machine, using 1Gbps Ethernet interconnect and running Yellow Dog Linux, c.2009.
TCD-SCSS-T.20121208.099	csTCDie GPU Cluster, 64-core/32-GPU/16-node cluster using 1Gbps Ethernet interconnect, c.2011.

References:

1. Wikipedia, *Beowulf cluster*, see: https://en.wikipedia.org/wiki/Beowulf_cluster

2. Brian Coghlan, JohnWalsh, Stephen Childs, Geoff Quigley, David O'Callaghan, Gabriele Pierantoni, John Ryan, Neil Simon, Keith Rochford, *The Back-end of a 2-Layer Model for a Federated National Datastore for Academic Research VOs that Integrates EGEE Data Managemen*t, Journal of Grid Computing, 8, (2), 2010, p341 – 364.

3. Coghlan, B.A, *OpsCentre 10Ge Optical Networking*, e-INIS Report, 23-Dec-2010,