EDIM1 Progress Report

Paul Martin, Malcolm Atkinson, Mark Parsons,
Adam Carter and Gareth Francis

December 21, 2011

Abstract

The Edinburgh Data-Intensive Machine (EDIM1) is the product of a joint collaboration between the data-intensive group at the School of Informatics and EPCC. EDIM1 is an experimental system, offering an alternative architecture for data-intensive computation and providing a platform for evaluating tools for data-intensive research; a 120 node cluster of 'data-bricks' with high storage yet modest computational capacity. This document gives some background into the context in which EDIM1 was designed and constructed, as well as providing an overview of its use so far and future plans.

1 Context

According to [Hey et al., 2009], there are now four paradigms of scientific research: the *empirical*, the *theoretical*, the *computational* and now that of *data exploration*. Whereas the theoretical paradigm attempts to formalise the rules underpinning empirical observations and the computational paradigm uses modelling and simulation to surpass the limits of human analysis, the *data-intensive* paradigm is concerned with the accumulation, curation, propagation and integration of the vast quantities of data produced by modern scientific instruments; for example, the Sloan Digital Sky Survey[1] generated 120 TB of processed data, whilst the LOFAR project[2] is expected to eventually yield 38 PB *per day*. In essence, researchers in many disciplines are no longer limited by their ability to measure the world around them, or formulate new theories, or even compute new complex models in simulation, but simply by their ability to take hold of and grasp the enormous volumes of information streaming through their fingers.

A significant part of supporting the 'fourth paradigm' of data-intensive research is the development of new tools for managing and operating over data. Another important part is the construction of hardware infrastructure which specifically supports data-intensive computation. It has been realised now that the requirements for high-performance data-intensive computing are not the same as those for traditional high-performance computing. In essence, data-intensive computing emphasises access to storage over processor cycles to the extent that data-intensive research can be better served by making more data available at once to many less powerful processors than by having many stronger

[1] `http://www.sdss.org`
[2] `http://www.lofar.org`

processors but less aggregate data availability [Szalay, 2011]. This is because computation is often enacted broadly over extensive corpora of information rather than in intense depth over relatively small datasets (indeed a significant part of the data-intensive research task is to perform computations over all data in order to identify the specific subsets of data upon which more conventional computational science should be targeted).

The 'classical' approach to large-scale computation usually involves the separation of data from computation (storage distinct from processing) — data is staged from a particular location onto and off of the computation platform as and when needed. As data volumes increase, algorithms fail to scale (often assuming in-memory processing to avoid disk access) and considerable time is spent waiting for data to be delivered to a processor. More recent data-intensive projects favour the use of a 'data brick' architecture (for example [Barclay et al., 2004, Szalay et al., 2009]), where a given node (of a cluster) has significant quantities of private storage alongside a modest processing capacity. Data is shuffled between nodes via network connections as necessary, but where possible algorithms and assorted computational processes are brought to the data.

The Edinburgh Data-Intensive Machine 'EDIM1' is the product of a joint collaboration between the School of Informatics[3] and EPCC[4], funded jointly by the EPSRC (EP/D079829, £100,000) and the University of Edinburgh (£150,000) for staff, equipment and operations. EDIM1 is designed to be more 'Amdahl-balanced' than existing data-intensive machines insofar as it offers the greatest possible capacity for applications to benefit from the parallelisation of any components where potential for such exists [Szalay et al., 2010]. EDIM1 also distinguishes itself by its use of commodity hardware to achieve the computational throughput of a much more expensive machine, being an attempt to bring high-performance data-intensive computation into the reach of smaller institutions and research groups in much the same manner as Beowulf clusters were for conventional compute-intensive tasks.

2 Design

EDIM1 is an experimental machine rather than a service. Its principle purpose is to be used in an extended investigation of data-intensive applications — in particular determining under what circumstances the 'data-brick' architecture used by EDIM1 offers substantial advantages over other more traditional machine-cluster architectures. Thus, the machine does not exist to compete with existing university or national resources (such as HECToR[5]), but instead to assist in learning more about data-intensive applications and to complement the resources already in existence.

Data centres absorb considerable amounts of electrical power, conferring considerable expense and practical limitations on the construction and location of such centres. EDIM1 is designed to have modest power consumption, generally achieved through the use of low-powered processors and solid-state disks for application logic.

[3] `http://www.ed.ac.uk/schools-departments/informatics/`
[4] `http://www.epcc.ed.ac.uk`
[5] `http://hector.ac.uk`

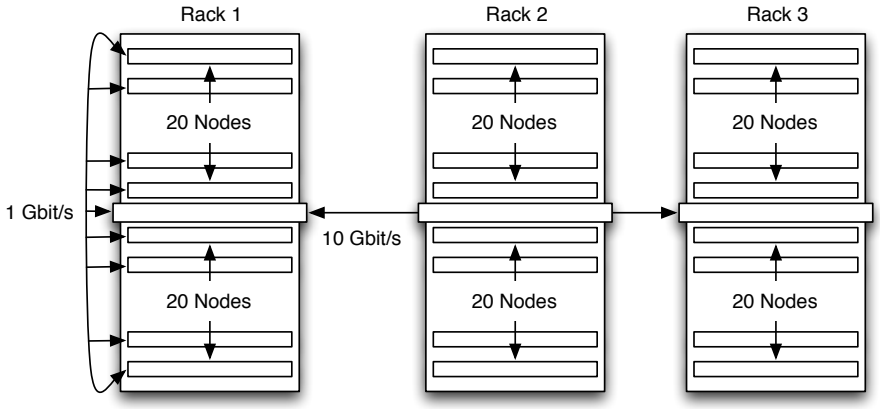

Figure 1: Representation of the physical layout of EDIM1.

Component	**Type**
Processor	2 x AMD Opteron 6128 (8 core 2 GHz)
Memory	32 GB DDR3
Controller	Adaptec 5805 SATA/SATA disk controller
Networking	Intel 10 Gb Ethernet network card
Storage	36 hot-swap drive bays,
	currently 5 x 2 TB HDD (Hitachi Deskstar 7K3000)

Table 1: Specification of EDIM1's data-staging node.

The basic topology of EDIM1 consists of 120 independent nodes arranged into 3 racks of 40 nodes. The nodes within each rack share an internal 1Gbit/s ethernet connection. Racks are connected together via a 10 Gbit/s channel. Access to EDIM1 and its constituent nodes is via a dedicated (and distinct) login node; EDIM1 also has a dedicated data-staging node, specified in Table 1, onto which data to be processed can be placed prior to their deployment onto nodes in the machine. Finally, a testing and development machine exists by which to permit the construction and testing of appliances (as described below) to be used within EDIM1 itself — this machine is a replica of EDIM1 itself, albeit with only a few nodes identical to those specified below.

The basic hardware specification of each node in EDIM1 is as described in Table 2. The choice of components is a balance of capability, cost and power consumption; striking this balance is an important part of fulfilling the goals of the EDIM1 project insomuch as EDIM1 is supposed to demonstrate the applicability of the 'data-brick' architecture to data-intensive computation performed on commodity hardware. The Intel Atom processor is low-powered and inexpensive; the hypothesis here is that other factors (significant storage per node, networking between nodes) will compensate for any loss in raw computational horsepower for data-intensive applications. An NVIDIA Ion graphics processing unit is included in every node in order to benefit from the aptitude of general purpose GPUs (GPGPUs) for low-level parallel computation [Owens et al., 2007]. It was deemed that 4 gigabytes of memory was essential for any serious data

Component	Type
Motherboard	Zotac ION-ITX-K 1.6 GHz Atom/ION Mini-ITX
Processor	Dual-Core Intel 1.6 GHz Atom
Memory	4 GB DDR3
GPU	NVIDIA Ion
Storage	1 x 256 GB SSD (RealSSD C300), 3 x 2 TB HDD (Hitachi Deskstar 7K3000)

Table 2: Specification of an EDIM1 node.

analysis and integration; a significant quantity (256 gigabytes) of solid-state storage provides fast random access for experiments in progress, as well as ensuring the quick loading and operation of a node's operating system and applications; finally, three 2 terabyte hard disks provide substantial storage space for experimental data. Each node shares with each other node in its rack a 1GB/s ethernet connection; each node also has a USB2 connector made available for direct data staging onto the node from an external drive (as an alternative to using the data-staging node, or directly downloading data from the internet).

Applications on nodes are assumed to be transitory, such that a set of nodes can be configured for some task, that task can be executed, and the nodes then reconfigured for some other task (or set of tasks) with minimal effort. Data from experiments is expected to persist on the machine, and is not purged when nodes are reconfigured for new experiments. This allows follow-on experiments to be conducted immediately without having to re-stage all of the data, and is believed to better represent data ecology in scientific domains where it is less expensive to bring computation (in the form of programs and scripts) to existing large data corpora than it is to transfer data to some service.

One of the agreed requirements of any middleware system used in EDIM1 was that it be as lightweight as possible so as to allow proper comparative evaluation of various applications in practice. The administration of EDIM1 and the installation of appliances are conducted using CentOS ROCKS[6], a Linux-based cluster-management system chosen for flexibility and stability. Custom software stacks (appliances) can be installed, replaced and reinstalled at will using ROCKS. This offers considerable flexibility in how different experiments can be accommodated, at the cost of greater preparation time for new types of experiment. Not using a full virtualisation system reduces overhead, which is important for evaluating the architecture and conducting performance comparisons between different experimental configurations.

Monitoring is performed using Ganglia[7], a scalable distributed monitoring system for various types of cluster. Access to EDIM1 is by negotiation with the data-intensive group, or their counterparts in EPCC. A suitable appliance is installed on a selection of nodes and access is then granted for a fixed period to a designated set of users, who are free to use the nodes as needed to execute their experiment.

[6] `http://www.rocksclusters.org`
[7] `http://ganglia.sourceforge.net`

Aug 2010	Permission granted to spend funds on machine granted by EPSRC and the University of Edinburgh. Prototype construction initiated.
Oct 2010	High-level design for machine completed.
Nov 2010	Procurement process initiated and specifications captured.
Feb 2011	Order placed with supplier.
Mar 2011	Amendments to order placed due to discontinued hard disk and motherboard models.
Mar 2011	Machine delivered and installed. Acceptance testing revealed performance issues.
May 2011	Networking upgrade complete (between racks).
Jun 2011	Data-staging node delivered and installed.
Jul 2011	Performance fix instructions for main machine provided by supplier.
	Fix for data-staging node provided by suppliers.
Aug 2011	Final replacement hardware supplied for performance fix.
Oct 2011	Performance fixes completed.

Table 3: Timeline of EDIM1 construction.

3 Application

It was necessary to liaise with a number of individuals both within and without the data-intensive group in order to identify suitable applications to be used as testbeds of the EDIM1 architecture; these liaisons will continue into the future. An overview of active and completed projects involving EDIM1 is given in Table 4. Notable projects include:

Astronomy Researchers from the School of Physics and Astronomy led by Thomas Kitching used EDIM1 to evaluate low-level data reduction of astronomical images, using data from the Canada-France-Hawaii Telescope Legacy Survey (CFHLS)[8]. Approximately 40 terabytes of raw data was transferred onto EDIM1, at which point the machine was used to remove image artefacts, correct for camera distortions imprinted by the telescope optics and remove cosmic ray signatures. EDIM1 was found to be able to complete the task in 75% of the typical time required by a compute-cluster consisting of 200 CPUs with 24 gigabytes of memory shared between every 8 CPUs, using only 80 nodes; much of this success could be attributed to the ease at which the task could be divided evenly between nodes of EDIM1.

Benchmarking During the installation and subsequent investigation of the performance of EDIM1, various benchmarks were explored. The emphasis on the benchmarking was focused on those aspects of performance that were observed as being below the specifications provided by component suppliers, and for this reason most of the effort was concentrated on single-node benchmarks.

[8] `http://www.cfht.hawaii.edu/Science/CFHLS/`

Many of these were low-level (for example reporting the time taken by disk utilities such as `hdparm` and `dd`). Other benchmarking procedures were tested (for example, for Hadoop[9]) but a full systematic benchmarking exercise using these benchmarks was delayed until certain disk performance issues were resolved. Further benchmarking will continue into the future.

Billion Triple Challenge The annual Billion Triple Challenge (part of the Semantic Web Challenge[10]) is concerned with finding new scalable semantic web applications which can effectively make use of a (provided) billion-triple RDF data corpus. A group of Informatics researchers (Andras Salamon, Ewan Klein, Stratis Viglas and Peter Buneman) enlisted EDIM1 as part of their efforts in 2011's challenge, and it is intended that EDIM1 be used for 2012's challenge.

DISPEL The ADMIRE (Advanced Data Mining and Integration Research for Europe) project[11] produced a standard platform for data-intensive computational research, part of which is the DISPEL (Data-Intensive Systems Process Engineering Language) workflow language. The ADMIRE platform is built upon OGSA-DAI (Open Grid Services Architecture — Data Access and Integration)[12], a distributed data access and management system. On EDIM1, several nodes have OGSA-DAI services installed and DISPEL gateways which allow data-intensive workflows to be executed on those nodes, assisting in the development of DISPEL by Malcolm Atkinson, Paul Martin (Informatics), Amy Krause (EPCC), Oscar Corcho (Universidad Politécnica de Madrid) and David Snelling (Fujitsu).

Eurexpress Eurexpress is a transcriptome atlas database for mouse embryones[13] which provides a large corpus of image data which requires data-intensive methods to organise, maintain and analyse. An ongoing project for the data-intensive group at Edinburgh is to train models over 350,000 (40 kilobyte raw) images from Eurexpress in order to perform automatic annotations — a task which requires deep analysis of a considerable portion of database and which EDIM1 is being enlisted to assist. So far, exploratory studies have been performed on a limited number of transcription factors and a limited subset of images, with greater experiments expected to be performed in early 2012 by Paolo Besana (Informatics) in collaboration with Richard Baldock and Ian Overton of the Medical Research Council, Human Genetics Unit.

[9] `http://hadoop.apache.org`
[10] `http://www.challenge.semanticweb.org`
[11] `http://www.admire-project.eu`
[12] `http://www.ogsadai.org.uk`
[13] `http://www.eurexpress.org`

Project	Investigator	Type	Status	Summary
Astronomy	Thomas Kitching	Evaluation	Completed	Evaluation of EDIM1 architecture for gravitational lensing applications in cosmology.
Benchmarking	Gareth Francis and Adam Carter	Benchmark	Ongoing	Exploration of EDIM1 architecture in general; measurements of processing, storage and I/O capabilities.
Billion Triple Challenge	Andras Salamon	Science	Ongoing	EDIM1 has been used in the development of scalable semantic web applications which use a billion-triple RDF data corpus.
DISPEL	Paul Martin	Science	Ongoing	DISPEL and OGSA-DAI services have been installed on EDIM1 in order to perform experiments in distributed knowledge discovery.
Eurexpress	Paolo Besana	Evaluation	Ongoing	An evaluation of processing of a significant corpus of microsopy images on the EDIM1 architecture.
VERCE	Alessandro Spinuso and Luca Trani	Science	Ongoing	Prototyping of services for the VERCE project; specifically the cross-correlation use-case.
OMERO	Donald MacDonald	Science	Early 2012	Investigation of image processing and subsampling for OMERO within EDIM1.
OSDC (Sector/Sphere)	Gareth Francis	Evaluation	Early 2012	Evaluation of Sector/Sphere on EDIM1 with aim towards a contribution to the Open Science Data Cloud.
SAGA	Ole Weidner	Evaluation	Early 2012	Evaluation of the use of SAGA on EDIM1.

Table 4: Completed, ongoing and pending research conducted using EDIM1.

Project Title	Student	Supervisor	Type	Summary
Scientific applications: exploiting the data bonanza. The microscopy case – using Rasdaman	Vamsi Kalyan Kamini	Paolo Besana	Evaluation	Explored the use of Rasdaman on EDIM1 as a means for more efficient storage and processing of biological image data [Kamini, 2011].
Scientific applications: exploiting the data bonanza. The microscopy case	Petros Parakevopoulos	Paolo Besana	Evaluation	Explored the use of Hadoop on EDIM1 as an alternative to OMERO for processing biological image data [Parakevopoulos, 2011].
Benchmarking an Amdahl-balanced cluster for data intensive computing	Omkar Kulkarni	Adam Carter	Benchmark	Evaluated EDIM1 as an Amdahl-balanced architecture both theoretically and synthetically [Kulkarni, 2011].
Programming abstractions for dynamic, distributed, data-intensive computing	Vinay Sudhakaran	Neil Chue Hong	Evaluation	Evaluated the use of Hadoop on EDIM1 for the cccgistemp implementation of the NASA GISS model for estimating global temperature change [Sudhakaran, 2011].
Testing high level astronomical image analysis	Mengda He	Thomas Kitching	Evaluation	Investigation of Hadoop on EDIM1 as a means to speed up the analysis of images using an algorithm for measuring the shape of galaxies in order to extract cosmological parameters.

Table 5: Overview of MSc projects conducted using EDIM1 in Summer 2011.

OMERO OMERO[14] is client-server software for the visualisation, management and analysis of biological microscope images. Very large (100K by 100K pixel) images can be uploaded and viewed within the OMERO server; OMERO converts these images into a pyramid-structure of consecutively sub-sampled images in order to efficiently send the rendered and scaled images to the user. This conversion process can take upwards of 30 minutes per image however, and EDIM1 will be used to investigate how best to distribute it on the data-brick architecture; if a significant improvement can be attained, then this will provide evidence for the suitability of the architecture for a wide range of analogous data-intensive problems. This project will be led by Donald MacDonald (Informatics) and Jason Swedlow (University of Dundee).

VERCE VERCE (the Virtual Earthquake and seismology Research Community e-science environment in Europe)[15] is making use of EDIM1 both as a test platform for prototypes of tools and services for computational seismology (using the ADMIRE platform) and as a possible resource in and of itself. Work on EDIM1 for VERCE is jointly led by Alessandro Spinuso and Luca Trani (Informatics / Koninklijk Nederlands Meteorologisch Instituut).

MSc Projects Thus far EDIM1 has been used in four MSc projects as summarised in Table 5. These projects have drawn upon framework technologies such as Hadoop and OMERO, database technologies such as Rasdaman[16], and codebases such as the Clear Climate Code[17] reimplementation of the GISS Surface Temperature Analysis scheme (GISTEMP)[18].

4 Discussion and Future Work

From our investigations of EDIM1, we have been able to make the following observations:

- Building a machine of this nature is not a trivial task. The use of cheaper, commodity processors and motherboards alongside more expensive disks can lead to problems with compatibility and reliability. Most of the compatibility problems experienced related only to performance, but the effects of this can be significant. Performance problems have now been resolved, but only after having liaised with the manufacturers of some of the hardware.

- Whilst it can usually be made to run on the EDIM1 with little or no modification, application code typically needs to be refactored to some extent to make the most of the architecture. Even when the amount of I/O is balanced to the amount of computation required, the I/O access patterns often do not fully exploit the machine. In order to fully exploit many disks, problems which are relatively small in terms of computation required have to be split between many processors. This requires the

[14] `http://www.openmicroscopy.org`
[15] `http://www.verce.eu`
[16] `http://www.rasdaman.com`
[17] `http://clearclimatecode`
[18] `http://data.giss.nasa.gov/gistemp/`

application to exhibit good scaling behaviour. This is a well-known issue in parallel computing but is potentially exacerbated by this architecture.

- Related to the previous point, it has been found to be more difficult than expected to find existing applications which benefit from EDIM1's architecture without modification. This is not due to a lack of data-intensive problems, but is more attributable to current coding practices which try to avoid intensive I/O. Close collaboration with scientists is required in order to frame problems in a manner which best makes use of this type of computational architecture.
- We tried to do something novel in terms of software configuration on the machine, rather than configuring it as a 'typical' high-performance cluster. In order to avoid moving around large volumes of data on EDIM1, we looked for ways to move computation to the data. The solution we went for (ROCKS) satisfied all requirements of flexibility and performance, but has proven difficult and time-consuming to configure and use in practice.
- Getting data onto a machine such as this is still hard. Thought needs to be put into the complete application workflow from the point where data is collected (*e.g.* telescopes or seismometers) to the point where it is ultimately analysed, visualised or used (which is probably a scientist's desktop machine). Dealing with the volume of data which *could* be processed on a machine such as EDIM1 requires both a high-bandwidth connection from the machine to JANET, but also sufficiently high bandwidth back-bone connections. In practice, a lot of data transfer is still done by transporting physical disks to the machine – a fairly labour-intensive and inelegant procedure.

Nevertheless, there remain a number of ongoing projects making use of EDIM1, and future projects are expected:

- There are plans for EDIM1 to be used within EUDAT (European Data Infrastructure)[19].
- As part of a possible contribution to the OSDC (Open Science Data Cloud), the use of Sector/Sphere[20] is to be tested on EDIM1.
- The use of SAGA (Simple API for Grid Applications)[21] is also to be tested on EDIM1.
- EDIM1 will figure prominently in future collaborations between the Data-Intensive Group within the School of Informatics and AIST (national institute of Advanced Industrial Science and Technology) in Japan.

EDIM1 is also figuring in new project proposals in preparation for submission in January 2012 — for example in:

- Astroinformatics in Cosmology, a European Research Council Synergy Grant application, involving investigators from both the School of Informatics and the School of Physics and Astronomy at the University of Edinburgh.

[19] `http://www.eudat.eu`
[20] `http://sector.sourceforge.net`
[21] `http://saga.cct.lsu.edu`

- Multiplying the Applications of Intelligence from Data for the benefit of European Nations (MAIDEN), a project seeking to use technologies such as DISPEL and ADMIRE in industrial-scale knowledge discovery applications on a variety of platforms.
- Mining Accurate Information from Data Securely, a project seeking to extend the ADMIRE platform, with particular interest in increasing the robustness and security of the platform.

References

[Barclay et al., 2004] Barclay, T., Gray, J., and Chong, W. (2004). Terraserver bricks — a high availability cluster alternative. Technical Report MSR-TR-2004-107, Microsoft Research.

[Hey et al., 2009] Hey, T., Tansley, S., and Tolle, K., editors (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research.

[Kamini, 2011] Kamini, V. K. (2011). Scientific applications: exploiting the data bonanza. the microscopy case – using rasdaman. Master's thesis, School of Informatics, the University of Edinburgh.

[Kulkarni, 2011] Kulkarni, O. (2011). Benchmarking an amdahl-balanced cluster for data intensive computing. Master's thesis, Edinburgh Parallel Computing Centre, the University of Edinburgh.

[Owens et al., 2007] Owens, J. D., Luebke, D., Govindaraju, N., Harris, M., Krüger, J., Lefohn, A. E., and Purcell, T. J. (2007). A survey of general-purpose computation on graphics hardware. *Computer Graphics Forum*, 26(1).

[Parakevopoulos, 2011] Parakevopoulos, P. (2011). Scientific applications: exploiting the data bonanza. the microscopy case. Master's thesis, School of Informatics, the University of Edinburgh.

[Sudhakaran, 2011] Sudhakaran, V. (2011). Programming abstractions for dynamic, distributed, data-intensive computing. Master's thesis, Edinburgh Parallel Computing Centre, the University of Edinburgh.

[Szalay et al., 2009] Szalay, A., Bell, G., Vandenberg, J., Wonders, A., Burns, R., Fay, D., Heasley, J., Hey, T., Nieto-SantiSteban, M., Thakar, A., van Ingen, C., and Wilton, R. (2009). Graywulf: Scalable clustered architecture for data intensive computing. In *42nd Hawaii International Conference on System Sciences*.

[Szalay, 2011] Szalay, A. S. (2011). Extreme data-intensive scientific computing. *Computing in Science and Engineering*, 13(6).

[Szalay et al., 2010] Szalay, A. S., Bell, G. C., Huang, H. H., Terzis, A., and White, A. (2010). Low-power amdahl-balanced blades for data intensive computing. *ACM SIGOPS Operating Systems Review*, 44:71–75.