DiGS: Distributed Grid Storage for UKQCD

DiGS is a grid application that combines disparate storage resources to form a unified 'data grid', capable of meeting the data management challenges of both QCD Physics and a wider scientific community.

Overview and Application

DiGS is a data management solution, developed by software engineers at EPCC, which combines disparate mass storage technologies (for example, RAID units or SAN systems) to provide a unified, multi-Terabyte facility, commonly referred to as a *data grid*. The generic nature of the DiGS functionality has been crucial to its application to a number of scientific fields, where there is a common requirement to archive terabytes of source data reliably, and provide secure access to the data from geographically dispersed locations.

The most significant installation of DiGS has been completed by the UK QCD Physics community (UKQCD), who use the system to archive valuable Lattice Gauge Configurations generated on the QCDOC supercomputer. They have a deployment of DiGS across seven sites in the UK and USA, furnishing approximately 80TB of capacity and – at the time of writing – hosting approximately 50,000 datasets.

DiGS is also being tested within the cell biology community. The Wellcome Trust for Cell Biology are, in consultation with EPCC, completing a scoping activity to establish the feasibility of deploying a DiGS-powered environment to collate microscope image data from centres in Scotland and North America.

Deployment Architecture

DiGS is built on top of the Globus Toolkit and elements of the EGEE application stack. At the heart of DiGS is the 'Control Thread', a service that continuously monitors the integrity of content on the grid and orchestrates data replication/back-up operations. The Control Thread is supported by the File Catalogue, which tracks the location of the (possibly many) instances of the data on the grid. The actual data is held on Storage Elements interfaced with GSI-FTP protocol for high speed file transfer. Complementing the File Catalogue is the Metadata Catalogue that maintains a database of scientific provenance for the individual datasets. User interaction with the system is secured using X.509-based authentication and message level security. To simplify administration, DiGS-powered resources synchronise authorisation information with a central EGEE VOMS server.

Looking to the future, we will introduce second generation Storage Elements, built on top of the SRM interface. We will also pursue greater interoperability with EGEE gLite grids.

DiGS is the new name for QCDgrid, the data management system for UKQCD.

References

[1] DiGS: http://forge.nesc.ac.uk/projects/qcdgrid/
[2] EGEE: http://www.eu-egee.org/
[3] Globus: Toolkit – http://www.globus.org/toolkit/
[4] QCDOC Supercomputer – http://ukqcd.epcc.ed.ac.uk/community/qcdoc/
[5] Wellcome Trust for Cell Biology – http://www.wcb.ed.ac.uk/

Figure 1: Deployment architecture for DiGS data grid.

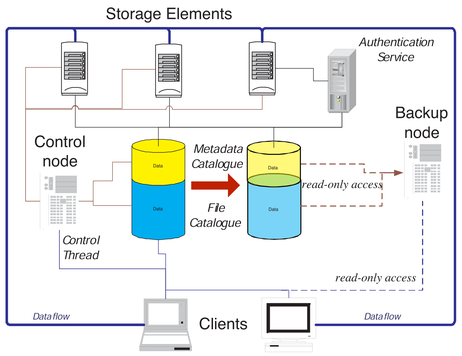
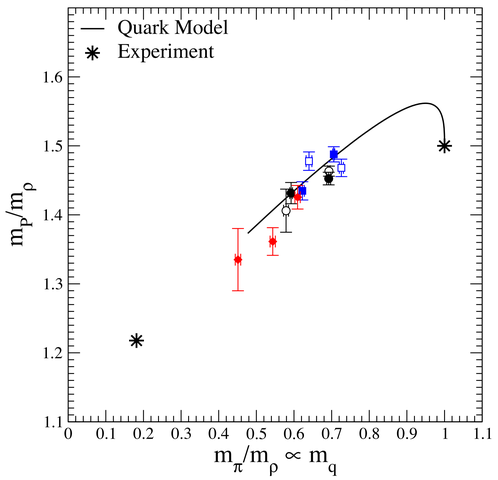


Figure 2: The Edinburgh plot showing the ratio of the proton mass to the rho meson mass as a function of quark mass. The solid line is a theoretical model. Symbols with different colours denote lattice QCD data at different lattice spacings, showing data that goes beyond the quark model for the first time. All data generated on QCDOC and stored on a DiGS data grid.



iti Techmedia

Condition-based monitoring in commercial farming

EPCC is working with eight other academic and commercial technology providers in ITI Techmedia's Condition-Based Monitoring (CBM) Programme, investigating the application of CBM technologies to commercial farming.
EPCC leads the biological modelling work, applying expertise in software, data management and data analysis to develop key intellectual property for the ITI CBM platform.

Condition-Based Monitoring

Condition-based monitoring (CBM) involves the collection and analysis of real-time data from a system and the consequent detection or prediction of the system's condition. CBM isn't new – process engineering, aircraft, modern cars all rely on these techniques to maintain safety or enhance performance – but recent strides in technologies such as low-cost, low-power sensors and ubiquitous wireless networks have broadened the horizons for CBM applications enormously.

CBM typically makes use of machine learning, expert systems and agent-based techniques to capture the normal expected behaviours of a monitored system and then tracks sensor measurements against their expected norms to predict the way the system will behave. The aim is to determine when the system begins to demonstrate anomalous behaviour, ideally well before it enters a failure state. In this way predictive maintenance can be used to increase system reliability and reduce upkeep costs at the same time.

Sensor-rich environments

Analyst predictions in 2006 suggest that the market for low-power, low-cost radio frequency identification technologies (RFID) is poised to explode. Couple this with the increasing availability of low-cost sensors such as silicon accelerometers and the ability to gather data from pretty much anywhere is just around the corner.

This is the so-called 'Sensor Grid', the next stage in the wide-area computing architecture that defines Grid computing. Leaving legal and ethical considerations aside this poses two key technical questions:
- how do we manage the massive increase in data these sensors will generate, and;
- what sense can we make of them?

Any CBM platform wishing to leverage the potential of a sensor-rich environment needs intelligent answers to both these questions.

Figure 1 illustrates the architecture of a potential Sensor Grid-based condition monitoring system. Wireless networks transmit data from groups of sensors to a hub which passes them along to the CBM platform itself. This platform forwards data to a large-scale data store while streaming the same data through its condition monitoring rules. Any significant changes in conditions detected trigger alerts to designated PDAs or other mobile devices, alerting operators that intervention is required. Meanwhile, new data in the storage array are being cycled regularly through a machine learning engine which provides for adaptive evolution of the detection algorithms within the CBM platform.

Working with ITI Techmedia

ITI Techmedia are an organisation which develops market-driven intellectual property in the areas of digital media and communications technology for the benefit of Scotland. ITI Techmedia are one of the three Intermediary Technology Institutions set up by the Scottish Executive with £450M over 10 years to develop a range of pre-competitive technologies with global market potential.

The Condition-Based Monitoring Programme (CBMP) is ITI Techmedia's sixth research and development programme. CBMP is a three-year, £4.75M programme which aims to apply sensors and networks technology to develop a condition-based monitoring platform with applications across a range of industries.

The Programme is initially focused on the commercial farming sector where the technology under development will enable the behavioural and physiological monitoring of beef and dairy cattle, allowing the animals' conditions to be tracked continuously. It brings together academic experts from computer science, veterinary science, farm husbandry and electronic engineering with technology and manufacturing firms to provide a broad church of expertise.

Fitter, healthier, more productive

EPCC have been working closely with groups at the University of Strathclyde, Scottish Agricultural College and Royal (Dick) School of Veterinary Studies at the University of Edinburgh to prepare the ground for the interpretation of large volumes of novel sensor data. Ongoing farm studies have begun to generate the first of potentially terabytes of sensor data from trial cows, and the challenge for data analysis experts at EPCC and Strathclyde is to search these large data volumes to find links from sensor readings to cows' conditions. The initial focus in this area will be on generating and processing as much data as possible for detailed analysis.

Figure 1: Architecture of a potential sensor Grid, based on the CMB platform.