

Close encounters with clusters of computers

To satisfy their ever-increasing demand for more and affordable computing power, particle physics experiments are using clusters of off-the-shelf PCs. A recent workshop at Fermilab looked at the implications of this move.

Recent revolutions in computer hardware and software technologies have paved the way for the large-scale deployment of clusters of off-the-shelf commodity computers to address problems that were previously the domain of tightly coupled multiprocessor computers. Near-term projects within high-energy physics and other computing communities will deploy clusters of some thousands of processors serving hundreds or even thousands of independent users. This will expand the reach in both dimensions by an order of magnitude from the current, successful production facilities.

A Large-Scale Cluster Computing Workshop held at Fermilab earlier this year examined these issues. The goals of the workshop were:

- to determine what tools exist that can scale up to the cluster sizes foreseen for the next generation of HENP experiments (several thousand nodes) and by implication to identify areas where some investment of money or effort is likely to be needed;
- to compare and record experiences gained with such tools;
- to produce a practical guide to all stages of planning, installing, building and operating a large computing cluster in HENP;
- to identify and connect groups with similar interest within HENP and the larger clustering community.

Thousands of nodes

Computing experts with responsibility and/or experience of such large clusters were invited. The clusters of interest were those equipping centres of the sizes of Tier 0 (thousands of nodes) for CERN's LHC project or Tier 1 (at least 200–1000 nodes) as described in the MONARC (Models of Networked Analysis at Regional Centres for LHC Experiments) project at "<http://monarc.web.cern.ch/MONARC>". The attendees came not only from various particle physics sites worldwide but also from other branches of science, including biomedicine and various Grid computing projects, as well as from industry.

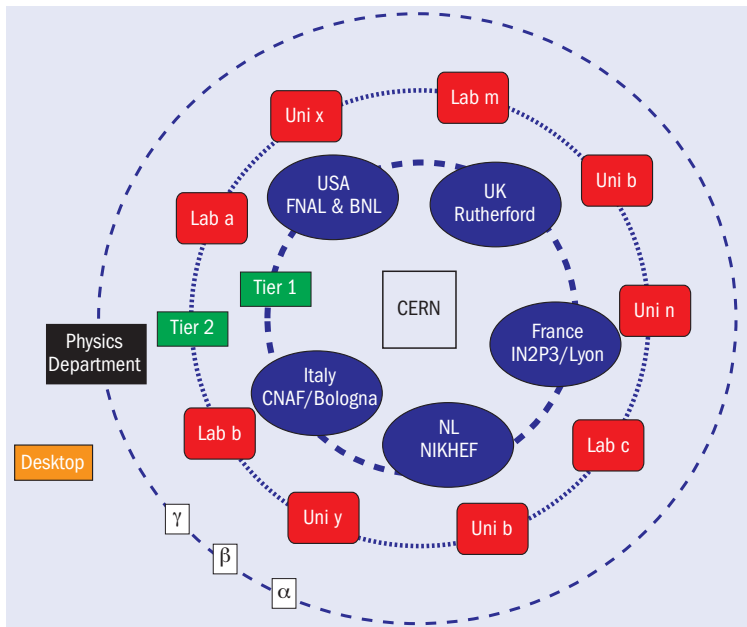
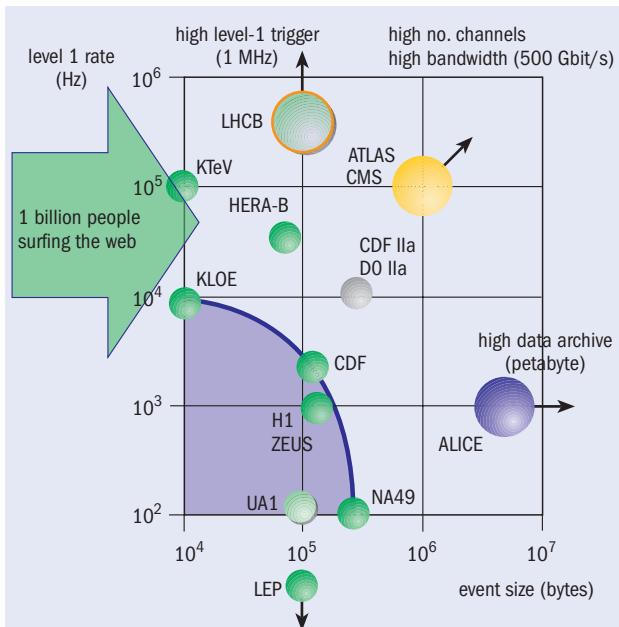
The attendees shared freely their experiences and ideas, and pro-



Presenting LHC computing needs – Wolfgang von Rueden, head of the Physics Data Processing Group in CERN's Information Technology Division.

ceedings are currently being edited from material collected by the convenors and offered by attendees. In addition the convenors, again with the help of material offered by the attendees, are in the process of producing a guide to building and operating a large cluster. This is intended to describe all phases in the life of a cluster and the tools used or planned to be used. This guide should then be publicized (made available on the Web and presented at appropriate meetings and conferences) and regularly kept up to date as more experience is gained. It is planned to hold a similar workshop in 18–24 months to update the guide. All of the workshop material is available via "<http://conferences.fnal.gov/lccws>". ▷

COMPUTING



Left: Particle physics experiments spearhead the demand for more computing power. Right: clustered computers. The MONARC project is a multilayered approach based on a large central site to collect and store raw data (Tier 0 at CERN), with multitiers (for example National Computing Centres, Tier 1), down to individual user's desks (Tier 4), each with data extracts and/or data copies and performing different stages of physics analysis.

The meeting began with an overview of the challenge facing high-energy physics. Matthias Kasemann, head of Fermilab's Computing Division, described the laboratory's current and near-term scientific programme, including participation in CERN's future LHC programme, notably in the CMS experiment. He described Fermilab's current and future computing needs for its Tevatron collider Run II experiments, pointing out where clusters, or computing "farms" as they are sometimes known, are used already. He noted that the overwhelming importance of data in current and future generations of high-energy physics experiments had prompted the interest in Data Grids. He posed some questions for the workshop to consider:

- Should or could a cluster emulate a mainframe?
- How much could particle physics computer models be adjusted to make most efficient use of clusters?
- Where do clusters not make sense?
- What is the real total cost of ownership of clusters?
- Could we harness the unused power of desktops?
- How can we use clusters for high I/O applications?
- How can we design clusters for high availability?

LHC computing needs

Wolfgang von Rueden, head of the Physics Data Processing group in CERN's Information Technology Division, presented the LHC computing needs. He described CERN's role in the project, displayed the relative event sizes and data rates expected from Fermilab Run II and from LHC experiments, and presented a table of their main characteristics, pointing out in particular the huge increases in data expected and consequently the huge increase in computing power that must be installed and operated.

The other problem posed by modern experiments is their

geographical spread, with collaborators throughout the world requiring access to data and computer power. Von Rueden noted that typical particle physics computing is more appropriately characterized as high throughput computing as opposed to high performance computing.

The need to exploit national resources and to reduce the dependence on links to CERN has produced the MONARC multilayered model. This is based on a large central site to collect and store raw data (Tier 0 at CERN) and multitiers (for example National Computing Centres, Tier 1 - examples of these are Fermilab for the US part of the CMS experiment at the LHC and Brookhaven for the US part of the ATLAS experiment), down to individual user's desks (Tier 4), each with data extracts and/or data copies and each one performing different stages of physics analysis.

Von Rueden showed where Grid Computing will be applied. He ended by expressing the hope that the workshop could provide

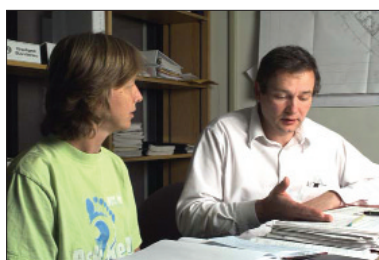
The other problem posed by modern experiments is their geographical spread, with collaborators throughout the world requiring access to data and computer power

answers to a number of topical problem questions, such as cluster scaling and making efficient use of resources, and some good ideas to make progress in the domain of the management of large clusters.

The remainder of the meeting was given over to some formal presentations of clustering as seen by some large sites (CERN, Fermilab and SLAC) and also from small sites without on-site accelerators of their own



Two "farms" of PCs at Fermilab. Such clusters have largely replaced multiprocessor mainframes.



Fermilab Computing Division head Matthias Kasemann discusses computing requirements for the CMS experiment at CERN's LHC with Vivian O'Dell.

The Grid: crossing borders and boundaries

The World Wide Web was invented at CERN to exchange information among particle physicists, but particle physics experiments now generate more data than the Web can handle. So physicists often put data on tapes and ship the tapes from one place to another – an anachronism in the Internet era. However, that is changing, and the US Department of Energy's new Scientific discovery through advanced computing program (SciDAC) will accelerate the change.

Fermilab is receiving additional funds through SciDAC, some of which will be channelled into Fermilab contributions to the Compact Muon Solenoid Detector (CMS) being built for CERN. A major element in this is the formulation of a distributed computing system for widespread access to data when CERN's LHC Large Hadron Collider begins operation in 2006. Fermilab's D0 experiment has established its own computing grid called SAM, which is used to offer access for

experiment collaborators at six sites in Europe.

With SciDAC support, the nine-institution Particle Physics DataGrid collaboration (Fermilab, SLAC, Lawrence Berkeley, Argonne, Brookhaven, Jefferson, CalTech, Wisconsin and UC San Diego) will develop the distributed computing concept for particle physics experiments at the major US high-energy physics research facilities. Both D0 and US participation in the CMS experiment for the LHC are member experiments. The goal is to offer access to the worldwide research community, developing "middleware" to make maximum use of the bandwidths available on the network.

The DataGrid collaboration will serve high-energy physics experiments with large-scale computing needs, such as D0 at Fermilab, BaBar at SLAC and the CMS experiment, now under construction to operate at CERN, by making the experiments' data available to scientists at widespread locations.

(NIKHEF in Amsterdam and CCIN2P3 in Lyon). However, the largest part of the workshop was a series of interactive panel sessions, each seeded with questions and topics to discuss, and each introduced by a few short talks. Full details of these and most of the overheads presented during the workshop can be seen on the workshop Web site.

Many tools were highlighted: some commercial, some developed locally and some adopted from the open source community. In choosing whether to use commercial tools or develop one's own, it should be noted that so-called "enterprise packages" are typically priced for commercial sites where downtime is expensive and has quantifiable cost. They usually have considerable initial installation and integration costs. However, one must not forget the often high ongoing costs for home-built tools as well as vulnerability to personnel loss/reallocation.

Discussing the G word

There were discussions on how various institutes and groups performed monitoring, resource allocation, system upgrades, problem debugging and all of the other tasks associated with running clusters. Some highlighted lessons learned and how to improve a given procedure next time. According to Chuck Boenheim of SLAC, "A cluster is a very good error amplifier."

Different sites described their methods for installing, operating and administering their clusters. The G word (for Grid) cropped up often, but everyone agreed that it was not a magic word and that it would need lots of work to implement something of general use.

One of the panels described the three Grid projects of most relevance to high-energy physics, namely the European DataGrid project and two US projects – PPDG (Particle Physics Data Grid) and GriPhyN (Grid Physics Network).

A number of sites described how they access data. Within an individual experiment, a number of collaborations have worldwide "pseudo-grids" operational today. In this context, Kors Bos of NIKHEF, Amsterdam, referred to the existing SAM database for the D0 experiment at Fermilab as an "early-generation Grid". These already point toward issues of reliability, allocation, scalability and optimization for the more general Grid.

Delegates agreed that the meeting had been useful and that it should be repeated in approximately 18 months. There was no summary made of the Large-Scale Cluster Computing Workshop, the primary goal being to share experiences, but returning to the questions posed at the start by Matthias Kasemann, it is clear that clusters have replaced mainframes in virtually all of the high-energy physics world, but that the administration of them is particularly far from simple and poses increasing problems as cluster sizes scale. In-house support costs must be balanced against bought-in solutions, not only for hardware and software but also for operations and management. Finally, delegates attending the workshop agreed that there are several solutions for, and a number of practical examples of, the use of desktops to increase the overall computing power available.

Alan Silverman, CERN and **Dane Skow**, Fermilab.