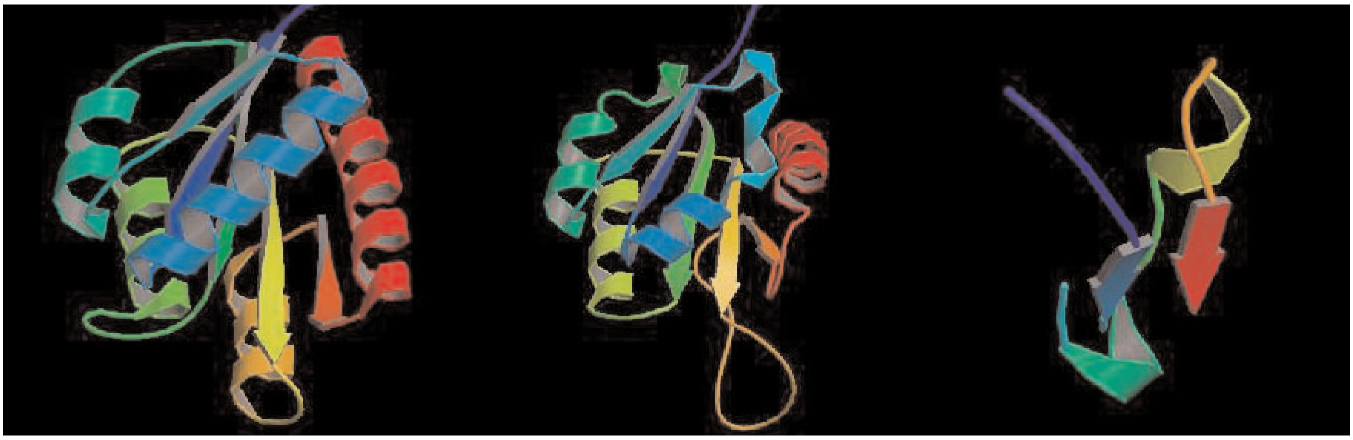
PDC Newsletter No 1 - 2003



Example of a description from a recent protein model Pmodeller (center) compared to an older model to the right and an experimentally determined model to the left.

Bioinformatics: Biology Plus Computer Science

The publication of the human genome in June 2000 was the culmination of more than a decade of work. The prime motive for the human genome project was clinical and therapeutic applications. Knowing the human genome changed the sciences of biology and medicine dramatically. It has fundamentally increased the information about the processes of life, but it has also implied handling of enormous amounts of data. The development of methods and computer technology were key factors to support the completion of the analysis needed to establish the human genetic map.

The publication of the human genome structure was a spectacular event, but the incentive to join biology and computer science was initiated already before that, especially since the first complete genome sequence was published in 1995.

The techniques used to analyze the genome have become more automated and time effective during recent years. High levels of automatization of DNA analyses have also been achieved, as described on page 6.

The capacity for processing and storing data as well as developing sophisticated mathematical methods therefore has been a major driving force in the successive development in what is now established as a specific field of science called bio-informatics. Huge amounts of data were generated and data from both measurements and calculations expanded extremely rapidly when scientists began to unravel the genomes of different organisms.

Bioinformatics is a fast-developing field, devoted to the interpretation of biological information related to DNA and proteins. It includes development of theoretical methods and data analysis focused on functional and comparative genome research, proteomics, structural genomics, system biology, and related fields.

…Continued on page 3

Global User Communities

PDC's mission is to facilitate the advance of Swedish research and education in disciplines that require high-end computational or storage resources. Traditionally, these disciplines have almost exclusively been within the sciences and engineering. However, due to improved instruments in many disciplines, digital technology, and pervasive communications networks the need for high-end resources is now emerging in a broad range of disciplines, due to rapidly growing data collections and improved models describing underlying phenomena. Concurrently, improved communication has made most disciplines global in the sense that collaborators are increasingly chosen based on shared interests rather than on geographic location, and research is pursued using data collections or unique instruments regardless of location.

This is clearly illustrated by the theme of this edition of our newsletter. Bioinformatics is an emerging discipline in which basic information produced in several laboratories around the world are used by a large number of teams to generate knowledge about structure and function of macromolecular complexes. Biology traditionally has been pursued in thousands of laboratories worldwide with little or modest sharing of data and information through means other than traditional publication in journals and conferences. However, this discipline is changing as well with digital data sets increasingly shared, as in the Encyclopedia of Life project (http://eol.sdsc.edu). Another example is various biomedical imaging efforts, such as the Biomedical Informatics Research Network (BIRN, http://birn.ncrr.nih.gov/birn). PDC has engaged in a number of activities in the past to prepare for this increasing emphasis on data, distributed collaboration and use of remote instruments for scientific discovery in its user community. The operation of high-performance computing and substantial storage recourses for the Swedish research community can no longer be viewed as a local activity, but need to be viewed as an element in a larger integrated environment, popularly often referred to as Grids as described in our previous newsletter. Grids seek to provide convenient and secure access to collections of resources such as computers, storage, data collections, instruments and sensor networks regardless of locations.

Finally, some tangible good news for PDC users. This summer we will make available a new cluster based on the IA-64 architecture that is widely expected to dominate computing systems for capability computing in the not too distant future. The PDC cluster has about two-thirds of a Tflop peak performance and in excess of half a TB of primary storage. Several recently acquired large scale systems for the scientific community are also based on this architecture, such as the terascale systems acquired by the NSF TeraGrid, the Pacific Northwest National Laboratory, the Ohio State Supercomputer Center and Rice University to mention a few.

To offer users opportunities to learn about these new developments several educational and training events and workshops are being developed. In July a Grid Summer School will take place in Italy, and in August NGSSC will offer a Grid course. Details of these opportunities appear elsewhere in this issue. We look forward to working with the Swedish academic community in the exciting and challenging transformation ahead.

Lennart Johnsson,
Director of PDC

In This Issue

PDC Newsletter No 1 - 2003

Published by PDC at KTH.
The center operates leading-edge, high-performance computers as easily accessible national resources. These resources are primarily available for Swedish academic research and education. PDC receives funding from the Swedish Research Council (VR), and KTH.
Publisher: Lennart Johnsson
Editors: Mats Höjeberg, Gert Svensson
E-mail: pdc-newsletter@pdc.kth.se
ISSN 1401-9671

To Contact PDC

Visiting address: Lindstedtsvägen 3, plan 5, KTH, Stockholm
Mail: PDC, KTH, SE-100 44, Stockholm, Sweden
E-mail: pdc-staff@pdc.kth.se
WWW: http://www.pdc.kth.se
Phone: +46 8 790 78 00
Fax: +46 8 24 77 84

...from page 1

Bioinformatics deals with analyzing, interpreting, organizing, and storing the enormous amounts of data from sequences and associated information, such as mutations, polymorphisms, expression patterns, three-dimensional structures, protein–protein interactions, metabolic pathways, just to mention a few examples. This research area requires interaction between computer science, mathematics, statistics, chemistry, biology, and medicine.

The Purpose and Challenges of Bioinformatics

The challenge is to assure access to the data archives of genomes and proteins, develop tools to work with these archives, and to formulate relevant scientific problems that can be addressed using the available data and tools.

Bioinformatics reflects a shift in the biological sciences from being observational to being deductive, and also becoming quantitative and precise in great detail. Our detailed knowledge of the human genetic map provides the foundation for what many hope will be a revolution in medical diagnosis and treatment.

Among other things, it opens up new ways to analyze the structure of proteins, which is critical knowledge for the development of new pharmaceuticals. The field also includes applications in biological information processing and modeling biological and behavioral processes. Today, bioinformatics is an indispensable part of all kinds of high throughput biology. In bioinformatics knowledge about relevant biological problems are combined with both physical-chemical and logical competence, often in close co-operation with experimentalists.

Cooperation Between SBC and PDC

Bioinformatics started to infect PDC in 2000 as described in the PDC Newsletter 1, 2000, page 4. The Stockholm Bioinformatics Center, SBC, a joint SU–KTH–KI initiative, was awarded a five-year grant from the Foundation for Strategic Research, SSF, in 1999. Cooperation between SBC and PDC was initiated by Gunnar von Heijne, head of the SBC, and Björn Engquist, then director of PDC in 2000. SBC is a national facility designed to provide a critical mass of bioinformatics and computer science expertise for high-level research, methods development, and advanced post-graduate training in bioinformatics. SBC and PDC were awarded a research grant from the Knut and Alice Wallenberg Foundation to build a cluster of PCs which has now reached a capacity of over 600 Gflops with more than 150 GB of prime memory distributed among more than 200 processors.

Research at SBC is focused on protein sequence and structure, molecular evolution, and modeling of cellular function. This implies analyzing and making predictions about protein structure from the amino acid sequence, aiding drug design for pharmaceutical development where analysis and comparison of whole-genome data are important.

Since this program was initiated in 2000, some important results have emerged, including the development of new methods for subcellular protein prediction, protein structure prediction, phylogeny reconstruction, and predictions of metabolic networks. Further information and references are available at the SBC website, http://www.sbc.su.se.

PDC Services for Handling Large Amounts of Data

The biological data are aggregated from genomes or DNA to small molecules in rapidly increasing quantities. For genomes there is now relatively complete and accurate data available. These can be used in predictive studies of more complex organisms such as proteins and other substances in models that have become reliable and effective. Such models make it possible to perform high throughput studies, which are necessary in order to provide information to the pharmaceutical development and also to the understanding of processes in biological systems and living organisms.

There is a need to handle, store, integrate, and analyze the data sets both from experiments in order to implement immediate interpretation of obtained data sets from theoretical work to develop tools for aggregate and predictive analysis of available information in databases all over the world.

A crucial problem, still approached by relatively few research groups, is to achieve effective methods of combining and comparing sequences of genomes or proteins. Combining several methods and comparing many combinations of methods have recently improved the capability to describe more complex biological systems. The mathematical methods are logically stable and the biological description reasonably accurate in order to get reliable predictions, meeting the requirement of further development in understanding biological systems and even living organisms. But the demand for computational skill and operational capacity also increases rapidly. This is the basic idea for the cooperation between SBC and PDC.

PDC has the capacity to house and operate large computer systems and to store very large amounts of data and to make them easily accessible for distributed use. Both these PDC capabilities are used by SBC for which PDC houses and operates a PC cluster with over 200 processors. SBC, the Uppsala Genome Center and other institutions also in other disciplines make use of PDC's storage capabilities.

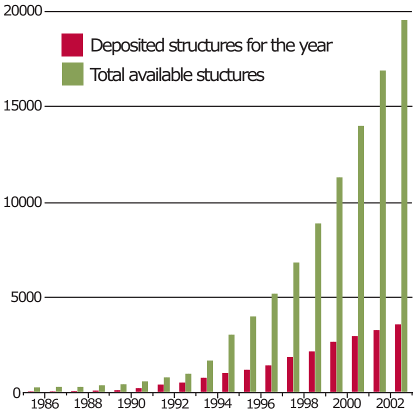
Biology and Computing

One major interest is to develop machine learning and evolutionary computation techniques to understand protein folding and structure. A protein starts life inside the cell as a simple linear chain of amino acids, but quickly and efficiently folds into a specific three-dimensional shape, which dictates how it interacts with other molecules inside or outside the cell. This interplay of proteins and other biomolecules is the very essence of life.

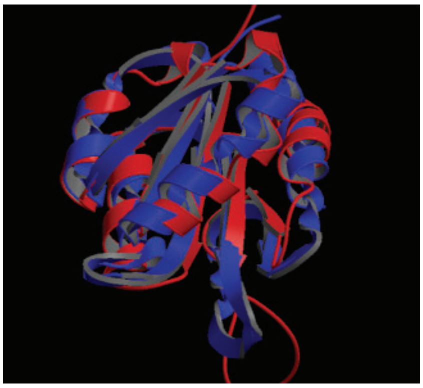
An organism's DNA specifies the exact sequence of amino acids for every protein. In turn, the sequence of amino acids specifies the three-dimensional shape of the folded protein. The folding process is difficult to observe experimentally and currently impossible to simulate accurately a computer. Therefore our limited understanding of protein folding has not yet resulted in computational prediction methods that can accurately produce a three-dimensional structure given a sequence as input.

The number of available structures has increased rapidly during recent years (see figure below). This rapid expansion has made it possible to develop benchmarking procedures to measure the performance of structure prediction methods. This is offered by the LiveBench Project, which is a program that provides evaluations of a large number of structure prediction servers from the point of view of the user. LiveBench offers a fast evaluation cycle that can help locate possible problems and tune methods for best performance.

Another benchmarking project is the Critical Assessment of Protein Structure Prediction (CASP), a biannual exercise where both automatic and manual prediction methods are compared. Recent progress has been achieved by so-called consensus methods, that are based on combining several individual methods. The most important feature of these methods is to discriminate between correct and false predictions. At the most recent CASP meeting in December 2002, the automatic methods were basically as good as the human judgments for the first time. Two of the three best consensus prediction methods were based on the Pcons algorithm developed by Arne Elofsson at SBC.



The number of structures in the Protein Data Bank from 1985 to the present.



Fold Prediction

Sequence information alone is often not sufficient to predict protein folds, since proteins with negligible sequence similarity can adopt the same fold. Fold prediction procedures have therefore attracted much attention recently. More than 70% of the newly determined protein structures today are found to correspond to a known fold. Having powerful methods for predicting the protein fold from the amino acid sequence would consequently be of great help. As function often is conserved within a fold, it also provides help for the prediction of function.

Today the complete DNA sequences of more than 150 genomes, ranging from microbes to humans, are known. In order to exploit this potential, genome sequence information must be decoded in terms of the molecular and cellular functions of the gene products. For many applications, in particular the development of new drugs, information on function must be complemented by knowledge of protein tertiary structure. Functional and structural properties are best detected by detecting homology between proteins of known structure and function. Arne Elofsson's group at SBC is developing such methods. Their latest methods (Pcons, Pmodeller and Pmembr) perform very well on detecting distantly related globular and membrane proteins respectively. A visual indication of this achievement is illustrated in the figure on page 1.

The Genome

A genome is the totality of genetic material (such as chromosomes, etc.) of an organism. Thus, the genome is not the same as 'all genes' but rather 'sequence of all DNA' wherein all genes can be found. The genome of an organism can be seen as the genetic blueprint for that organism. Knowing the complete DNA sequence of the genome is the first step in the complete mapping of the constituents, structures, and processes of the organism. It is a necessary (but not sufficient) requirement for understanding an organism. Biological processes must be described in a formalized way in order to build computable databases and this implies also having time as a fundamental coordinate.

SBC Services

SBC also operates a Web service to make its methods available for other researchers. Some of these services are hosted by PDC but are developed by SBC scientists and staff. The service include fold recognition, transmembrane predictions, and protein localization with several methods, programs, and databases made freely available.

SBC develops, in close collaboration with researchers at the Center for Biological Sequence Analysis at the Danish Technical University, new bioinformatics methods for the prediction of subcellular protein localization and for the identification of integral membrane proteins and prediction of their topology and three-dimensional structure. From a functional genomics perspective, the prediction of the subcellular localization of novel proteins can provide vital clues to their function. SBC's recently developed predictor TargetP is the first well-performing "integrated" prediction method that can

readme

Updated Kerberos software

The Kerberos software on the PDC hosts has been upgraded to Kerberos v5 compatible versions (Heimdal). In most cases this change is transparent to users, but there is one exception. If you need to renew your Kerberos tickets while logged on to a PDC system and need a long-lasting ticket, the command line is slightly changed:

Old syntax: `kauth -l -l`
New syntax: `kauth -l 1month`

Another slight change is that the default time unit used when specifying ticket life time is now seconds and not minutes.

Applying for class accounts now easier

PDC can provide access to the latest high-performance computing technology to the students in your class. Make it easier for them to understand the challenge in and opportunity of current technology by allowing them access to the same up-to-date high-performance computers that are available for Swedish scientists.

PDC recently streamlined the process for applying for class accounts. Not only can you, as course leader, enable all your students' accounts at one time, but you can also make special requests for such things as disk space, processing priority, etc. Go to the page for applying for accounts, http://www.pdc.kth.se/support/accounts.html, and follow the link for "class accounts" to apply.

To be part of this initiative your course should follow good academic standards and you need to justify that access to a national computer resource would be beneficial to your students. Examples of such is need for massively parallel computers, specific hardware architectures of interest to your students, specialized software not available locally to your students, or need for some other specialized resource not available to you locally.

Decisions on applications for course accounts are based on the availability of relevant resources, and the applications are ranked based on the justification supplied.

Future access to vector computers

The last available vector computer resource in Swedish academic computing is finally reaching the end of its life. If you are one of the remaining users relying on this service (selma.pdc.kth.se), we encourage you to contact pdc-staff@pdc.kth.se and discuss how to provide for your future computing needs. The final date when the system will be taken out of service is not decided yet, but it might be sooner than you would like.

Computing time on SGI system boye

Due to the continued efforts in providing our users with modern computer resources, the future of the SGI system boye.pdc.kth.se is under investigation. If you are relying on this particular system for your work, please contact pdc-staff@pdc.kth.se so that we can assure you uninterrupted access to appropriate resources for your work.

Major compiler update

On the IBM SP Nighthawk system compilers has been upgraded to XLF 8.1 and VAC 6.0. This upgrade provides OpenMP 2.0 support for Fortran and OpenMP 1.0 support for C and C++. More documentation is available at the PDC software Web pages.

discriminate between multiple subcellular locations. TargetP is continually developed and extended.

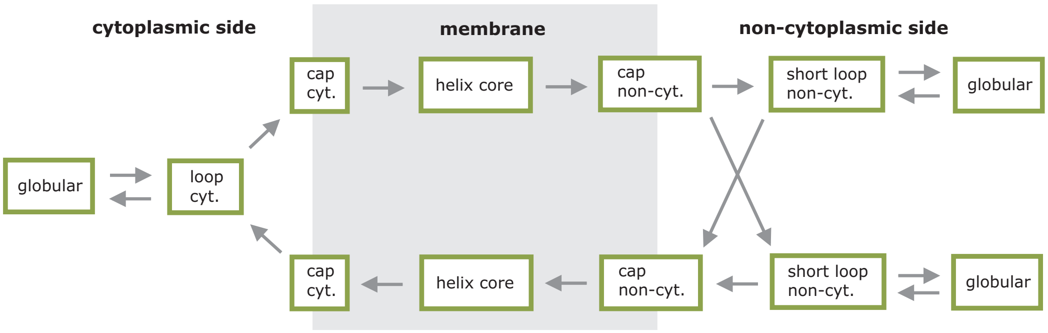
The identification of membrane proteins and prediction of their topology is an equally important first step in largescale functional genomics projects, in particular as membrane proteins are considered among the most important future drug targets. SBC has recently developed a well-performing hidden Markov model (TMHMM) to identify membrane proteins and to predict their topology (as is indicated in the figure below). TMHMM will be further improved by inclusion of multiple alignment information, among other things.

The most successful methods are provided as Web-based servers, which can be found on the SBC Web site http://www.sbc.su.se/service.

Further Challenges

The major challenge for bioinformatics is to synthesize an overall picture of the fundamental life processes out of the mass of experimental data that is being produced by biologists.

Metabolic and signaling networks must be understood in terms of their function in the organism and in relation to the data we already have. This requires combining information from a large number of sources: classical biochemistry, genomics, functional genomics (e.g., microarray experiments), network analysis, life process descriptions, and simulation. A theory of the cell must combine the descriptions of the structures in it (genome, proteome, subcellular structures, etc.) with a theoretical and computational description of the dynamics of the life processes.



The basic HMM architecture of the TMHMM membrane protein topology predictor.

Uppsala Genome Center: A Service Center for Genotyping and Sequencing

Genetic studies, experimental as well as theoretical, are often performed with an expectation to find genes, which cause or contribute to genetic disorders. Analysis of the genome are also performed in evolutionary studies searching for variation in the genome.

The need for more automation and time-effective techniques in experimental studies have led to the development of highly automated capillary electrophoresis instruments. This type of instrument provides opportunities both to commercial companies and academic laboratories to examine DNA with several orders of magnitude higher throughput than earlier instruments. Mapping and sequencing projects can be performed in cost-effective and timesaving ways. A genomic study generates and accumulates large amounts of data. The data has to be easily accessible for further analysis over long periods of time. Thus, genetic studies demands significant amounts of archival storage.

The Uppsala Genome Center is a service facility providing services for genotyping and sequencing projects mainly to the academic community. The center is located at the Rudbeck Laboratory at Uppsala University. The center was established in 1998 as a resource facility for running large genetic mapping projects mainly within the National Genome Program financed by the Foundation for Strategic Research, SSF.

The present throughput of the Uppsala Genome Center is more than 2 millions genotypes per year. The center has an automated procedure for amplification of microsatellite loci and pooling of PCR (Polymerase Chain Reaction) products. All analysis are performed by a capillary electrophoresis instrument, ABI PRISM® 3700 Analyzer.

The Uppsala Genome Center has been involved in a large number of genotyping projects including mapping of complex disorders in humans, and laboratory and domestic animals. The aim of these projects was to map loci for inflammatory diseases in man, rat, and mouse. Projects to map QTL (Quantitative Trait Loci) in domestic animals such as pigs and chickens, were also included.

In 1999 the Uppsala Genome Center established a collaboration with PDC for long time storage of data. This is a convenient way of data storage eliminating the problems with restrictions in local memory capacity. Generated data is automatically transferred at night to PDC via an in-house server. Normal workflow generates 1 GB of data to be stored per week. Via Internet the data stored at PDC is accessible for the staff at the Center but not directly for its customers. Data is generally delivered to customers on a CD or transferred via Internet.

The Center has not been using the capacity of processing data at PDC. In the future this might be an option for some of the statistical analyses performed with generated genetic data.

The Uppsala Genome Center is open to all academic and commercial users. Visit the web site for further information: http://www.genomecenter.uu.se

Education and Training

The need of bioinformaticians in both academia and industry is much greater than the available number of researchers. Due to the world-wide lack of researchers in bioinformatics, growing this discipline is a very challenging task that must address educational and training issues head on. The formation of centers such as SBC and the Uppsala Genome Center are key elements in addressing this issue as are collaborations between academia and industry.

Wordlist

Genotype: i) the genetic composition of an individual; ii) the types of a gene or a DNA sequence found at a locus in an individual.

Sequencing: determining the order of the nucleotides ("building blocks") in the genome.

Capillary electrophoresis: a technique to separate pieces of DNA in an electric field. Fluorescent labeled DNA is transferred inside a thin capillary and detected by a laser and registered by a CCD camera at the capillary end. Separation is according to the length; shortest DNA copy is detected first.

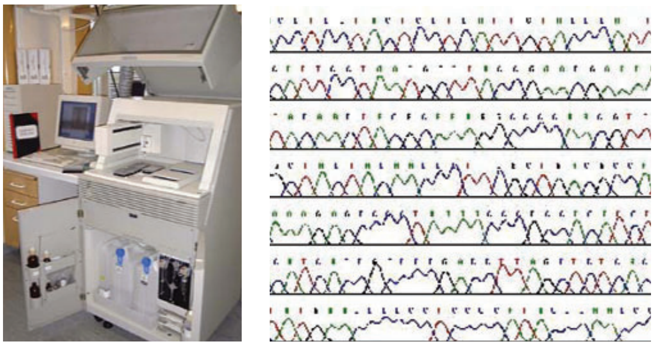
Locus: a unique chromosomal location defining the position of an individual gene or DNA sequence.

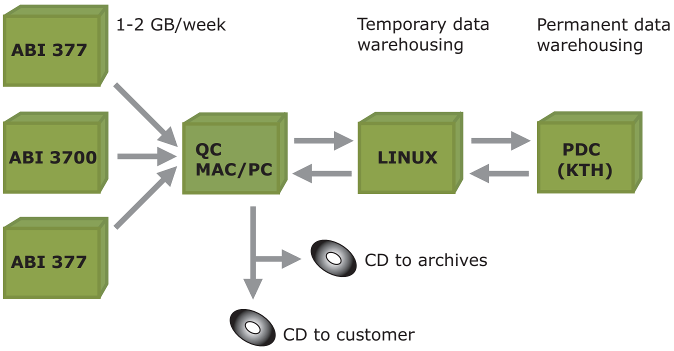
Amplification/PCR: a method to generate millions of copies of short pieces of DNA. Generates amounts of DNA possible to analyze.

PCR-product: a DNA sample after amplification/PCR.

Microsatellite: a repetitive sequence in sets of two, three or four nucleotides. Useful as genetic markers in mapping of genetic diseases.

Complex disorder: a disorder including both genetic (more than one gene) and environmental factors.





The capillary electrophoresis instrument (left) registers spectrums of fluorescent labelled DNA (middle), which are analyzed at the laboratory and thereafter transmitted to PDC for permanent warehousing.

Calendar

At PDC, KTH, Stockholm, Sweden

- June 15-18, 2003: A Workshop in Geometric and Global Properties in PDE with applications
http://www.math.kth.se

- August 18-29, 2003: PDC/NGSSC Summer School: "Introduction to High-Performance Computing"
http://www.pdc.kth.se/training/2003/SummerSchool

- August 18-26, 2003: NGSSC Grid Computing Course: "Grid Computing"
http://www.pdc.kth.se/training/2003/GridComputing

At NSC, Linköping, Sweden

- October 2003: 4th Workshop on Linux Clusters for Super Computing (LCSC); http://www.nsc.liu.se/lcsc

Elsewhere

- June 14-19, 2003, Albufeira (Algarve), Portugal: Advanced Environments and Tools for High Performance Computing, EuroConference on Problem Solving Environments and the Information Society
http://www.esf.org/euresco/03/pc03139

- June 22-24, 2003, Seattle, Washington: The Twelfth IEEE International Symposium on High-Performance Distributed Computing; http://www-csag.ucsd.edu/HPDC-12

- June 23-25, 2003, Seattle, Washington: Global Grid Forum - GGF8 (2003); http://www.gridforum.org, http://www-csag.ucsd.edu/HPDC-12/ggf8.html

- June 23-26, 2003, San Jose, California: ClusterWorld Conference & Expo
http://www.linuxclustersinstitute.org/Linux-HPC-Revolution

- June 23-26, 2003, Monte Carlo Resort, Las Vegas, Nevada: The 2003 International Multiconference in Computer Science and Computer Engineering
http://www.ashland.edu/~iajwa/conferences

- June 24-27, 2003, Heidelberg: 18th International Supercomputer Conference ISC2003
http://www.isc2003.org

- July 7-11, 2003, Portland, OR: O'Reilly Open Source Convention; http://conferences.oreilly.com

- July 13-25, 2003, Vico Equense, Italy: International Summer School on Grid Computing 2003
http://www.dma.unina.it/~murli/SummerSchool

- July 27-31, 2003, San Diego, CA: SIGGRAPH 2003
http://www.siggraph.org/s2003

- August 24-27, Reykjavik, Iceland: 21st Nordunet Networking Conference; http://www.nordunet2003.is

- August 26-29, 2003, Klagenfurt, Austria: Euro-Par 2003: International Conference on Parallel and Distributed Computing; http://europar-itec.uni-klu.ac.at

- August 27-29, 2003, Espoo, Finland: CAVE-Programming Workshop; http://eve.hut.fi/cavews2003

- September 2-5, 2003, Technical University Dresden, Germany: ParCo2003, Parallel Computing 2003 Conference; http://www.tu-dresden.de/zhr/ParCo2003

- September 3-4, 2003, Birmingham, United Kingdom: LinuxWorld UK; http://www.linuxworld2003.co.uk

- Sep 29-Oct 2, 2003, Venice, Italy: EuroPVM/MPI 2003: The European PVM/MPI Users' Group Conference
http://www.dsi.unive.it/pvmmpi03

ROYAL INSTITUTE
OF TECHNOLOGY

PDC
Newsletter
No 1 - 2003

New Linux Cluster at PDC.

New Linux Cluster at PDC

PDC is installing a new computer system, a HP Intel Itanium2 cluster composed of 90 dual-processor nodes for a total of 180 processors. The 900 MHz 64-bit processors provide 7.2 GFlop/s per node for a total of 648 GFlop/s. There are 6 GBytes memory per node for a total of 540 GBytes. A Myricom switched communication fabric will provide the users with a high-performance internal network.

This new cluster follows the tradition at PDC to provide systems with large memories and high memory bandwidth. The system will run the Linux operating system; and PDC will install the utilities and applications familiar to PDC users, making this an easy-to-use environment for scientific research.

The installation of the system has just begun, and deployment is planned for July. The new computer resource will gradually replace the Power2SC and PowerPC part of the IBM SP. Allocations granted on the present resources (IBM SP) will be transferred to the new resource as the improved system becomes available for production use.

Introduction to High-Performance Computing

The traditional PDC/NGSSC Summer School takes place August 18-29 at KTH.

This is an intensive two-week summer school course, for a limited number of participants. A number of topics will be covered in overview lectures, in-depth technical lectures, and hands-on computer lab sessions. For more information, visit the Web site:
http://www.pdc.kth.se/training/2003/SummerSchool