

Organic Meets Inorganic

Linux Bridges the Gap

Linux captures the hearts of biology researchers.

Head of the National Centre for Biological Sciences (NCBS, www.ncbs.res.in)? It is located in India's silicon city, Bangalore, and is a part of the Tata Institute

of Fundamental Research (TIFR). At NCBS, researchers use "...experimental and computational approaches to the study of molecules, cells and organisms." Their aim is simple: to understand biology at each of these

TUX HERO(INE)

Dr R Sowdhamini has been researching in the area of protein sequence analysis and structure prediction since 1992, ever since her post-doctoral fellowship in the UK. When she moved to the NCBS to take an independent position in 1998, she wanted the laboratory to apply the techniques they were using on a large number of sequences—like the whole genome databases—and develop user-friendly computer algorithms and databases.

“We had a sensitive understanding of the bottlenecks and problems in protein structure prediction as users, and we have tried to consider all these points in developing our Web servers and constructing our databases.

Application to genome databases is a huge escalation (computationally demanding) and hence, as part of the learning curve, we began to implement and employ a cluster of computers,” she explains.

In 2001, when the equipment grant was realised, important decisions had to be made about the type of computers, along with specifications in terms of memory, motherboard, space requirements and



Dr R Sowdhamini, with the lab's computer clusters in the background

switch configuration. At this time, when P4 machines were just being released, the cost of Athlon machines was quite high and much caution was required not to purchase too many nodes to build one cluster, according to Dr Sowdhamini. “It was a great experience -- purchasing computers, installing clusters; and in the cluster management, to oversee the systems administration required to look after multiple users. We also learnt to manage sub-derived databases that required the calculation and maintenance of a large number of files by means of dynamic Web page creation.”

Even if the lab had computer staff to assist the researchers in maintaining the clusters, it still required substantial monitoring. She says the most active users of the clusters are those responsible for detecting and reporting problems. “Most of my students immediately showed interest to participate in this valuable learning experience. Apart from the purchase of clusters, systems administration has always been an enjoyable role for me—it was important to thoughtfully administer the rules owing to the presence of multiple users,” she quips.

levels to advance an integrated view of life processes.

Naturally, the centre houses all the necessary facilities that research scientists may need, which certainly includes computing power. In fact, their needs can be compared to the requirements of large-scale enterprises. So, what kind of hardware and software do they use to carry on with their work?

Dr R Sowdhamini's Lab No 25 at NCBS, part of the Computational Biology group, uses computation to analyse genome sequences, study protein structural similarities, and works in developing tools to aid such analyses. Likewise, it did not come as a surprise that this lab houses four clusters. The facilities include a total of 64 rack-mountable machines of different capacities within the four clusters, consisting of close to 200 processors. The clusters include Athlon and Opteron machines—some of which are dual processors, and several are nodes with four processors each. The dual processors are connected by a gigabit switch, while the clusters consisting of four processors each, are connected by an Infiniband switch.

Where is Linux?

Impressive, but where's the Linux angle, you ask? Well, all the four are Linux clusters. That's not all. Dr Sowdhamini

clarifies, “We also provide seven databases on protein sequences and structures, and nine Web servers on the public domain developed using open source technologies (Linux, Apache Server, MySQL, Perl, Python, Ruby and Java) and running on a CentOS server.”

These clusters cater to the computational needs of the two groups for running computer-intensive applications. More than 25 computer applications are installed on the clusters, which includes some in-house programs and several public domain software. Some of these computer applications involve sensitive algorithms that are geared to query connections between protein sequences and structural databases. Just to give you an idea, molecular dynamics of a small protein, for a modest conformational virtual excursion, may take more than 12 hours of parallel computing involving as many as 40 processors.

Computer applications to establish relationships between protein sequences and structures using sensitive sequence searches and molecular dynamics simulations are I/O and CPU-intensive. In addition, these applications can typically give rise to huge output files and hence are space-demanding! “Due to these reasons and also since there are about 25 users between the two laboratories, we have dedicated two NAS storage solutions for file

saving and management,” Dr Sowdhamini explains. “We have a dedicated Linux dual-processor Web server for our lab databases and Web servers... there are more than 3 million visits to our public domain resources from different IP addresses around the world.”

However, all this is available only now. In 2001, the lab only had one Linux cluster consisting of 11 dual-processor Athlon machines after starting from scratch. But why Linux, anyway?

Dr Sowdhamini quips, “Linux computers were our obvious choice since they outweigh others like Silicon Graphics and Sun machines, in terms of their cost-effectiveness, reliability and scalability. Other options, like branded machines (IBM and HP products), were phenomenally more expensive.”

So was cost the only factor for going with Linux? Says Dr Sowdhamini, “Several applications that are required by us are highly suited for installation in Linux machines. We also found the system software, OS configuration and cluster management software highly suited for patches and upgrades in Linux machines.”

Challenges with Linux

So was it all a smooth ride or were there pitfalls along the way? The professor quips, “In the beginning, several issues with the Linux cluster were new, rather unanticipated and some were quite challenging. The maintenance of a cluster, that requires strong cooling and hence more power, needed close attention and was almost a full-time job for one person. Our laboratories did benefit by the presence of other clusters, like those of Dr Upinder Bhalla’s lab, in NCBS, to build the facility. However, it was quite novel to deal with 22 processors, 25 users and 25 different software in our lab cluster when we did not have a role model with exactly that number of users and software.”

To make matters even more complicated, she says that each of these software ran through a different backbone script. Some applications ran as batch jobs while some others were truly parallel. “In our earlier clusters, the cluster management software was often not robust enough to handle multiple jobs run by multiple users. This meant that while we were eager to use the full resource power of the Athlon machines, we did not want to overheat the machines. Despite a 2U architecture and eight fans within a node, the machines used to get heated rapidly.”

She adds, “With time, the purchase and use of additional computer clusters meant that not only did we have to move with the technology but also go through different vendors. Thankfully, the prices of computers and storage disks are being slashed, as well. We are pleased that not all our clusters were built in one go.”

Regarding expectations from the implementation, Dr Sowdhamini says that starting from 11 nodes to 64 nodes, the expectations were to exhaustively search the

THE LEARNING EXPERIENCE


Dr R Sowdhamini says: “The developers of the clusters also felt and expressed their learning experience. In the initial phase, the most important objective for us, as a team, was to keep the cluster with all its nodes living and going when the system went full-steam. In the next few clusters, we could not only enjoy the speed of execution, but we were guaranteed of the basic stability of the systems to intensive CPU usage. The latest generation clusters are much more advanced with key features such as improved air circulation, redundant power supply and an automatic power-off when the systems get overheated! Our infrastructure and architecture teams also learnt to improve the cooling, cabling and dustless airflow within the rooms to deliver a better environment for the clusters. Now, we can proudly claim a state-of-the-art cluster room and machines.”

entire sequence database. At one point, this meant three months of continuous use of the clusters to generate the desired results. “But, our cluster usage was productive in several projects as evidenced by more than 20 peer-reviewed publications that have appeared in international journals (the list can be found at caps.ncbs.res.in/pubs/cluster_pub.html). Likewise, there are more than 10 peer-reviewed publications from our laboratory that report the availability of databases or Web servers (list at caps.ncbs.res.in/pubs/databases_servers_pub.html),” the professor clarifies.

The final cut

This sort of an implementation is a waste if people aren’t aware of it—worse still if the very management of the centre is among those ignorant. So was the management aware of the achievement? “Our management is well aware of our successful implementation of our clusters—in our area, this is also evidenced by research publications using the cluster facility,” she asserts. “Indeed, our management is supportive and happy about our implementation, and soon we will be moving to a bigger cluster room within the centre that will be four times the current capacity!”

Of course, all this doesn’t come cheap. The lab needs to obtain more financial support from internal and external grants, and purchase more clusters, she adds. “We need to improve the explicit link between the laboratories by a fibre optics cable—this should be increasingly easy for maintenance, with time.”

Good to see Tux being part of what Dr Sowdhamini feels is a state-of-the-art cluster that the centre is proud of. Amen! 

By: Atanu Datta, LFY bureau