Parallel Numerics '05, 217-221 M. Vajteršic, R. Trobec, P. Zinterhof, A. Uhl (Eds.)
Chapter 7: Systems and Simulation ISBN 961-6303-67-8

Bioinformatics Application Oriented IT Deployment Model

Nikola Pavković, Karolj Skala*, Valentin Vidić, Zorislav Šojat

Rudjer Bošković Institute,
Center for Informatics and Computing,
Bijenička 54, 10000 Zagreb, Croatia

Modern scientific research methods increasingly depend on some kind of supercomputing infrastructure. As the new distributed programming paradigm found its place among academic and scientific circles, high-performance parallel computing is becoming a very important scientific tool. On the other hand, we must consider some limitations that we might face in the near future. We would like to show the influence of Moore's law on the future of scientific research which depends on high-performance computing infrastructure. Through the presentation we will show our vision of a new model of exploiting and sharing the computing resources. The presentation will also cover a short technical description of what is going on in our research lab regarding this issue.

1 Introduction

e-Science is used to represent the increasing scientific global collaboration, of people and shared resources, that will be needed to solve the new problems of science and engineering. These e-Science problems range from the simulation of whole engineering or biological systems, to research in bioinformatics, proteomics and pharmacogenetics. The information technology infrastructure that will make such collaboration possible in a secure and transparent manner is referred to as the Grid. Both e-Science and the Grid have fascinant technical (infrastructure) as well as scientific (applications) aspects.

The two key technological drivers of the IT revolution are Moore's Law [1] (the exponential increase in computing power and solid state memory) and the very intensive increase in optical network communication bandwidth.

*Corresponding author. E-mail: skala@irb.hr

Scientists are now attempting calculations requiring orders of magnitude more computing and communication than was possible only a few years ago. They are satisfied by the widespread deployment of cheap clusters of computers at university or institute campus environment at research group level. Moreover todays experiments generate several orders of magnitude more data that has been collected in the whole of human history. The genome sequence data is increasing at a rate of four times each year and that the associated computer power required to analyse this data will only increase at a rate of two times per year, still significantly faster than Moore's Law. In the next decade we will see new experimental facilities coming on-line, which will generate data sets ranging in size from hundreds of terabytes to tens of petabytes per year. There are many examples that illustrate the spectacular growth forecast for nanosecond and nanometer scale scientific data generation. We have many rapidly growing databases in the field of bioinformatics. The data in these cases, unlike in some other scientific disciplines, is complex mix of numeric, textual and image data.

In order to handle this 'explosion' in storage and computing demands, we have to consider the sustainability of today's high performance computing deployment models.

2 "Bored" Computers

It is well known that evolution of software in some way dictates the rate new hardware is being bought. Every few years a new generation of PCs with increasingly powerful processors are being acquired within scientific institutes and universities, and are put under users' desks to serve as personal workstations. We made a short research to see how much the expensive and powerful hardware is really being exploited within the organisation that owns it.

The result was very interesting. Typical processor usage graph can be seen on figure 1.

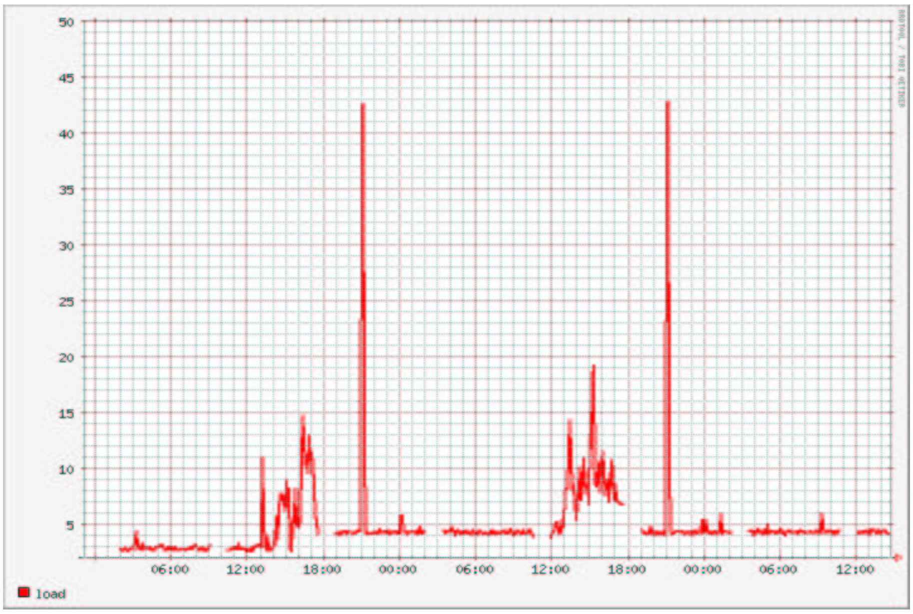

Figure 1: Processor usage graph

It is clearly visible that the average processor utilization is far less than 10 percent during working hours, and less than 5 percent during non-working hours. We might say that our computers are 'bored'. The reason for this boredom is primarily the lack of integration model that will allow to exploit all the spare cpu cycles on the 'bored' workstations.

3 Existing IT infrastructure deployment model

Nowadays, the IT infrastructure deployment and exploitaion model can be described with one word. That is "dedication". All the systems that we build, we dedicate them to a specific purpouse i.e. mail server, web server, personal workstation etc. However, this model has some advantages, but noone asks if it is optimal i.e. sustainable in the new distributed computing era. It is obvious that in a short time from now, new highly demanding applications will emerge, and the underlaying hardware infrastructure is going to be short, or at least, the existing deployment model is going to be extremely expensive.

4 The new model

At the Rudjer Boskovic Institute, we are working on implementation of a inovative concept of 'harvesting' idle processor cycles from pretty powerful workstations located within the Institute. The prime target application that is about to run on this system is BLAST (Basic Linear Alignment Search Tool) [2], which is mostly used by biologists to find similarities among genome and protein sequences. The very important issue here, is that BLAST has a great potential of becoming one of the most resource-consuming application in the near future. Since BLAST itself is actually a database lookup algorithm, it can pretty easily be parallelized, moreover the concurent task-chunks can run completely independant from each other. These characteristics are the key for integration of idle workstations to form a cheap but extremely powerful infrastructure for large BLAST database lookups. Because of distracting the massive BLAST jobs from expensive, dedicated supercomputing facilities, this model of IT infrastructure integration will cause significant savings in future high-performance computing investments. Let's consider the cost of a supercomputing facility (i.e. high-performance cluster) needed for a number biologists using BLAST tool in their research. The accumulate price of a 'dedicated' system (for a 3-year runtime):

$ 600.000,00	Computing hardware (512 nodes of 2 GHz)
$ 20.000,00	Network equipment
$ 10.000,00	Central node (large SCSI disks, lot of RAM, 2 processors)
$ 30.000,00	Storage Area Network
$??.???,??	Server room
$ 48.000,00	Maintainence
$ 700,000.00	or more

Let's consider the expenses of a system with similar performance, built on the new model:

$ 0,00	Computing hardware
$ 10.000,00	Network equipment
$ 10.000,00	Central node
$ 30.000,00	Storage Area Network
$ 0,00	Housing
$ 48.000,00	Maintainence
$ 100,000.00	or less

5 Conclusions

With implementation of the ’new model’ we expect to achieve significant savings while also providing our scientists powerful computing infrastructure at the same time. This model however does not solve all the problems with lack of resources, but significantly improves the price/performance ratio of a company’s IT infrastructure. Other details on http://dcc.irb.hr

References

[1] http://www.intel.com/research/silicon/mooreslaw.htm

[2] NCBI BLAST - http://www.ncbi.nlm.nih.gov/BLAST/

[3] http://dcc.irb.hr