

## Grid Computation – The Fastest Supercomputer in the World

By Chao-Hsu Yao

### Abstract

A supercomputer is a computer designed to do large-scale and complicated computation by using many CPUs (Central Processing Units) simultaneously to perform parallel processing. Grid computation comes from the idea of super computation; it uses the internet to connect computers worldwide, to create a virtual supercomputer. Grid computation saves money and space, while a supercomputer is costly and wastes space. However, security is a concern because sharing resources always increases the risk of a computer virus or worm.

### Introduction

#### ➤ *History*



The ENIAC (Electronic Numerical Integrator and Computer), the first large-scale general-purpose electronic computer, in 1946  
<http://www.library.upenn.edu/exhibits/rbm/mauchly/jwmintro.html>  
Department of Special Collections,  
Van Pelt Library, University of Pennsylvania

It has been almost 80 years since the first computer was invented. It was able to perform digital calculation and operated with many large sized vacuum tubes. This old computer wasted a lot of space (normally two rooms) and could only perform simple calculations. With the invention of small transistors to replace large vacuum tubes, and the progress of semiconductor manufacturing technology, the size and processing speed of the computer has greatly improved, as circuits are etched on a chip less than one micron wide. Today a personal digital assistant (aka PDA or palmtop) is the same size as a calculator with almost the same capability as a desktop personal computer, which is lighter, uses less electric power, and is less costly than a large computer.

#### ➤ *Parallel Computation*

Engineers have been seeking ways to minimize computer size and speed up processing since the invention of the first computer. Therefore, central processing unit (CPU) design has long been important in Materials Science, Engineering, Electronics Engineering, and Computer Engineering. It has been a challenge to make the CPU smaller and smaller by improving manufacturing technology. One major problem is that the heat generated from a CPU increases as its size decreases, forcing processing to slow down. To solve this size-and-speed issue, parallel processing, which uses multiple CPUs to perform parallel computations and decreases processing



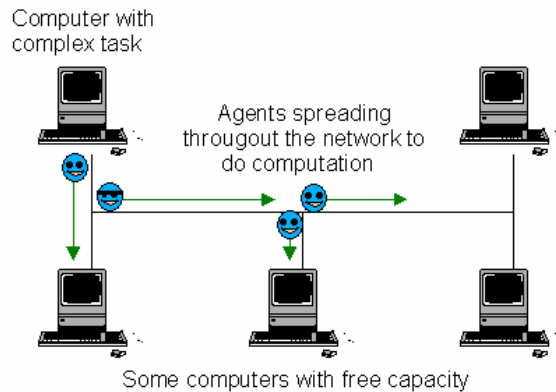
Blue Gene, the fastest supercomputer in the world, can do 135.5 trillion calculations a second  
<http://news.bbc.co.uk/2/hi/technology/4379261.stm>  
BBC News

time, was invented. A supercomputer, normally containing thousands of CPUs, has been used to perform complicated calculations and simulation.

➤ *Grid Computation*

Building a supercomputer facility is a huge and costly undertaking. In the early 1990s, the distributed computing system was invented at the University of California, Irvine, which connected several personal computers to perform parallel computation, instead of building an expensive supercomputer. With the invention of the internet, which connects millions of computers in the world through servers using a special protocol (aka Internet Protocol, or IP,) grid computation was invented based on the design of the distributed computing system, using the internet's powerful communication ability to connect online computers to perform parallel computation.

**Distributed Computing using Mobile Programs**



<http://www.projectory.de/kaariboga/usage/distributedcomputing.html>  
Kaariboga Mobile Agents

➤ *Comparing the Supercomputer with Grid Computation*

A supercomputer has one operation system, which must be able to support parallel computation, while grid computation operates under many different operation systems, which do not need to be parallel computation supported. With grid computation, several internet protocols must be followed for communication between the computers. However, a supercomputer is just ONE computer with multi-CPU, while grid computation uses MULTI-computers connected to each other through internet protocol to simulate a supercomputer's function.

Obviously, a supercomputer with 1000 CPU's runs faster than a thousand connected computers with single CPU, because each CPU running grid computation spends extra time handling internet protocol.

**Theory**

➤ *What kind of processes can be done in parallel?*

To solve a problem using grid computation, the process must be divisible into several subprocesses, and each subprocess should be independent. For example, using an editorial company with 10 employees, if we would like to calculate the average production number per employee per day, we may have 10 CPUs to calculate each employee's average daily production number in one month, separately, at the same time, and use one CPU to calculate the 10 employees' average daily production number. It

only takes two steps to finish the job, while it takes 11 steps using a single CPU. (See Fig. 1)

Single CPU



10 CPU's

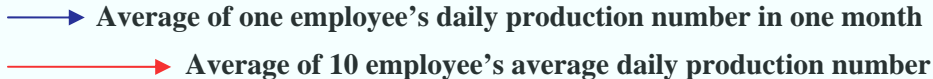
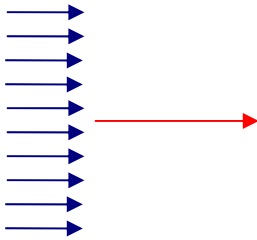


Fig. 1

However, if the process cannot be divided into several subprocesses, the parallel computation won't be performable. If some of the subprocesses are dependent on previous ones, the parallel computation will still be performed, but won't greatly speed up the outcome. For the above example, the red arrow stage (averaging 10 employees' daily production numbers in one month) cannot be started, until all blue stages (averaging each employee's daily production numbers) have been completed.

Vector computation is one of the most popular parallel computations, and has long been used in solving scientific problem. A vector  $V$  containing  $n$  elements is normally expressed as

$$V = \{v_1, v_2, \dots, v_n\}$$

To perform addition for 2 vectors  $X$  and  $Y$ , we will need to do

$$X + Y = \{x_1, x_2, \dots, x_n\} + \{y_1, y_2, \dots, y_n\} = \{x_1+y_1, x_2+y_2, \dots, x_n+y_n\}$$

Doing this sequentially, it takes  $n$  steps to complete the job. However, using an  $n$  CPUs machine to perform parallel computation, it only takes one step to complete the job.

The Bitonic Sorting Algorithm is a very fast sorting method using parallel computation to sort bitonic sequences. A bitonic sequence is a series of numbers, for which

- the sequence is monotonically increasing and is monotonically decreasing; or
- there exists a cyclic shift of indices such that the above is satisfied. (See Fig. 2)

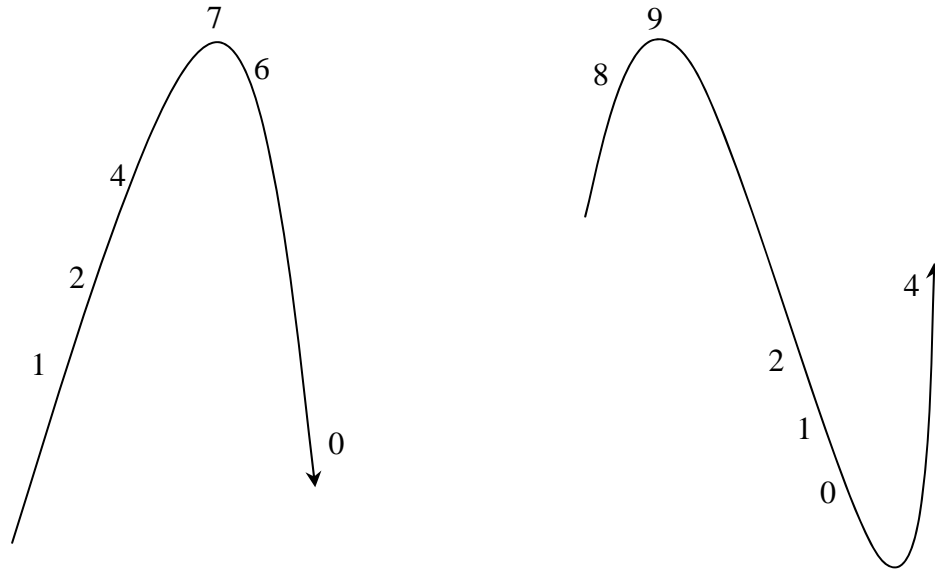


Fig. 2

For example, to sort 8 bitonic numbers using 4 CPU's, in the first step we compare the 1<sup>st</sup> and 5<sup>th</sup>, 2<sup>nd</sup> and 6<sup>th</sup>, 3<sup>rd</sup> and 7<sup>th</sup>, and 4<sup>th</sup> and 8<sup>th</sup> numbers and switch them if they are not in order. In the second step, we compare the 1<sup>st</sup> and 3<sup>rd</sup>, 2<sup>nd</sup> and 4<sup>th</sup>, 5<sup>th</sup> and 7<sup>th</sup>, and 6<sup>th</sup> and 8<sup>th</sup> numbers and switch them if they are not in order. In the 3<sup>rd</sup> step, we compare 1<sup>st</sup> and 2<sup>nd</sup>, 3<sup>rd</sup> and 4<sup>th</sup>, 5<sup>th</sup> and 6<sup>th</sup>, and 7<sup>th</sup> and 8<sup>th</sup> numbers, and switch them if they are not in order. The total parallel process can be shown in Fig. 3. It only takes 3 steps while sequential comparison using a single CPU could takes up to 12 steps to complete the job.

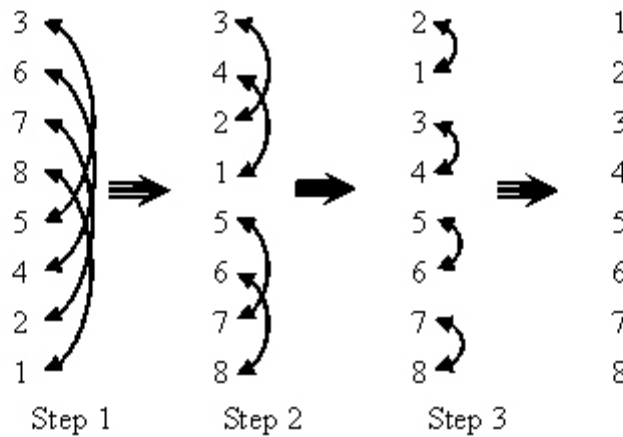


Fig. 3

➤ *How can computer resources be shared on line using internet connections?*

Unlike most companies and academic institutions, which run computers in a Local Area Network (LAN) environment, grid computation runs processes under an internet (global) environment. Therefore, a specific domain must be set up to restrict resource sharing to those with authorized access. This domain is called virtual organization (VO, aka the Grid), which is outside the LAN environment.

People who join a VO will be able to share some (but not all) of their computer's resources (storage drive, memory, or CPU...) with others in the same VO. Normally a web-based interface is designed for users to perform their requests.

When a computer joins the VO, it will start running a subprocess after receiving the request online from another computer. The benefit of grid computation is that when users log in to the internet, they may not use their computers' CPUs all the time. For instance, when they are reading news on a website, it may take them 10 minutes to finish reading an article, while the CPU stays idle. Other users will try their best to utilize this computer's CPU to run the subprocess in parallel while it's idle. Imagine how many computers there are running parallel processing in a VO, even if most of them provide only a few resources. A proper design could speed up the whole process.

## Applications

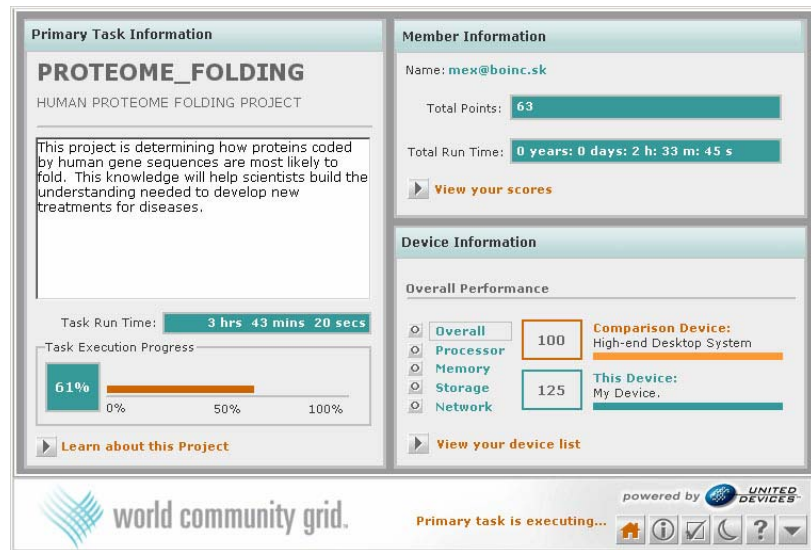
To perform grid computation, the process must be divisible into several subprocesses and run in parallel. The following are some of famous projects that have been designed for grid computation.

➤ *Human Genome Project (HGP)*

The human genome is composed of 24 distinct chromosomes with about 3 billion DNA base pairs organized into 20,000~25,000 genes [1]. To identify these genes and determine the sequences of 3 billion DNA base pairs, running a computer simulation would be expensive and time consuming. Coordinated by the U.S. Department of Energy (DOE) and the National Institute of Health (NIH), the Human Genome Project was completed in 2003, three years ahead of the target goal. The institutes involved in this research are Wellcome Trust, Sanger Centre, and HUGO Gene Nomenclature Committee. The Wellcome Trust Sanger Institute has accomplished almost 1/3 of the total work. The project identified about 20,000~25,000 genes in human DNA, determined the sequences of the 3 billion chemical base pairs that make up human DNA, stored this information in databases, improved tools for data analysis, transferred related technologies to the private sector, and addressed the ethical, legal, and social issues (ELSI) that may arise from the project [2].

➤ *Human Proteome Folding Project (HPF)*

After the Human Genome Project was completed, scientists wanted to understand the function of human proteins, which affect human health, to discover the cure for diseases such as AIDS and cancer. Today, only the function and structure of 30 % of human proteins are known [9]. To identify all human proteins could take



Graphic interface for the world community grid  
<http://www.boinc.sk/gfx/articles/world-community-grid-running.png>

up to 1,000,000 years using the most advanced personal computer to perform the analysis. Therefore, the Human Proteome Folding Project (HPF) was started and ran on two computational grids; World Computing Grid (WCG) [10] and United Devices' grid.org [9], which tried to identify all human proteins' functions and structures in a short time. The institutes that participated in this research include the Institute for Systems Biology, the University of Washington, Seattle, and the IBM Corporation. For more information about HPF, visit Institute for Systems Biology website: [http://www.systemsbiology.org/Scientists\\_and\\_Research/Technology/Data\\_Visualization\\_and\\_Analysis/Human\\_Proteome\\_Folding\\_Project](http://www.systemsbiology.org/Scientists_and_Research/Technology/Data_Visualization_and_Analysis/Human_Proteome_Folding_Project).

➤ *World Community Grid*

The World Community Grid is an experimental project led by IBM, which accepts volunteer members. It is a non-profit organization, which welcomes anyone in the world to donate some computing resources when staying online but doing nothing. It supports all kinds of research that benefits humanity, at no cost. To participate in World Community Grid you can download the software from their website (<http://www.worldcommunitygrid.org/>) and install it. The software is free and secure. The current projects running are Help Defeat Cancer, FightAIDS@Home, and Human Proteome Folding - Phase 2 [5]. You may also submit your project proposal to them by filling out the application form online.

➤ *Computational Chemistry*

Chemical reactions or molecular behavior can be huge and complicated processes. Chemists have been trying to determine molecular structure, simulate molecular behavior, and predict the reaction processes. Computational chemistry has been operational for a long time; however, some chemistry problems, like quantum mechanics,

would take hundreds of years to simulate on a personal computer. Therefore, grid computation plays an important role in computational chemistry, which not only saves equipment costs but also processing time. Computational Chemistry Grid (CCG, <https://www.gridchem.org/>) is one of the most important virtual organizations, which provides all necessary software and resources for computational chemistry.

➤ *Business Computation*

Grid computation is not only used in science, but also in business computation, where all corporate resources can be pooled so they can be processed efficiently in parallel, according to the business demand. Based on this design, enterprise level business-to-business (B2B) collaborations will be the virtual organization, which handles resources management [3].

Oracle has developed the most famous database management system in the world, and most enterprises like its reliable data management ability and powerful data query process. The Oracle 10g, in which g stands for “grid,” has become the first database management system for grid computation. The Oracle 10g runs all database systems in a virtual environment (grid) where all systems are considered a resource pool, using resources efficiently and dynamically for business needs [6].

➤ *SETI@Home Project*

Searching for extraterrestrial intelligence (SETI), is a compelling scientific research topic. SETI@Home, directed by UC Berkeley, utilizes grid computation technology to analyze space-based radio signals collected from a radio telescope, at Arecibo, Puerto Rico. This project uses a new platform, Berkeley Open Infrastructure for Network Computing (BOINC), to support the research. This platform will automatically update without having to download new versions.

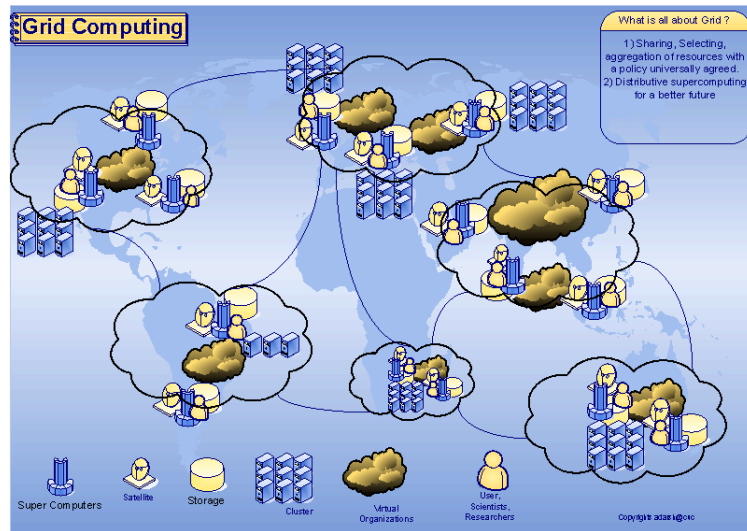
Sun Microsystems Inc. has made a great contribution to the SETI@Home project by providing UC Berkeley with required equipment and software [8]. By downloading a screensaver program bundled with required software, almost a half million personal computers have been connected, through this technology, to perform grid computation for the SETI@Home project.

➤ *Other Applications*

Grid computation can also be used in financial modeling, earthquake simulation, and climate/weather modeling, which are complex processes requiring an intricate infrastructure. A dynamic grid environment, which can perform parallel processing under a collaborative network, must be created to deliver the information [7].

## Grid Computation Software

To perform grid computation for joined computers online through the internet, the software which supports grid computation framework must be installed on each computer inside the VO. The software not only handles information queries, storage management, and processing scheduling, but also does authentication and data encryption to ensure information security. Some famous software that supports grid computation is Java CoG Kit [4], UNICORE (UNiform Interface to Computing REsources) [11], and IBM Batch-on-Grid [12]. Microsoft is also working on a new version of the Windows system that supports grid computation [13].



Grid computing worldwide  
<http://www.adarshpatil.com/newsite/research.htm>  
 Adarsh Grid Computing Research

## Security Problems

### ➤ *Authentication*

Grid computation shares resources online through the internet, so anyone may access shared resources. Therefore, information security has been very important in the grid computation area. The basic idea of controlling access to shared resources is through authentication. The simplest authentication design is to set up a username and password for the user to join a VO. Over time, the authorization framework and architecture design have been popular research topics in grid computation to ensure better information security.

### ➤ *Cryptography*

To prevent unwanted users from stealing information, encryption has been widely used for data transmission. A special cryptography can be designed for each VO to transfer data online. Even if information is stolen by unwanted users, they must have the cryptography scheme or key to decode the data in order to view it.

### ➤ *Hackers, Worms, and Viruses*

A VO can have millions of computers running parallel processes to carry out high performance computation. Therefore, it is likely that some of them were infected by computer viruses and worms, or hacked by intruders. The VO may not be damaged if

## CSA Discovery Guides

<http://www.csa.com/discoveryguides/discoveryguides-main.php>

Released November 2006

only some of the computers were hacked and part of the information was stolen. However, the computer may be infected with an unwanted program (computer virus) that spreads to infect other computers inside the VO, and finally causes all of the information to be lost. It may be a good idea for all users inside a VO to perform a virus scan before receiving information from other users. Even if the VO has a very safe authentication process, however, the virus can still find a security hole to get in, or spread by e-mail and message passing software like ICQ or MSN.

Some computer worms, which may not be designed for stealing information, can also infect the VO, increase the CPU load, and hinder performance.

### Future work

Grid computation is proven to be an expensive supercomputer for doing parallel computation. Though grid computation is still in the development stage, and most projects are still voluntary and experimental, it has developed very quickly and more and more scientists have worked hard to improve its efficiency and security. The most powerful way to use grid computation would be to use all the computers in the world online. There are millions of computers online in the world everyday, some of which may just stay idle, while others are busy running programs 24 hours per day. If we can share all of the computer resources worldwide to perform some complicated computation, we will save a tremendous amount of money and time. However, with more and more computers involved in grid computation, the security problems will become increasingly serious. How to design the largest global virtual environment for grid computation is the most important issue when trying to employ all possible computer resources on earth.

All figures developed in-house at CSA

### References

1. International Human Genome Sequencing Consortium. (2004).
2. Human Genome Project Information, [http://www.ornl.gov/sci/techresources/Human\\_Genome/home.shtml](http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml)
3. UNCW Grid Computing Project, <http://www.gridnexus.org/web/index0.php?session=business>
4. Cog Kit Home Page, <http://www.globus.org/cog/>
5. World Community Grid – Home, <http://www.worldcommunitygrid.org/>
6. Dan Kusnetzky, Carl W. Olofson, Oracle 10g: Putting Grids to Work
7. Gregor von Laszewski, Grid Computing: *Enabling a vision for collaborative research*
8. GRIDtoday: Sun Gives Equipment to SETI's Grid Computing Project, <http://www.gridtoday.com/04/0112/102491.html>
9. The Human Proteome Folding Project, <http://www.grid.org/projects/hpf/>

10. Human Proteome Folding Project Overview,  
[http://www.worldcommunitygrid.org/projects\\_showcase/archives/viewHpfOverview.do](http://www.worldcommunitygrid.org/projects_showcase/archives/viewHpfOverview.do)
11. UNICORE, <http://www.unicore.org/unicore.htm>
12. New IBM Software Brings Autonomic Computing to Grids, <http://www-03.ibm.com/press/us/en/pressrelease/19636.wss>
13. Darryl K. Taft, Microsoft Brings .Net to Grid Computing,  
<http://www.eweek.com/article2/0,4149,39279,00.asp>