

# History of Grid Computing

December 2007

Michel Guillet

Helsinki University Of Technology

Grenoble Institute Of Technology

email: michel.guillet@gmail.com

**Abstract**—This article gives an overview of grid computing. Definition of grid computing is given with its composition and its differences between traditional solutions. Assets of grid computing systems are exposed and its issues are evoked. The article relates the history of these systems and the tools used. It also gives a detailed example of a grid computing project and lists some projects with their respective goals. At the end is a view of future of grid computing.

## I. INTRODUCTION

Computers are great assets in the scientific research. Nowadays, theories demand a lot of calculation and we could not do it without computer. Computers are also very useful to process data.

But, sometimes, the calculation is so big that it could take days or years for regular computers. This is why distributed computing has been invented. Supercomputers, clusters and grid computing are a part of distributed computing. Grid computing is the more recent implementation of distributed computing.

## II. WHAT IS GRID COMPUTING ?

Grid computing is a technology of distributed computing like cluster or supercomputer.

### A. Definition

1) *Theory*: Many definitions about grid computing can be found. The CERN[1] (European Council for the Nuclear Research) consider grid computing as *a system for sharing computer power and data storage capacity over the Internet*. Another definition can be : *it is a technology that enables resource virtualization, on-demand provisioning, and resource sharing between organizations*[2]. One concept stands out between these definitions : grid computing is about setting up different and distant computers together in the aim of adding their resources

2) *Current applications*: Nowadays we can find two type of grid computing, which correspond to two types of grid : local and global grid. A local grid can be associated to a local network such as a network of a small company. A global grid would be associated to a WAN, eg. : Internet.

People do not use all the resources of their computer. A lot of computing power is left unused and the idea of grid computing is to provide this unused power to one global system.

There are many current applications. The most popular one is certainly grid computing for scientific project.

There is a lot of scientific projects on the Internet gathering domain like : mathematics, physics, biology, and chemistry. Other application can be found, such as separate compilation of source code, video rendering, 3D rendering, etc. However, these last applications are most likely to run on a local grid.

### B. Composition

The regular architecture of a grid computing system is composed by :

- one or several servers used to distribute work and collect the results owned by the organization of the project.
- A network. The grid system runs over this network. It allows each component to communicate. It can be a local network or a larger network such as Internet
- Multiple nodes running client application. Usually nodes are regular PCs or workstations.

Among the grid computing projects over Internet, certain projects keep statistics. They can also have account management. It can be used to classify and make a ranking list of the entire participants. Teams of participants can also be formed.

### C. Difference between grid and cluster or supercomputer

Cluster and supercomputer are entirely designed for high performance computing. The software is designed to optimize the use of every node and CPU. The hardware can be designed specifically to this purpose. For instance, supercomputers are computer with thousands of CPUs and are often manufactured in one single copy. A cluster is a set of powerful servers with a very fast network.

On the other hand, nodes of a grid computing system are regular PCs. They are linked with poor network (slow with delay). And they provide unused CPU time to the system. But... there are hundreds of thousands of them. And this makes the strength of a grid computing system. Like for a P2P system, the efficiency improves with the number of participants.

## III. WHY GRID COMPUTING ?

Grid computing appears because of the limitation and the difficulties encountered in the setup and implementation of solutions based on clusters and supercomputers.

### A. A cheap solution

Both of supercomputers and clusters are expensive. These solutions require a specific area to install the computing structure. They also demand a lot of electrical power to compute and to cool the whole system. These are the main costs when those kinds of systems are running.

PCs are cheap and widespread. A cluster needs powerful servers with high-performance network. A supercomputer is a very specific system, with a lot of CPU, fast memory and so on. So, they are very expensive. Clusters exist for this reason, it is cheaper to assemble powerful server than design a whole new supercomputer. With grid computing, you use other PCs around the world. The grid system do not own the PC but PC's owners share their CPU time with the grid system.

The remaining costs, for a grid computing system, are the software development and the system to collect results.

### B. Easy to develop

Developing software for supercomputer is not an easy task. These monsters of computation have a lot of shared processes, memory, etc. Software designers and developers have to be very careful with concurrency problems. For a cluster, application has to be optimized on the network side.

There is no such problem with grid computing. Concurrency issues are not a problem. The network is slow, but we know it and it is not an important part to the efficiency of the application. A lot of tools exist to design grid system. The grid client is running on regular PCs so it is easier to debug than a complex program running over multiple nodes and/or CPUs.

### C. Great Scalability

The design of a supercomputer or a cluster is quite frozen. You cannot easily add nodes or CPUs. In clusters, adding nodes implies to resize the network. Grid computing systems are not designed for a fix number of clients. If you need another client : just connect a new PC to the network, install the client software, run it and that's it ! In the case of a worldwide system, a simple website where people can download the client is needed and a proper server to distribute work and collect results.

Supercomputers can have tens of thousands of CPUs (130 000+ for Blue Gene[3]). Clusters can have thousands of computers, and several CPUs on each computer. The size of grid computing system can reach millions of devices.

## IV. ISSUES OF GRID COMPUTING

Grid computing does not solve all the problems. This solution can not replace traditional solutions such as clusters or supercomputer. Compared to these solutions, grid computing has some weaknesses.

### A. Based on a slow infrastructure

A grid computing system uses a network to allow communication between the different nodes. One factor of efficiency is the speed of this network. The faster it is, the more powerful

the system will be. In the case of clusters, communications between CPUs or node are very fast. But for a grid computing system running over the Internet, it is slow. Though the average speed of connections to the Internet have been improved these past years, it is still very slow compared to local network and even slower compared to a memory bus. The results need to be transmitted quickly with small delay. In grid computing delay is an issue because a node could be geographically everywhere on the planet.

### B. Project specific

Due to a slow communication infrastructure, grid computing system is not efficient for every project. If the computing needs a lot of intermediate results dependent on each other, then we would rather use a cluster or a supercomputer. These systems provide very fast communications between nodes and can even share memory between node and CPU.

On the other hand, there are computing projects that allows a splitting of the computing in independent parts. In this case, there is no need to communicate intermediate results, all the computing is done locally and only the final result is sent. For instance, grid computing has been used to crack cryptographic keys. This project used a brute force method trying every possible key. Grid computing was a good solution : the system gave a set of keys to the nodes to try and the result was sent back. This project managed to crack 64-bit RSA key in less than five years[4].

For this kind of work, grid computing systems are very efficient and can be compared to the most powerful super-computer or cluster.

### C. Security

A lot of grid computing projects are based on voluntary participation. Everyone who has a PC with an Internet connection can download the grid client and start to work on the project like anybody else. This is one of the reasons why grid computing projects are popular.

But like every popular system on the Internet, it is likely to be attacked. It happened already with the SETI@Home project[5]. People can send fake results to the system in order to crash it or to gain credit. In this case, the system is corrupted and so are the results. In clusters or supercomputer, there is no trust issue. These systems are working over a local environment. Every node or CPU is trustworthy. On the Internet, this is much more different. Internet provides anonymity, and the end-user cannot be trusted.

People came with an easy solution to this leak, but it divides the computing power by at least two. One only has to send the work twice (or more) to two random nodes. Then, if both of the results matched it is a valid sent result ; otherwise one of the nodes (or both if you are unlucky !) is malicious or defective. In this way you make sure that every results is valid. Furthermore this solution provides you the ability to identify malicious or defective nodes. If a node sends several time results, which are identified as corrupted result, you can tell that there is something wrong with this node.

## V. HOW DID IT BEGIN ?

### A. Origin of Grid computing concept

The name *Grid Computing* comes from Ian Foster and Carl Kesselman also called *Fathers of the grid*. They use this name as a metaphor. They wanted to use computer power as simple as you can plug an electrical device on the power grid. The word grid also evokes the fact that the computer can be anywhere on the Internet.

The idea was to use PCs' unused CPU time to work on some project. This was called CPU scavenging (also named cycle-scavenging or cycle stealing). This is possible because there are a lot of PC's connected to the Internet and doing nothing. In this way, during the night, lunchtime or work, the CPU time is used to work on a given task. Then the results can be sent to the network.

These *Fathers of the grid* also led the development of the first solution to build grid system, called Globus Toolkit.

### B. Tools

Design and build a grid computing system is not an easy task. There are some solutions to make it easier. The most known distributed computing project is SETI@Home. It was also one of the first volunteer grid projects and its design did not include security measures. So people tried to cheat in order to win more credits. Some also sent falsified works. The University of California, Berkeley design BOINC (Berkeley Open Infrastructure for Network Computing) to solve these security leaks. At first BOINC was designed for SETI@Home but, now, there is a lot of grid computing projects using this platform.

BOINC is a free software. Commercial softwares also exist such as :

- Xgrid designed and developed by Apple and running on the Mac OS X platform.
- Sun Grid Engine (ex-CODINE) designed and developed by Sun Microsystems and running on multiple platforms.
- Globus Toolkit designed and developed by *Fathers of the grid* and running on multiple platforms.

### C. SETI@Home

This was one of the first popular grid computing project. This project was released to public in May 1999[6]. SETI is an acronym for *Search for Extra-Terrestrial Intelligence*. One of the main goals of this project is to search for someone sending us some radio signals from up there. The radio data come from Arecibo radio telescope. Once sent to the SETI@Home facility, it's split into small chunks and sent to every PC connected to the Internet and running the SETI@Home client. Then, the client processes the data searching for radio signal that cannot be assimilated to the deep space's noise.

As you know, this aim is still not reached today. Even if some data opened to hope as in September 2004, nothing serious has been found yet.

In fact, the main goals of this project was also to scientifically prove that grid computing over the Internet is possible.

This goal has been reached. The BOINC platform is now used by a lot of different projects. There were 5.2 millions of computers running SETI@Home and it is the most important nowadays. The project has accumulated 2 millions years of computing time and it has a record in the Guinness World Records (Largest computation in history).

The SETI@Home project has recorded his performance peak in October 2007 ; it was able to compute about 275 Teraflops. For comparison, the Blue Gene supercalculator (designed by IBM) has 478 Teraflops computing peaks.

### D. Some other projects

There is a lot of other computing projects over the web. Most of them are beneficial to the humanity (fight against cancer, finding IT, etc.) and that may be the reason of their popularity.

Here is a short and uncompleted list of famous projects you can participate on the Internet:

- FightAIDS@Home : This project is based on the World Community Grid (IBM) platform. The project is about testing how well certain molecules bind to the HIV. It runs on Windows, Linux, Mac OS X, and SGI. This project was released to public in 2003.[7]
- Folding@home : This project is based on the World Community Grid (IBM) platform. It computes how do molecules interact with each other to form functional molecules. This is used to understand the development of diseases such as cancer, BSE, Alzheimer's disease. Folding@Home became in September 2007 the fastest computing system ever with a 1224 teraflops peak. It is a very popular project and it is multiplatform : Windows, Linux, MacOS, PS3. This project was released to public in October 2000.[8]
- distributed.net : Two projects for this community : one mathematical and one cryptographic. The first consists into finding numbers with the *Golomb Ruler* characteristic. This problem has no solution but to calculate every possibility. The other project was about breaking RSA Key's. They succeeded to do it for key's length of 56(250 days to find it), 64 bits (1757 days to find it). 72 bits key cracking was in progress but RSA decided to shutdown its support to the project[4].
- Genome Comparison : This project is based on the World Community Grid (IBM) platform. It was released to public on the 20th December 2006 and finished (with success) on the 21st July 2007.[9]

## VI. WHAT ABOUT THE FUTURE ?

Electronic manufacturer design faster components every month. Nowadays, one CPU is more likely to have two cores or even more. PCs with several CPU on the same motherboard also exist. PCs become very powerful. At the same time, Internet becomes also faster and faster. With DSL technology recently and FTTH in the future average connection has more bandwidth with less delay. This is why, grid computing can grow in term of computing power and efficiency.

Former project, such as SETI@Home were using only the CPU of the PCs. For instance : recent projects, like Folding@Home, also use the GPU of the graphic card. It appears that these processors were even better than classic CPUs (tens time faster)[10]. This has an explanation : GPU are designed for parallel computing and people use less their GPU than there CPU.

In the high technologic world, the current trend is to talk about mobility. Mobility implies that every electronic device is connected to others devices and can be linked to the Internet. This means that more computing power will be available from grid computing. Video game consoles (which can be very powerful), cell phones, media centers, PDAs, audio/video players are eligible to take part in a grip computing challenge. Mobile devices have seen their computing power increase drastically these last years and can run easily serious programs.

## VII. CONCLUSION

Over the years, grid computing has proven his efficiency, especially with large-scale project over the Internet such as Folding@Home or SETI@Home. Its utility is not limited to big scientific project ; grid computing can also be used in companies with a sufficient IT infrastructure.

Every future electronic device will contain one or more processors and can take part into distributed computation. The increasing speed of all the devices and of their links between each other will assure grid computing to be a good alternative to other solution of distributed calculation.

## REFERENCES

- [1] CERN. (2007) What is the grid? [Online]. Available: <http://www.ctan.org/tex-archive/macros/latex/contrib/IEEEtran/>
- [2] R. W. P. Plaszczak, *Grid computing*, 2005.
- [3] (2007) Top500 website. [Online]. Available: <http://www.top500.org/>
- [4] distributed.net official website. [Online]. Available: <http://www.distributed.net>
- [5] D. D. P. Anderson. (2003) Public computing: Reconnecting people to science. [Online]. Available: <http://boinc.berkeley.edu/madrid.html>
- [6] (2007) Seti@home official website. [Online]. Available: <http://setiathome.berkeley.edu>
- [7] (2007) Fightaids@home official website. [Online]. Available: <http://fightaidsathome.scripps.edu/>
- [8] "Folding@home official website," 2007. [Online]. Available: <http://folding.stanford.edu>
- [9] Genome comparison official website. [Online]. Available: <http://www.dbm.fiocruz.br/GenomeComparison>
- [10] (2007) Folding@home client statistics by os. [Online]. Available: <http://fah-web.stanford.edu/cgi-bin/main.py?qttype=osstats>