Towards Cloud Computing: Opportunities and challenges for e-Science

Dr. Fabrizio Gagliardi

EMEA & LATAM Director, Technical Computing, Microsoft Research
Fabrizio.Gagliardi@microsoft.com

Abstract

A new science paradigm has emerged in the last few years referred to as electronic Science (e-Science). It extensively uses simulation techniques based on software modeling which run on distributed computing infrastructures. In addition, it makes use of huge amounts of distributed and shared data captured by instruments or sensors and/or stored in databases, analyzed to provide new results for science. This distributed HPC and data environment allows sharing the acquired knowledge, accessing remote resources and enabling world-wide scientific collaboration.

The EU Datagrid project was one of the first examples of Grid computing, which was followed by the EGEE (Enabling Grids for E-science) project. The High Energy Physicists at CERN and the research community were among the first to adopt Grid computing. The two EGEE project phases have up to date integrated more than 50.000 CPUs in Europe and beyond, and 20 Peta Bytes (millions of GigaBytes) of storage, serving multiple application communities including HEP, Bioinformatics, Astrophysics, Computational Chemistry, Earth Sciences, Fusion. Some business/industrial applications are also adopting this distributed HPC computing model such as the automotive, finance, multimedia, and there a few examples of e-Government ones such as in the civil protection area.

Hence, Grid computing has delivered an affordable and high performance computing infrastructure to scientists all over the world to solve intense computing problems within constrained research budget. Business or industrial entities have also used similar technologies to increase the usage of their computing infrastructure and reduce their total cost of ownership (TCO). In addition, Grid computing is leveraging the advanced research networks to deliver an effective and irreplaceable channel for international collaboration.

Issues which hinder wider adoption of grid technologies in e-Science and industry have to do with the cost of operations and the overall complexity of the Grid, which aims at delivering secure and reliable services over multiple different administrative domains. The EGEE project, for instance, is spending more than €30 million per year operational expenditures (around half is covered by the EC) to run and support the 50.000 CPUs infrastructure (operations, middleware development and certification, application support, training, dissemination, etc.) Power consumption and heat dissipation are also becoming a new factor that needs to be considered seriously.

Elastic computing, Computing on the Cloud, Data Centres and Service hosting are offering on demand CPUs and storage with affordable pricing. As an example, Amazon Elastic Computing (EC2) offers a "small instance" CPU hour for $0,10, and is quite easy to use. In addition, it charges $0,10 per GB for data transfer in and $0,18 for data transfer out (of their systems). For datasets that stay at the Amazon system, one has to use the Amazon Simple Storage Services (Amazon S3), which have an additional cost of $0,15 per GB per month. With some rough calculations (as presented by EGEE management at a recent user project forum) it can be proved that already the Amazon services would be a bit more cost effective for EGEE, if all the EGEE job processing were performed in Amazon. Of course other considerations to do with the long term preservation and ownership of data can still make the in-house approach of EGEE more preferable. Other major stakeholders in the market such as SUN, Google and IBM are moving in the same direction towards offering similar services of on-demand computing, storage and hosting. In addition, many multi-cores and CPU accelerators promise potential breakthroughs, without needing to rely on computer clusters and the grid.

Microsoft is actively investigating this field, and the Technical Computing activity in Microsoft Research, is supporting e-Science initiatives in collaboration with leading scientists around the world to enable easier and better scientific discovery. We need to advance in making computing easy to use for the scientists to concentrate their energy in real science and not the computing tools.