BSC becomes part of the European Training Network in Big Data

The Storage Systems Group of the Barcelona Supercomputing Center (BSC) is participating in the Marie Curie Innovative Training Network titled "BigStorage: Storage-based Convergence between HPC and Cloud to address Big Data Challenges". The main goal of BigStorage is to train future data scientists in holistic, interdisciplinary approaches so that vast amounts of data can be fully exploited. This requires a reworking of the storage architectures underpinning high-performance computing (HPC) and Cloud infrastructures, with a focus on meeting highly ambitious performance and energy usage targets.

In BigStorage the BSC Storage System Group will lead the Use Case Analysis and Evaluation work package. The objectives are to undertake an in-depth review of the technical and architectural needs for data storage in four use cases (smart cities, the Human Brain Project, the Square Kilometre Array and climate science) to consolidate sets of requirements. These will be fed into other work packages. Furthermore, the team will also determine suitable benchmarks and specifications that reasonably represent these use cases and will be used to assess the BigStorage solutions produced by the other work packages.

"Data scientists will be some of the most sought-after researchers in the near future. These scientists need to have a multidisciplinary view of the whole data stack, from the way data is kept in persistent-storage devices to the algorithms needed to analyse the data, while bearing in mind how to take advantage of the potential parallelism. The objective of BigStorage is to train future data scientists who benefit from this global view of the data world," says Toni Cortés, Storage Systems Group Manager at the Barcelona Supercomputing Center.

For those interested in being part of this European Training Network, two research positions are available at BSC:

- PhD position on HPC-IO
- PhD position on Data placement via code inspection

The Storage System Research Group at BSC is developing a distributed storage platform to store and share data between applications. A distinguishing feature of this platform is that, from the point of view of the applications using it, data is stored in the form of objects, which include data, code (methods manipulating the objects) and behaviour policies that are also stored together with the data. The main purpose for storing methods in the platform is to bring execution close to the data in order to avoid unnecessary data transfers from the data store to the application that is executed on the client. The BigStorage team aims to take advantage of this feature to analyse the different stored algorithms to improve how data is placed and prefetched, and how parallelism is extracted.
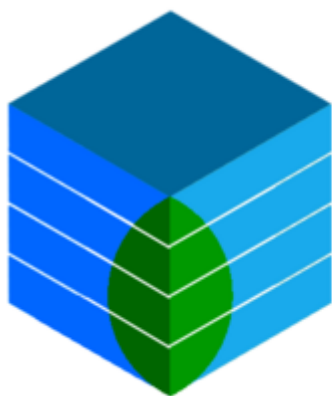
Within this network, BSC also aims to evaluate how algorithm programmers can prompt the system to improve energy efficiency and performance.

About European Training Network BigStorage

Currently, there is a lack of professionals with knowledge relating to the storage, management and analysis of Big Data; in addition, there is a gap between infrastructures for dealing with Big Data and applications using these volumes of data. In 2011, the McKinsey Global Institute published a study that found that, by

2018, there could be a shortage of up to 190,000 data scientists in the United States, representing a 50 percent to 60 percent gap between supply and demand. Similarly, European officials estimate that 300,000 data scientists will be needed in Europe in the forthcoming years. Other reports, such as those from PRACE and ETP4HPC, have also emphasised the need for skills in HPC, Cloud computing, storage, energy, or Big Data to maintain Europe's economy. In this context, a major goal of this project is to make a substantial contribution to the process of training these future experts.

More information: http://bigstorage-project.eu

Barcelona Supercomputing Center - Centro Nacional de Supercomputación

---