Energy aware big data management and processing

The development of big data application will be faced to the same energy challenge that the HPC domain : giving access to very large volume of data and performing complex processing within a reasonable energy budget. Even if big data applications are usually more distributed than HPC applications, it is desirable that:

- Not any location of the big data infrastructure consume more than tens of MW
- The energy consumption of the whole infrastructure is optimized

The objective of energy optimization must be one of the main criteria to choose an architecture for the big data application. To progress in efficient big data infrastructure some models must be developed that will help to gather the different essences of the production of data, the data features, the data processing and more globally the data life cycle. Based on these models, decisions on architectures can be taken in order to achieve this objective of energy efficiency.

The models that are needed are quite complex and cover different aspects of big data applications.

Data creation

Models that can represent the different source of the data

- Sensors, scientific instruments
- Behavior monitoring objects
- Published data (web, social network)
- Simulation data
- …

Meta data attached to data

In addition with the data themselves, that are important meta data to take into account for optimizing the architecture of the big data infrastructure

- Data access (type of network and cost to access data)
- Data relationship (locality and/or temporal proximity can be important)
- Dynamic nature of data
- Noise and uncertainty of data
- Privacy and security attributes
- …

Application processing patterns

Models that can represent the pattern of the most common processing algorithms:

- Statistic algorithms
- Correlation algorithms
- Machine learning algorithms
- Data base query processing algorithms

- Visualization algorithms
- …

Life cycle

Models that can represent :

- Iteration of the application (like updating statistic,…)
- Permanent storage requirements (for legal reasons, for scientific purposes…)
- Causality requirements (data set leading to result retrieval,…)
- Erasement of data
- …

The development of the different models will give the ability to assess different architectures for the big data application and to choose the one that leads to a reasonable energy efficiency while achieving the objective in performance.

The models can be coupled with simulation of the architecture tools to help to take the right decision in terms of :

- Distribution versus centralization
- Type and location of the storage devices
- Type and location of the computing devices
- Security management
- …

In summary, big data application energy optimization will be an important issue and will require important research to be conducted.

JF Lavignon

Bull

ETP4HPC Chairman

February, 2014