A distributed storage system for use in the Multiuser Bioinformatic's Laboratory of Embrapa

Leandro Cintra
Embrapa

Computation is currently acquiring a new status in the science, in a process which is commonly named e-science. Computational methods are changing from simple tools in the science process to be really integrated at the scientific method. As an example of that assertion we can consider the growth in the generation of data in many scientific fields. Biology is one area which a revolutionary process is in course, with new methodologies and technologies that can produce a great amount of data about many biological processes. The idea is to analyse the data to generate knowledge in biology, and to do that it is necessary a lot of storage space to data. Dedicated storage devices are very expensive, in part because they have a high confiability level and high throughput. These are indispensable characteristics in many computational applications, but there are cases where the storage capability is more important and so, it is necessary investigate others storage ways. Distributed file systems (DFS) have been used by a long time in problems where the confiability is less important than the capability of storage and the possibility of distributed processing. The indexing of the web, developed by companies like Google and Yahoo, is an example of these problems. In biology, principally in the omics areas, we have similar storage cases where to work with DFS is promising. For example, consider the case where a local copy of a public dataset is necessary to be used in scientific experiments, or the case where a great intermediate amount of data is generated at one scientific experiment. In this work we are proposing use the clusterFS or CEPH distributed file systems to construct a distributed storage system based in common hardware. The system will not be expensive because it is based in commonly available hard disks and computers. The resilience of the system is obtained in function of the redundancy. The throughput will be limited by the network connection, because the low rates recuperation of common disks will be compensated by the fact that data will stay distributed in the system and their access will be made using a lot of parallel hard disk devices. So the system will not have a high performance, but there are many scientific problems which are cpu bound. This means that little time is used in reading information and a lot of time is used in processing it. For those cases, high throughput is not necessary and the storage capability is more important in the system. With this work we would like to have a resilient and scalable distributed storage system adapted to be used in the analyses of biological data.

Keywords: Distributed Storage Systems, Distributed file Systems
Concentration area: Inst

⬅