An distributed environment for data storage and processing in support for bioinformatics analysis

Leandro Cintra

Embrapa - Brazilian Agricultural Research Corporation

Abstract

Nowadays, new technologies on data generation are amplifying in an unthought manner the amount of biological information accessible for analysing. This present us scenarios at which storage spaces and processing capabilities are bottlenecks on the computational system. The issue related with processing can be addressed with a computer cluster in which is possible to execute the tasks of an analysis in parallel. However, the storage issue is not well addressed for huge amounts of information with traditional tools. Normally, the clusters systems use NFS (network file system) to provide an unique information repository in the computation environment and this will have some disadvantage: a) storage space is limited by the capability of the server, which means that it will not be sufficient for those cases with great storage demand b) data throughput is limited by the server capability c) and all the system will be nonoperating if the file server go down by some reason. In this work, we investigate the use of a distributed file system (DFS) for data storage; associated with a distributed resource management (DRM) for control the parallel tasks execution. With a DFS system the environment can scale for petabytes of storage and operate in high throughput. Our environment was configured with six machines each one with 4xIntel Xeon E5-4620 (32 cores), 512Gb Ram and 887Gb of usable storage space in RAID 6. They were connected with a Gigabit ethernet network. These nodes with low storage capability were used as testing and specialized storage nodes are being provided with about 40Tb each one. We used the GlusterFS system as DFS and the Gridengine system as RDM. Bioinformatics tools with intensive IO activities were used in benchmarks of the system. Among them are blast, interproscan, SAM/BAM tools and the genome assembler MaSuRCa. Our tests were made considering local file system, NFS file system and GlusterFS file systems. The results indicated that the distributed storage system is stable, resilient and has potential to be used in production environment. Parallel environments based on processing of distributed tasks with distributed file systems are a promise approach for bioinformatic's demands and would be considered in projects working with big amounts of biological data.