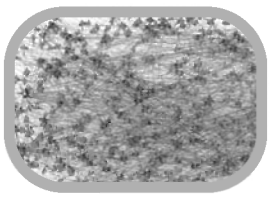
Visualization of large distributed data in Neuroscience

Introduction: The Human Brain Project is an exemplary case of *data driven science*. As a result, the large amount of complex data produced -at both, experimental and computational levels- poses important challenges from several points of view:

1. Data storage and sharing: mainly an infrastructure issue. It requires high performance computing (HPC) capabilities for data storage and mechanisms to distribute and share data between sites.

2. Analysis and knowledge extraction: special methods, sometimes still a subject of research, are required to gain knowledge from very large sets. *Visualization* techniques take advantage of the power of the human visual system to help gain insight from data, and become specially useful when it is not well known how to analyze data, requiring interactive exploration.

Despite the research done in large scale data handling, not much attention has been given to the low latencies and bandwidth required for analysis and visualization in an *interactive supercomputing* scenario. In this work, we explore how dCache - a distributed file system for very large data - can be successfully applied to several **use cases** in the HBP. Here we concentrate in the following ones:

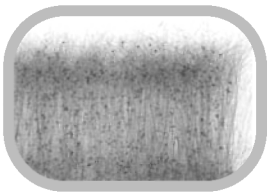


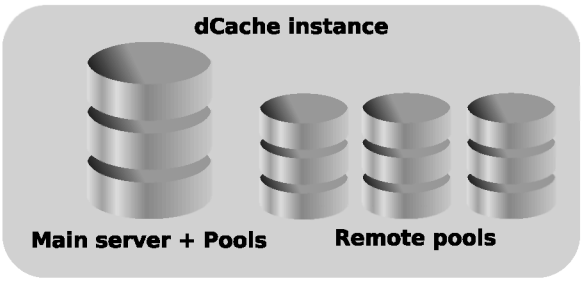
1 Web based access - MSPViz

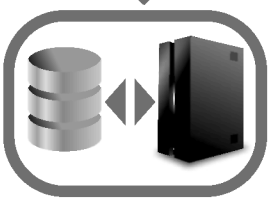
WebDav enables straightforward access to the filesystem of a dCache instance and has the advantage that web applications run requiring only a web browser. The price to pay is potentially expensive client-server data transfers. We studied this use case using MSPViz, a web enabled tool for schematic and abstract visualization of a synaptic plasticity model known as MSP (Model of Structural Plasticity).

2 NFS access - RTNeuron

This use case stresses the access to bigger data sets with different granularities. We will evaluate how the NFS service provided by dCache performs when accessing thousands of files of a few MB and thousands of small parts of a very large file, considering remote and local pools. RTNeuron, a programmable rendering engine for interactive visualization of neuronal microcircuity is our choice for this test case because it covers a wide range of access patterns.







3 Efficient pool usage

The distributed nature of dCache allows efficient utilization of the network bandwidth available at each site. Allocating the computational resources close to the data is essential to achieve good efficiency by accessing local pools. Remote visualization technologies are necessary to improve the response time. Only final images are sent to clients. Another aspect that needs to be explored is to allow data sources, such as simulators, to directly write to dCache storage space.

4 Collaboratory Integration

The Collaboratory is a key component inside the HBP platforms. As more analysis and visualization tools - which can be either running as web clients, or remote services - are integrated, it is crucial to provide a seamless way to connect data sources and sinks in the user's workflow. We will integrate the browsing of a dCache instance directory into the HBP Collaboratory, so applications and services can be directly fed with URLs, without the user worrying about where the data are physically located.

Future: Continuing with the collaborative trend inside the HBP initiative, we will use the proposed infrastructure for the application of collaborative visualization techniques centered in the analysis of complex neuroscientific data. This will allow several groups of neuronscientist, hosted in remote locations, to work collaboratively in the visual analysis of neuroscientific data.

Juan Pedro Brito Méndez (juanpedro.brito@upm.es)
Universidad Politécnica de Madrid, Task 7.3.2 & 7.6.4
Juan Hernando (EPFL), Benjamin Weyers (RTWH),
Vicente Martín (UPM)

Acknowledgments: This project has received funding from the [European Union's Horizon 2020 research and innovation programme] under grant agreement No 720270 (HBP SGA1)