DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Massively Parallel Spatial Indexing

Keywords: Scientific Data Management, Spatial Indexes, Neuroscience

Problem: Scientists in all kinds of disciplines like biology, chemistry, physics etc. produce vast amounts of data through experimentation and simulation. The amounts of data produced are already so big that they can barely be managed. And the problem is certain to get worse as the volume of scientific data doubles every year. In the DIAS laboratory we are working on next generation data management tools and techniques able to manage tomorrow's scientific data.

We work with neuroscientists in the Blue Brain Project (bluebrain.epfl.ch) to manage the vast amounts of data they produce. Their research, modeling and simulating a fraction of the rat brain, already produces gigabytes of data. With the recent upgrade of their computing infrastructure (IBM Blue Gene/P), the volume of data will soon be in the order of terabytes.

Current solutions are inadequate to manage this data volume and we are thus investigating new methods to index and store it in order to provide efficient access. A particular problem we are currently addressing is the retrieval of objects in space, i.e., accessing neurons based on their position. While it is simple to index several thousand neurons, we will have to do it for several millions or even billions of neurons.

To this end we have devised a new indexing approach called FLAT. With it, the brain data can be indexed efficiently, but more importantly, spatial range queries ("What elements are in a given region of the brain?") can be executed substantially faster.

Project: In this project the student will parallelize an already existing spatial indexing approach (FLAT) to run in a shared nothing cluster. The goal is to find a suitable distributed computation framework, to identify the parts of the FLAT which can be parallelized and finally study the performance of the resulting implementation.

Plan:

1. Identify available approaches for parallel spatial indexing
2. Identify parts of FLAT which can be parallelized
3. Study available frameworks to distribute FLAT in a shared nothing cluster
4. Implement FLAT with the previously identified distributed computation framework
5. Evaluate performance and compare to existing approaches

Knowledge: Knowing C/C++ and Java would be helpful.

Supervisor: Prof. Anastasia Ailamaki (anastasia.ailamaki@epfl.ch)

Responsible collaborator(s): Thomas Heinis (thomas.heinis@epfl.ch), Farhan Tauheed (farhan.tauheed@epfl.ch)

Duration: 4 months