Data Management System for Distributed Virtual Screening [1]

Ting Zhou†, Amedeo Caflisch‡
Department of Biochemistry, University of Zürich, Switzerland
Email: †t.zhou@bioc.uzh.ch, ‡caflisch@bioc.uzh.ch

Introduction

To manage HTD data effectively and efficiently, we have developed a distributed virtual screening data management system (DVSDMS) by integrating the open source database MySQL into high throughput docking. The essential concept of DVSDMS is the separation of data management from the main docking and ranking applications. DVSDMS can be used to dock millions of molecules effectively, monitor the process in real time, analyze docking results promptly, and process up to 10^8 poses by energy ranking techniques. The benchmark shows the DVSDMS running on a low cost Linux PC can distribute about 60 molecules per second.

Advantages

1. It separates data management from the main docking and ranking applications.
2. It is based on master-worker scheme. No balance problem.
3. It can overcome weaknesses of master-worker scheme easier.
   - SQL language and object-relational mapping
   - Many useful tools of the database for maintaining
   - Mature techniques for improving the performance
4. Python scripts is developed for connection all software packages.

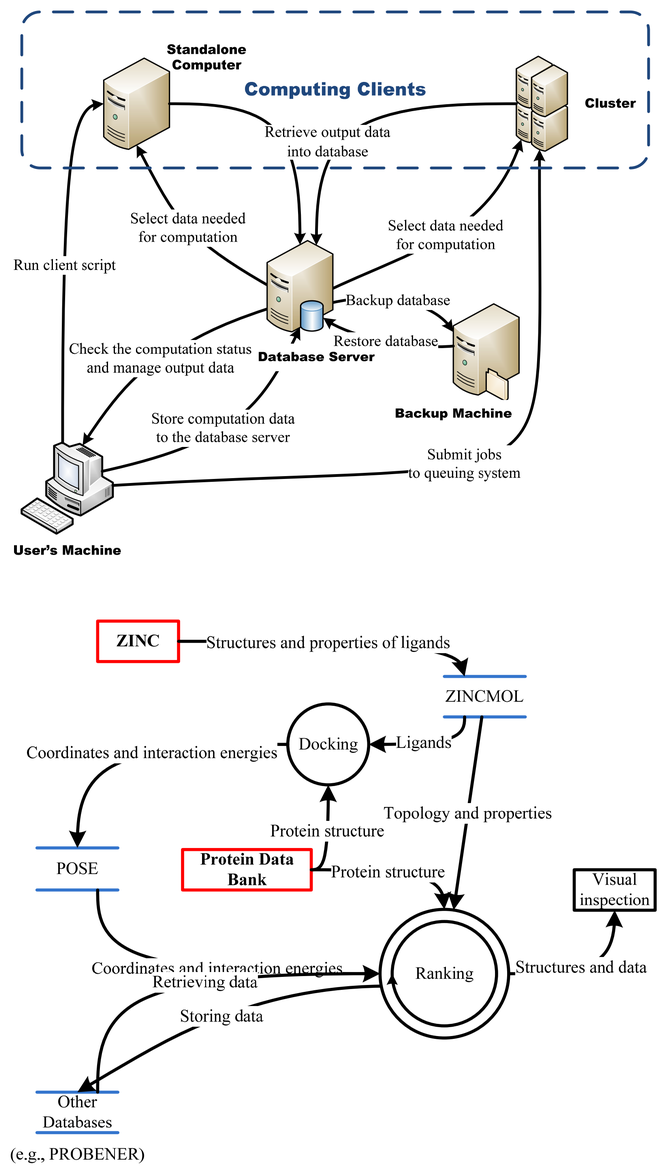
Performance Tuning

- Using database index
- Partitioning large tables
- Storage engine
- Local cache and bulk update
- Compressing molecule files

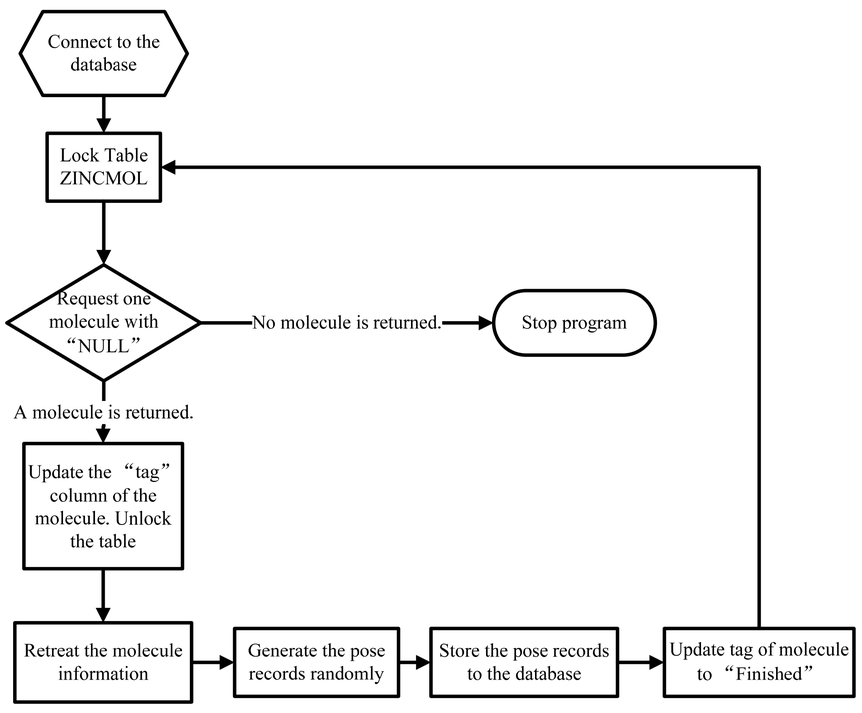
Software and its function

Software Name	Function
DAIM	Calculate the properties of molecules
AutoDockTools	Convert molecule file into pdbqt type
AutoDock	Dock small molecules into receptor
CHARMM	Add hydrogen atoms, and minimize structure
Witnotp	Convert molecule file types among mol2, pdb,
MOPAC	Calculate QM energies used for ranking
MySQL	The database management system
SQLAlchemy	The database toolkit for Python
Elixir	A declarative layer on top of SQLAlchemy

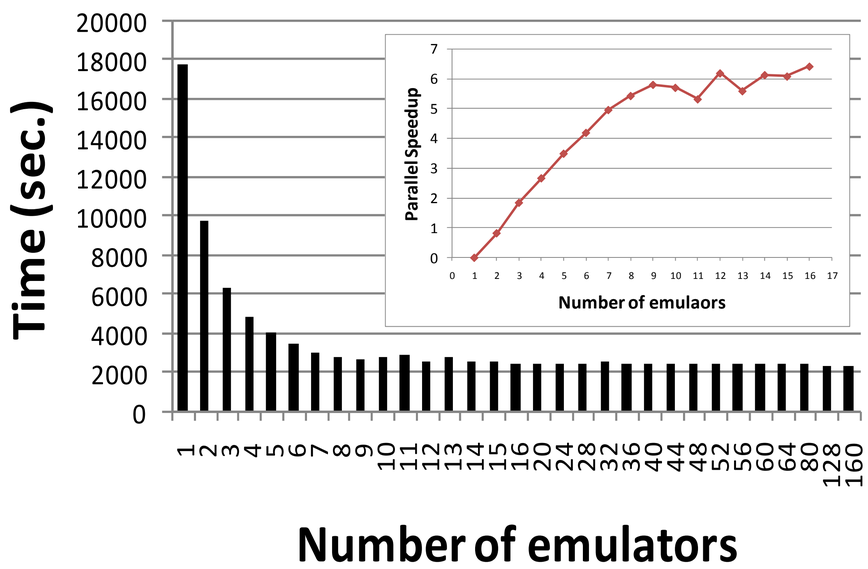
Schematic data flow in DVSDMS



Benchmark of the rate of distribution of molecule for DVSDMS



The flowchart of docking emulator



The duration of emulated docking of 128000 molecules with different numbers of emulators.

The maximal rate of distribution of molecules (maxRoD) is about 60 per second. The emulated docking does not require and CPU time, which is essential to estimate the maxRoD on the computer clusters we could access (<200 nodes).

[1] Ting Zhou, Amedeo Caflisch, J Chem Inf Model. 2008 Dec 12