

## Brian Moldover

---

**From:** B-Tech Consulting, Ltd. [B\_Tech\_Consulting\_Ltd@mail.vresp.com]  
**Sent:** Friday, February 19, 2010 12:55 PM  
**To:** brian@b-techconsulting.com  
**Subject:** Virtual Screening in the Cloud (Newsletter 02/2010)

[Click to view this email in a browser](#)



3701 Market Street  
5th Floor, Suite 119  
Philadelphia, PA 19104  
Office: 215.989.4880

[www.b-techconsulting.com](http://www.b-techconsulting.com)  
[info@b-techconsulting.com](mailto:info@b-techconsulting.com)

## Turning research data into knowledge for new discoveries

### Virtual Screening in the Cloud

Cloud computing is basically internet-based access to compute resource, whether it be servers, CPU cycles, disk storage, or networks. From my jaded point of view I'd call it 'an update to the old time-sharing systems. Needless to say I've been somewhat skeptical about putting my data out there where I'm not "in control".

We've been doing quite a lot of work in computational chemistry, virtual screening and docking to validate a set of compounds identified by a client as potentially binding to a particular receptor. This type of work is very computationally intensive, and a typical run can consume a multiple CPU cluster for days on end. On the other hand, it does not require much disk i/o or network bandwidth.

As this project expanded, we simply ran out of CPU cycles. The option to expand the compute server or buy a cluster was evaluated and rejected. Not that the cluster was so expensive - each 16 core compute server is only about \$4500 - but housing it in the data center was \$1000/month. It was hard to justify that expense unless the server was running nearly 24/7, which never happens.

We looked into the Amazon EC2 'cloud' computing environment, and it looks like it will be an ideal solution to our problems. In order to use the environment, you create an 'image' of the operating system and all of the software tools needed, called an "AMI". This took a day, most of which was setting up security and learning the environment. We used SSH/SCP and public/private keys, so security is not an issue.

### Experimental Design

There were three parts to the test.

### About the Company:

B-Tech, Ltd. provides contract research, consulting, and analytical services. We have worked extensively on the identification of differentially regulated genes using the Affymetrix and Illumina platforms, high-throughput DNA sequencing, SAGE and pathway analysis.

B-Tech has extensive experience in chemoinformatics, especially in structural databases and small molecule libraries.

We are also highly skilled in data integration, statistical analysis, and data presentation and visualization.

We are able to accommodate both large scale long-term projects as well as an individual experiment. All work is tailored to the needs of the researcher, and is a highly interactive process. In this way we ensure that the analytical methods will provide the best answer to the scientific questions being asked.

A typical analysis is completed within 2 weeks from receipt of data. If you have special needs, such as grant or meeting deadlines, we will do our best

1. Setting up the EC2 environment for virtual screening, which involved setting up security and creating a Linux-based system image with the screening tools and associated software that could be remotely installed into an Amazon EC2 server.
  - Docking software was Autodock, Scripps Research Institute
2. Benchmarking performance and cost using a small test set of compounds that had been previously screened on a local server. The test set was 24 compounds that we believe bind to HMG-CoA reductase. They varied in size from mw 221.2 to 1464. Obviously the latter is not really a potential drug, but it was one of the compounds of interest, so we carried it through.
3. Benchmark using a larger set of 10,000 compounds derived from the ZINC database, subset #3, druglike compounds.

#### Quick summary of results:

Original benchmark was screening on the local server: 2.13ghz quad-core, 4GB RAM rack mounted running CentOS 5.4. Disk and I/O are negligible, the entire dataset and results are less than 3 megs. Memory also is not that critical, the code is very efficient and the limiting factor is CPU. Original test took 17.5 hours and all 4 cores ran at almost 100% for the entire time.

Amazon EC2 environment (what they call a High-CPU Extra Large Instance): 7 GB of memory, 20 EC2Compute Units (8 virtual cores with 2.5 EC2 Compute Units each), 1690 GB of local instance storage, 64-bit platform. An EC2 compute unit is equivalent to a 1.0-1.2 ghz Xeon processor. Cost for use of this machine is \$0.68/hour, plus data transfer and storage fees.

**Test 1 completed in about 8.75 hours. Total cost was \$6.08:** \$5.95 for computer time, \$0.13 for monitoring, and two cents for 5 gigs of disk space! The majority of time was spent on the largest ligands. In summary, I could do virtual screening for around 18 hours a day 365 days a year and break even on the cost of purchasing a system. And we have not fully optimized the process.

Test 2 used 4 Amazon EC2 instances, same parameters as above. We took advantage of spot pricing at \$0.28 (average) per instance. Ligand set was 10,000 compounds from Zinc (subset 3) and same receptor. A script was used to do rudimentary load balancing across the instances and collate results. **Test 2 completed in time was 32.8 hours. Total cost was \$38.33.**

Where we go next with this is to use a larger test set and a different receptor. Build an environment with load balancing, and taking further advantage of what Amazon calls 'spot pricing' – the ability to bid on unused capacity at a lower cost. Dynamically allocate instances based on user criterion such as cost, time, spot pricing, etc. Build a GUI interface to set up docking and manage the process, results, etc.

to accommodate your requests. Our work is affordable, professional, and presented in a format designed for the bench scientist.

We currently perform contract research and support for several major universities, and we welcome reference checks. We have also worked closely with several top-tier pharmaceutical companies, which can also provide references.

#### About the Founder:

*Brian Moldover, Ph.D.* I have worked in bioinformatics since 1993, where I was a Research Fellow at NIH working in the Human Genome Project. Since then I spent 15 years working in the pharmaceutical industry for companies such as Warner-Lambert, Pfizer, Aventis, and Schering AG. I have held positions from Sr. Scientist through Vice-President of Global Research Computing. I have extensive knowledge of genomics and bioinformatics research, as well as significant business experience.

Because of my long industry experience, you can expect any work we perform to be done professionally, securely, and timely.

This is now a service we can offer, and are collaborating with both commercial and academic groups on expanding this technology and actively using it in research. Pretty exciting, and if you want even more information, just ask!

-brian

[Forward this message to a friend](#) | phone: (215) 989.4880

---

If you no longer wish to receive these emails, please reply to this message with "Unsubscribe" in the subject line or simply click on the following link: [Unsubscribe](#)

---

B-Tech Consulting, Ltd.  
3701 Market Street  
5th Floor, suite 119  
Philadelphia, Pennsylvania 19104



[Read](#) the VerticalResponse marketing policy.

No virus found in this incoming message.

Checked by AVG - [www.avg.com](http://www.avg.com)

Version: 8.5.435 / Virus Database: 271.1.1/2695 - Release Date: 02/19/10 07:34:00