

SimBioSys®

Technical Note

## Speed and Acceleration Options in eHiTS 2009

### Reference

Zholdos Z. eHiTS on the Cell B./E. -- Revolutionary Hardware Technology Opens New Frontiers in Molecular Modeling, SimBioSys Inc. white paper, 2007.

[http://www.simbiosys.ca/science/white\\_papers/eHiTS\\_on\\_the\\_Cell.pdf](http://www.simbiosys.ca/science/white_papers/eHiTS_on_the_Cell.pdf)

Zholdos Z. The fast and the furious: compare Cell/B.E., GPU and FPGA. SimBioSys blog, 2008

<http://www.simbiosys.ca/blog/2008/05/03/the-fast-and-the-furious-compare-cellbe-gpu-and-fpga/>

### Study Overview

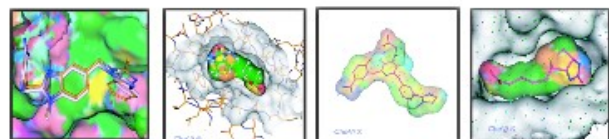
The trade-off between speed and accuracy is an inherent constraint on the reach of computational chemistry. In the docking paradigm, which is in itself a simplification of a complex problem, practical considerations determine the acceptable time-scales for docking, and within this time-frame, developers have to make conscious decisions as to what approximations can be done that would not undermine the viability of the results. The level of description of interactions, the flexibility of the ligand and the receptor, the conformational sampling approach and granularity, are just a few of the items that can be examined individually on the accuracy scale.

Although computational chemists have always pursued optimization of algorithms, for decades, Moore's law had provided an exponential computational acceleration much needed by most applications. In recent years, however, the CPU clock speed curve has levelled off, and processor manufacturers have turned to packing several cores on a single chip in order to address consumers' expectations for acceleration. At the same time, applications developers have started to give more attention to efficiency of codes and resources handling in order to achieve high performance computation. Field Programmable Gate Arrays (FPGAs), and Graphics Processing Units (GPUs) are gaining ground as mainstream platforms for scientific applications, and increasing attention is given to the emerging technology of IBM's Cell Broadband Engine. For reasons that are described in the references above, the Cell processor has been identified by SimBioSys as the ideal platform to deliver significant docking speedup without compromising the accuracy of the eHiTS approach.

The purpose of this study was to examine the speed of eHiTS 2009 with a particular emphasis on hardware based acceleration achieved on the PlayStation 3. A software-based acceleration available for conventional processors is discussed in the appendix, for the completeness of this discussion. Since docking speed depends on the ligand size, and the size of the binding pocket, datasets that are broad enough to cover a wide range of targets and ligands were chosen. Scaling of the speedup with ligand size and complexity, and with the level of accuracy is discussed below.

### Methods

To gather as many data points for this study as possible, we combined several data sets that are used



SimBioSys®

## Technical Note

routinely in SimBioSys for testing. Naturally, the sets differ significantly in the types of targets and ligands they cover, but since the aim is simply to time the docking, those aspects can be overlooked for the purpose of this study. For the most part, the data is available as pdb files of ligand-receptor complexes, or as pdb files for receptors and mol2 files for ligands, and therefore in most cases we used the following command lines:

```
ehits.sh -complex complex_file.pdb
```

or,

```
ehits.sh -receptor receptor_file.pdb -ligand ligand_file.mol2 -accuracy 1
```

where the default accuracy (accuracy 3) is shown in the first example, and the lowest accuracy is used in the second.

The eHiTS 2009 docking jobs ran on Sony's PlayStation 3 and on an Intel Pentium 4 3.00GHz (one hyperthreaded CPU).

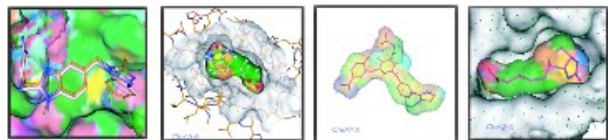
### Data

Over 6500 complexes have been used for this speed test. They offer a thorough sampling of ligand sizes and complexity. Table 1 shows the distribution of ligands according to their number of rigid fragments, and their number of rotatable bonds. Clearly, not all combinations are equally represented, but the massive amount of data is sufficient to highlight trends and capabilities.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
1	337																					
2		484	166	198	24	57	7	4	6	4	9	7										
3			446	252	366	70	56	21	6	18	3		1			3	1					
4				194	288	219	398	179	61	46	13	13	6	2	2							
5					94	124	142	114	387	89	31	11	24	3	17					1	2	
6						42	71	63	66	125	41	65	16	11	11	16				6	2	
7							14	19	44	41	52	124	30	31	16	13	8	3	1			1
8								13	8	9	33	26	73	72	23	11	4	1	1	1		
9									1	3	18	20	25	27	10	20	15	6	11	3	6	
10										1		4	4	16	15	14	8	9	10	4	11	
11											1			1	7	10	5	4	7	21	4	
12															8		7		8	8	7	

Table 1 Number of ligands sampled under each of the combinations of number of rigid fragments (left column) and number of rotatable bonds (top row).

The size of the binding pocket is another key parameter in determining the complexity of docking problems. However, since this property can be controlled by the user by determining a clip box, and by fixing the box margins, we did not analyze the results based on the volume of the binding site, and this issue will be addressed in a separate technical note that will focus on various aspects of binding site identification.



SimBioSys®

## Technical Note

### Results

The speedup factors and the average docking times for the data set are presented in Table 2. The numbers agree with results obtained with smaller datasets that show a clear upward trend of speedup factors with increasing accuracy level.

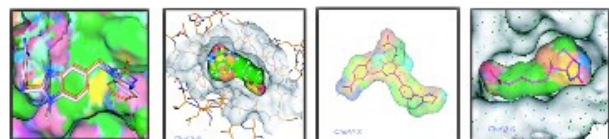
Accuracy Level	Measured Speedup Factor	Average docking time on the PS3
1	9.05	16 sec
3	12.4	37 sec

*Table 2 Average docking time and speedup on the PS3 compared to Intel Pentium 4*

To put the results in context, consider the following description of the eHiTS algorithm:

1. Ligands are divided into rigid fragments and connecting flexible chains.
2. RigidDock: each fragment is docked independently everywhere in the binding pocket.
3. PoseMatch: A graph matching algorithm reconstructs the ligand by matching the fragments.
4. Local energy minimization of the poses within the binding pocket.
5. Final pose scoring and ranking.

The difference between the accuracy levels in eHiTS is the number of fragment poses that are carried from RigidDock to PoseMatch, and the number of matched poses that are sent to optimization. Although eHiTS docks the fragments independently everywhere in the pocket, for combinatorial reasons, it is impractical to use all the docked poses. Based on scoring of the fragments' positions, as well as on considerations of diversity, eHiTS will choose representative subsets of the results of RigidDock and PoseMatch for further evaluation. The number of solutions carried from steps 2 and 3 in the algorithms are shown in Table 3. More detailed information is available in the user manual.



SimBioSys®

## Technical Note

Accuracy level	Number of solutions out of RigidDock	Number of solutions out of PoseMatch
1	2500	100
2	3500	275
3	4500	525
4	5500	850
5	6500	1250
6	7500	1725

*Table 3 Interpretation of accuracy levels in eHiTS 2009.*

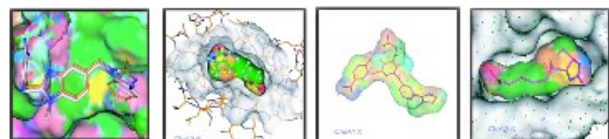
The various components in the algorithm are not equally speeded up on the Cell processor due to algorithmic design and hardware constraints. The following table gives a snapshot taken at some point during the development of eHiTS 2009 with the speedup factors corresponding to various steps in the algorithm. Notice that those numbers have not been evaluated on the final release of the software.

Function or code component, module	speedup on PS3
Scoring function (with rotamer optimization)	27X
Rigid Fragment Docking	21X
Pose Matching	13X
Conformation Minimization (ligand on its own)	24X
Final Pose Optimization (ligand in the active site)	22X

*Table 4 Speedup of components of the eHiTS algorithm on the Cell processor.*

The speedups in Table 4 do not translate directly to total speedup of the code due to read/write operations and other considerations. Nevertheless, the factors in Table 2 are significant and imply that a single PS3 is capable of replacing more than 10 Linux nodes that are more expensive, and involve much higher power consumption.

The dependence of the speedup on ligand complexity is demonstrated in Table 5 and Table 6. The observed trend is an increase of the speedup factor for increasing number of rigid fragments, and increasing number of rotatable bonds. The dependency with respect to the number of rotatable bonds appears to be more significant and can be rationalized using Table 4, which shows higher factors for scoring and minimization steps compared to rigid docking and pose matching.



SimBioSys®

## Technical Note

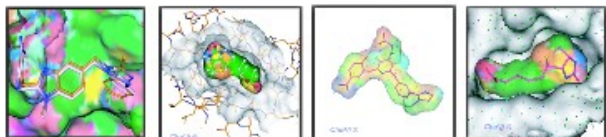
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	4																				
2		6	6	6	6	7	10	12	17	17	15	20									
3			6	7	7	8	7	7	9	12	16		15			15	18				
4				7	7	7	8	8	8	8	9	11	11	12	9						
5					7	7	7	7	8	8	8	9	10	8	8				11	14	
6						7	7	7	8	7	8	8	9	8	9	9			9	8	
7							7	7	8	9	8	9	6	9	9	9	10	10	9		10
8								8	8	8	9	8	9	9	9	8	9	10	9	10	
9									3	7	8	8	9	8	8	9	10	9	9	9	9
10										9		8	9	9	9	9	10	9	10	11	10
11															8	9	10	8	11	9	9
12															8		10		10	10	11

Table 5 Average speedups achieved on the PS3 at accuracy 1.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	6																				
2		8	7	7	7	9	12	15	21	21	18	21									
3			6	8	8	9	8	6	10	17	26		22			14	22				
4				7	8	8	9	11	10	11	12	16	19	19	14						
5					8	8	8	9	10	10	9	11	14	10	10				38	19	
6						9	10	9	10	8	10	11	11	11	12	12			16	7	
7							9	10	10	11	10	10	12	12	10	11	8	16	15		27
8								13	10	11	12	13	11	12	12	11	9	7	18	18	
9									2	10	11	10	16	12	11	13	12	13	10	13	12
10										10		15	14	11	16	16	16	16	11	19	12
11														9	12	15	15	11	19	16	15
12															11		16		15	11	14

Table 6 Average speedups achieved on the PS3 at accuracy 3.

To visualize the above results, the noise introduced by the large variance in the number of samples for each data-point had to be reduced. We adopted a simple nearest-neighbours weighted averaging scheme to produce Figures 1 and 2. The figures demonstrate the observed trends in terms of speedup factors with respect to the accuracy level and ligand complexity.



SimBioSys®

## Technical Note

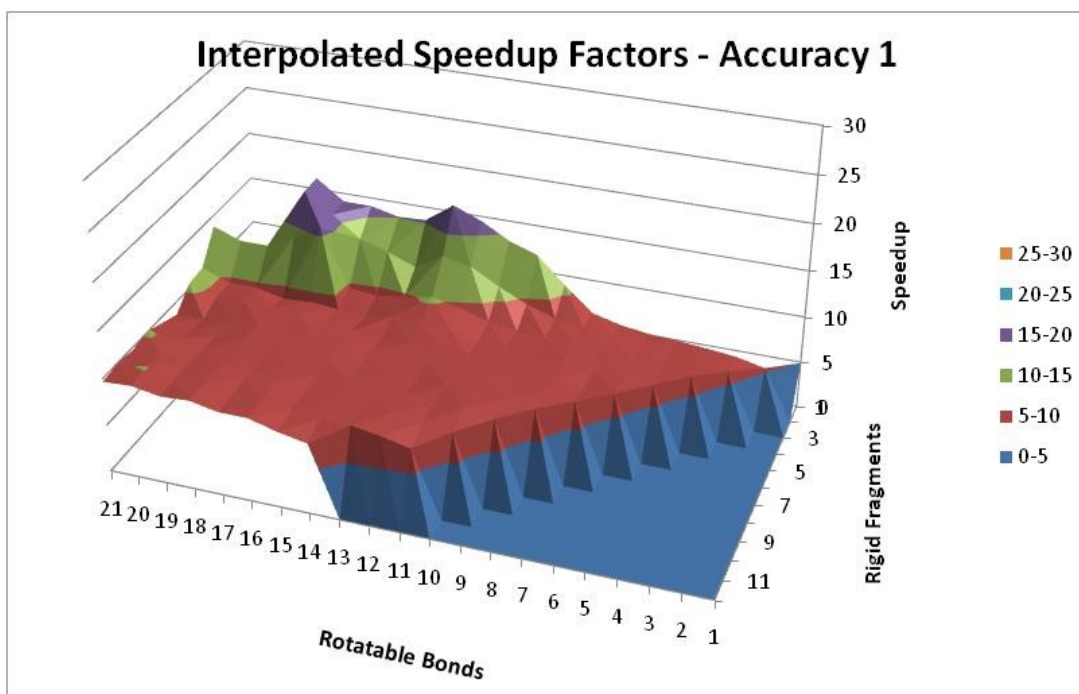


Figure 1 Speedup dependence on ligand complexity at accuracy 1.

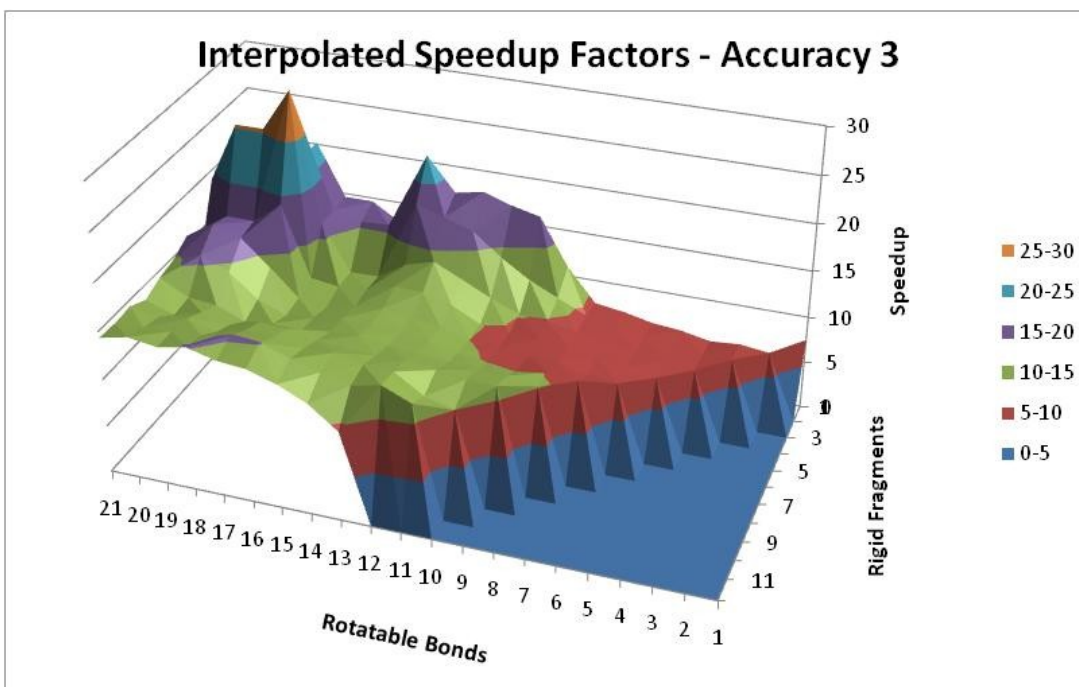
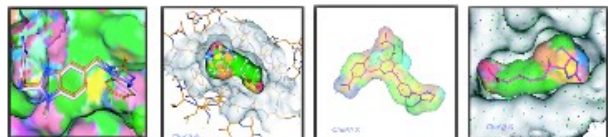


Figure 2 Speedup dependence on ligand complexity at accuracy 3.



SimBioSys®

## Technical Note

### Conclusion

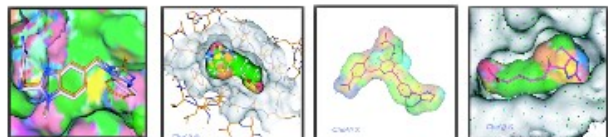
The above results demonstrate the dramatic hardware acceleration achieved using the Cell technology. Cell Processors are also notoriously efficient in their energy consumption, providing an exceptionally high performance per Watt, thus reducing the environmental footprint of the computational work. This is manifested in the Green500 list for 2008 which is topped by seven Cell-based supercomputers.

<http://www.green500.org/lists/2008/11/list.php>

The low energy consumption reduces the operational costs on top of lower hardware costs for cell-based systems. The following table offers a rough estimate of the savings associated with operating PS3 clusters.

Costs	100 CPU cluster	Cell/B.E.
Hardware	\$50K-\$100K	\$4K
Electricity (3 years)	\$45K-\$90K	\$3K
Total:	\$95K-\$190K	\$7K

The Cell processor is also available on IBM Blades in the QS20's series. Each blade center includes two cell processors with an extended amount of memory, and therefore offers an even higher speedup of docking.



SimBioSys®

## Technical Note

### Appendix – Acceleration using fragments SQL Databases

eHiTS' fragment-based algorithm invites a significant acceleration in screening scenarios by taking advantage of the reoccurrence of rigid fragments in different ligands. Whenever various ligands are docked into the same receptor, the RigidDock step can be carried out only once for each type of fragment. The **-usedb** flag available only for traditional CPUs invokes the use of an SQL database that stores docked poses for rigid fragments. As more ligands are docked, more fragments are stored in the database, and if encountered in a new ligand, they can be reused. The use of the SQL database dramatically reduces docking times as is shown in Figure 3. In this example, the docking speed increases substantially already after a few hundreds of ligands were docked. The average docking speed almost doubles after 4000 dockings.

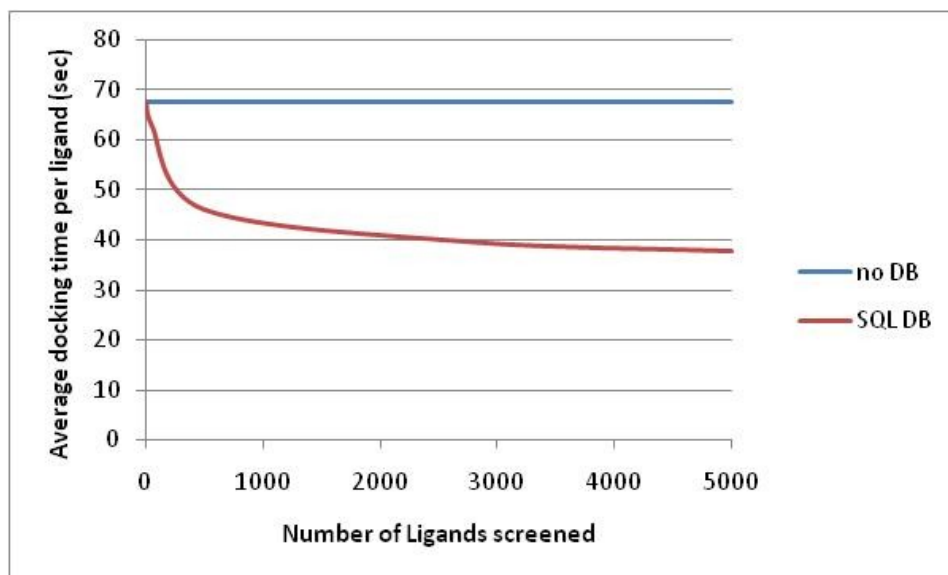


Figure 3 Docking acceleration using fragments SQL database.