Acceleration of Quantum Chemistry on GPGPU architecture towards faster virtual screening

Yuki Furukawa, Ryota Koga : X-Ability Co.,Ltd. Koji Yasuda : Nagoya Univ

E-mail : rkoga@x-ability.jp

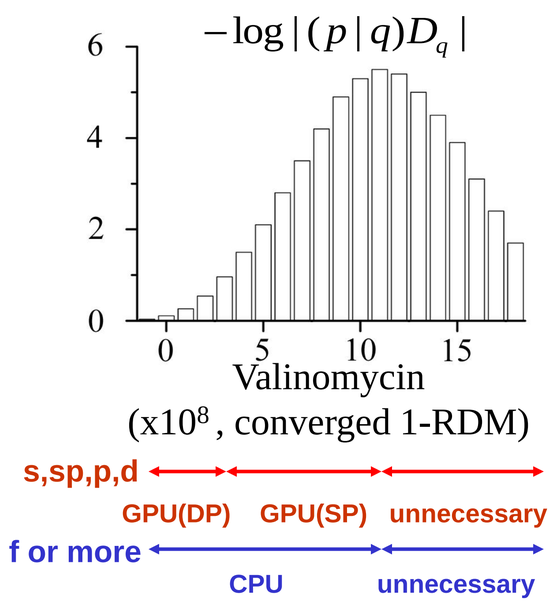
Purpose of this Research

1. To virtually screen biologically important molecules by using Gaussian, Fragment MO (electronic-structure calculations) by using GAMESS, Many-core processors such as Graphics Processing Units will be used to drastically accelerate heavy computations.

2. To analyze performance limiting factors of calculation and to develop better algorithms suitable for many-core processors.
3. To develop easy-to-use and commercial quality software.

Method (electronic-structure calculation)

1. Total energy and the gradient calculation with s, p, d basis functions are accelerated by GPUs.
2. Based on Yasuda [2008], most time-consuming steps (calculation of Coulomb, Hartree-Fock exchange, and exchange-correlation matrices…) are executed on GPUs. Remaining parts are calculated on CPUs. The Electron Repulsion Integrals (ERIs) and the Coulomb matrix in terms of primitive Hermite Gaussian basis are evaluated by GPU.
3. Use single precision arithmetic as much as possible for best speedup, without degradation of accuracy. The Schwartz upper bound of each ERI is first evaluated and ERIs which are small in magnitude are calculated on GPU with single-precision. The rests (10% of ERIs) are calculated on GPU with double precision.



4. Take special care for task scheduling and host-GPU communication time to keep as many CPUs and GPUs busy. Currently communication time between CPU and GPU was only about 10 % of GPU computation time. Multi-GPUs are used when available. DDI/OpenMP parallel version available for GAMESS, Linda parallel version available for GAUSSIAN.
5. Because of the strictly limited fast storage (e.g. registers) available on GPUs, conventional algorithms to evaluate Hermite ERIs do not run fast. We developed the special McMurchie-Davidson algorithm (submitted to JCTC).
6. We developed GPU computation library (XA-CUDA-QM) to accelerate famous QM software (GAUSSIAN, GAMESS-US…). All the functionality of these software are kept while the most important parts are accelerated by GPUs.

Summary

As the key technology of virtual screening of large molecules, DFT calculation by GAUSSIAN and FMO calculation by GAMESS were successfully accelerated by using GPUs. The calculated results were essentially the same, indicating cost and power efficiency of GPU-accelerated virtual screening.

Since the computational time of DFT was drastically reduced, the reduction of inter-node communication becomes necessary to run on massively parallel GPU clusters.

Thanks to matured QM software we can combine QM/MM (ONIOM…) and solvation models (continuum model, energy representation…) easily.

Practical application to catalytic reactions is now within the scope.

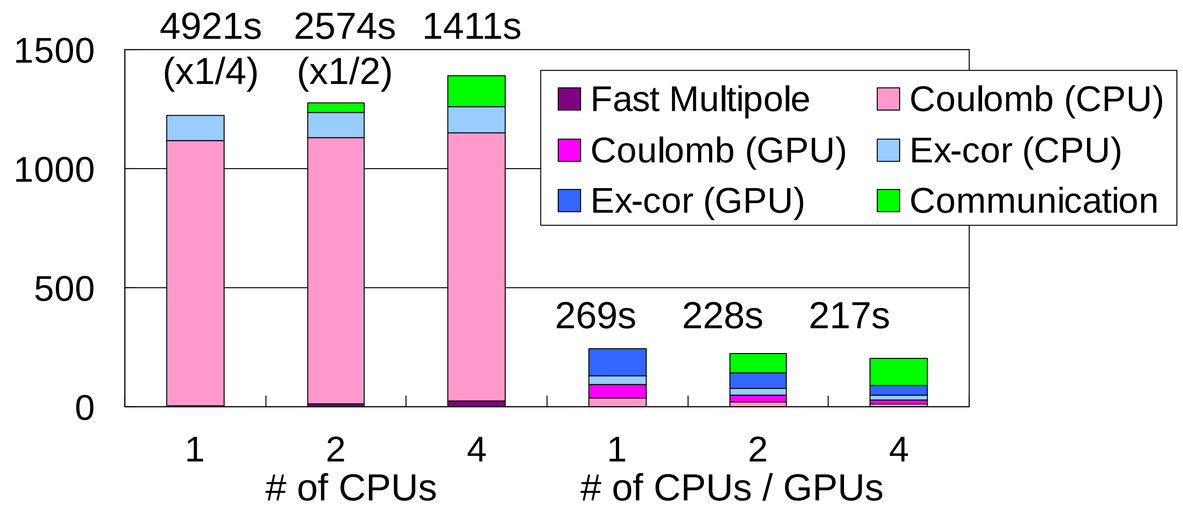
Results & Discussion

Table 1-1 : Performance of ERI calculation (% of theoretical peak)

Angular momentum *LP, LQ*	GPU single precision (NVIDIA GTX580 1581 GFLOPS)	CPU double precision (INTEL core i7 3930K 1 core 26.4 GFLOPS)	
		present	conventional
0,0	398 (25)	4.7 (18)	4.8 (18)
3,3	930 (59)	11.5 (45)	2.9 (11)
4,4	639 (40)	12.8 (50)	2.8 (11)
5,4	652 (41)	10.2 (40)	2.7 (10)

Our new ERI algorithm runs much faster on both CPU and GPU than the conventional (GAUSSIAN) one. It is four times faster for high-angular basis functions even on the same CPU because of the complete AVX vectorization.

Figure 1-2 : DFT calculation time (sec) of $Pt_{13}(H_2O)_{15}$ platinum clusters by GAUSSIAN



Most time-consuming part (Coulomb) was drastically accelerated by GPUs and the communication time dominates even on four nodes.

Table 2-1 : FMO2 monomer energies of each residue of Insulin(2HIU) by GAMESS (FMO-RHF/6-31G, a.u.)

Residue	E(CPU)	ΔE(GPU)
1(GLY)	-442.916776617	5×10^{-9}
5(CYS)	-1254.889732125	4×10^{-9}
9(ILE)	-347.549172133	1×10^{-8}
Total	-21635.448865252	3.3×10^{-7}

The dipole moment was exactly the same.

Total time using 2 nodes of system 1 was 3279.9 sec with CPUs, 790.5 sec with GPUs.

Table 2-2 : FMO environmental electrostatic potential and total acceleration rate of Insulin (2HIU)

Basis set / method	System 1		System 2	
	ESP	Total	ESP	Total
6-31G / FMO2, Single Point	13.8	4.1	3.3	2.1
6-31G*/ FMO2, Single Point	16.6	4.4	4.7	2.5
6-31G / FMO3, Single Point	19.5	3.7	4.8	2.3
6-31G / FMO2, Gradient	10.6	2.6	2.9	1.4

System 1: Intel Core i7-3930K @3.2GHz (6 core) , GTX580(cuda core 512) x 2, 32GB mem, CUDA 4.0

System 2: Intel Xeon E5-2650 @2.00GHz (32core) , a TESLA K20m(cuda core 2496), 64GB mem, CUDA 5.0 x 2 nodes

The calculation of ESP consumes more than 75% time. GPU calculates it 20 times faster than multi-core CPUs for system 1. The direct SCF method is unfavorable for each fragment calculation because they are so small that we can store all ERIs in a host memory.

Because of large number of CPUs in system2, the acceleration ratio was lower than that in system 1. Each fragment molecule is assigned to a GPU and is calculated independently, so we can increase performance by adding new GPUs.