Acceleration of Quantum Chemistry and Chemical Similarity calculations on GPGPU architecture towards faster virtual screening

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

E-mail : rkoga@x-ability.jp

Background

1. Biologically important macromolecules, such as proteins, are within the scope of electronic-structure calculation by using FMO because of the improvement of computational resources.
2. LINGO [Vidal et al. 2005] is a computational method to measure chemical similarity by comparing SMILES representations of two molecules.
3. GPGPU is successfully applied for many kinds of problems in high performance computation.

Purpose of this Research

First, based on Yasuda [2008], we develop XA-CUDA-QM to accelerate famous QM softwares (GAMESS-US etc) by mainly using NVIDIA GPU. We show the excellent FMO performance by XA-CUDA-QM. Second, we develop a LINGO similarity calculation method to utilize multiple CPUs and GPUs at the same time, based on SIML [Haque et al. 2010] originally developed to accerelate LINGO similarity calculations by using GPGPU.

Coding Policy

1. Total energy and the gradient of under FMO approximation.
2. Most time-consuming steps (the evaluation of Coulomb, Hartree-Fock exchange, and exchange-correlation matrices…) are executed on GPUs. Remaining parts are calculated on CPUs.
3. Use single precision arithmetic as much as possible for the sake of effective speedup, without degradation of accuracy.
4. Take special care for task scheduling and host-GPU communication time to keep as many CPUs and GPUs busy.
5. Modify LINGO original-code for multi-CPUs/multi-GPUs.

Algorithms

Electrostatic potential (Coulomb or J-matrix) :

The two-electron integrals and the J-matrix in terms of primitive Hermite Gaussian basis are evaluated by GPU (Direct J Engine).

$$J_{ab} = \sum_{cd} (ab|cd) D_{cd} = \sum_{p} E_{ab}^{p} \sum_{q} (p|q) D_q$$

Schwarz upper bounds are used to reduce the number of ERIs to be calculated, but we don't use integral symmetry (e.g. (p|q)=(q|p)).
The McMurchie-Davidson algorism is used to evaluate Hermite ERIs.
The communication time between CPU and GPU to send / receive p, q, D_q, J_p was only about 10 % of GPU computation time.

The 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 total count of ERIs) are calculated on GPU with double precision.

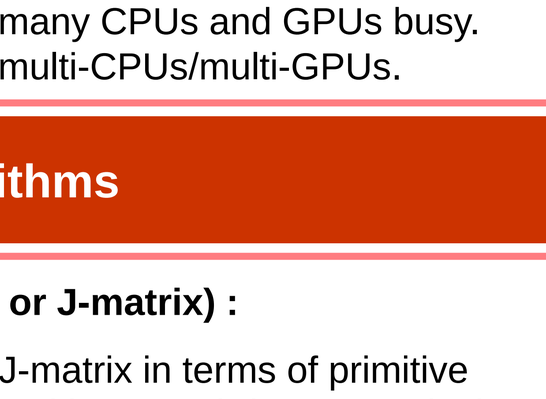
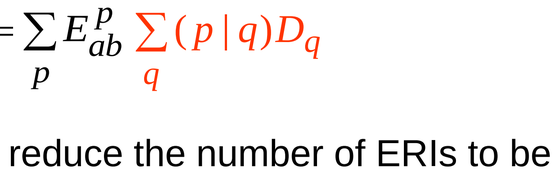


Fig 1. Valinomycin
(x 10^8 , converged D)



Environmental Electrostatic Potential (ESP):

J-matrix algorithm is reused.
This is the mainly time-consuming step of FMO.

LINGO algorithm on multi-CPUs and multi-GPUs:

Original algorithm implemented in SIML [Haque et al. 2010] is based on a single-CPU and single-GPU architecture. We have modified the source code so that LINGO algorithm can run parallel on multi-CPUs and multi-GPUs at the same time. In our implementation, calculation workload is divided into each set of CPU-GPU process with utilizing MPI.

Results & Discussion

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

Residue	E(CPU)	ΔE(GPGPU)
1(GLY)	-442.916776617	0.000000005
2(VAL)	-308.573606501	0.000000009
3(GLU)	-457.046042916	0.000000011
4(GLN)	-437.255799945	0.000000002
5(CYS)	-1254.889732125	0.000000004
6(CYS)	-1254.930631919	0.000000003
7(THR)	-344.327645726	0.000000007
8(SER)	-305.393476307	0.000000004
9(ILE)	-347.549172133	0.000000010

1. The value of dipole moment is exactly the same.
2. Total Monomer Energy
   E(Original) : -21635.4488652520
   E(GPGPU) : -21635.4488649211
   ΔE : 3.31×10^{-7} a.u.
3. Calculation Time (sec) :
   Original : 3279.944
   GPGPU : 790.527

Intel Core i7-3930K @3.2GHz (6 core) , GTX580 x 2, 32GB, CUDA 4.0 + same CPU, mem, compiler + GTX580 x 4 (2 nodes parallel)

Table 1-2 : FMO environmental electrostatic potential and total acceleration rate of Insulin(2HIU)

Basis set / method/ GPU	ESP	Total
6-31G / FMO2, Single Point / GTX680	13.8	4.11
6-31G*/GMO2, Single Point / GTX680	16.6	4.43
6-31G / FMO3, Single Point / GTX680	19.5	3.74
6-31G / FMO2, Gradient / GTX680	10.6	2.55
6-31G / FMO2, Single Point / GTX580	14.1	4.20

Intel Core i7-3930K @3.2GHz (6 core) , GTX580 x 2, 32GB, CUDA 4.0

The calculation of ESP consumes predominant time (over 75%) so that FMO2 accelerates so much. The direct SCF method is unfavorable for them because fragments are so small that we can store all ERIs in a host memory.

Table 2 : LINGO Similarity calculation time (16,384 x 16,384 chemical compounds all-to-all)

	Time(msec)	Speed up
1CPU	53,488	**x1.00**
4CPU	14,097	**x3.79**
60CPU (*)	2,640	**x20.26**
1CPU+1GPU	1,980	**x27.01**
2CPU+2GPU	1,010	**x52.96**
3CPU+3GPU	713	**x75.02**
4CPU+4GPU	529	**x101.11**

Intel Core i7-3930K @3.2GHz (6 core) , GTX580 x 2, 32GB, CUDA 4.0
(*) Only 60 CPU by Intel Quad Core XeonE5450 3.0GHz (many core), 32GB

Summary

The total energy of FMO and the dipole moments were essentially the same. GPGPU calculation of ESP becomes about 20 times faster than multi-core CPUs calculation.

LINGO calculation method that utilizes 4CPUs and 4GPUs improves calculation speed by more than 100 times compared to the speed calculated by one CPU. The calculation results were all coincident.

Future Theme

1. GPGPU acceleration of ERI with higher basis (metal protein).
2. Combination with molecular dynamics method
3. Massive parallelization on GPGPU cluster
4. LINGO similarity calculations on millions of compounds

(*) The super-computing resource was provided by Human Genome Center, Institute of Medical Science, University of Tokyo http://sc.hgc.jp/shirokane.html