Analisis Kinerja *Matrix Multiplication* Pada Lingkungan Komputasi Berkemampuan Tinggi (Cuda Gpu)

[1]Machudor Yusman,[2]Anie Rose Irawati, [3]Achmad Yusuf Vidyawan

[1]Jurusan Ilmu Komputer FMIPA Unila
[2]Jurusan Ilmu Komputer FMIPA Unila
[3]Jurusan Ilmu Komputer FMIPA Unila

Abstract

The increase of number and size of data, resulted the increase in the user needs for the ability of a computer to process large data. The new paradigm, parallel computing, is proposed to handle the problems which teaches that substantial-job can be split up into marginal-job by increasing the number of workers. One of the method is cluster computing which is using more than one processor to handle single process. Even it showed a significant increase in computing than the conventional one, the high price to build a cluster system becomes a constraint.

This study uses one of parallel computing method that is GPU computing and compares the result to cluster computing. GPU computing uses Graphics Processing Unit (GPU) to compute in parallel. The result of this study shows that by using GPU computing the use of processor can be maximized, and it shows that it has more capability in matrix multiplication than cluster computing.

Keywords: *cluster computing, CUDA-GPU, matrix multiplication, parallel computing*

1 Pendahuluan

Pertumbuhan teknologi saat ini semakin meningkat secara pesat dari waktu ke waktu. Penggunaan komputer dalam kehidupan sehari – hari mengakibatkan ketergantungan pengguna dalam setiap masalah yang dihadapi. Kompleksitas masalah yang dihadapi oleh pengguna memberikan implikasi jumlah data dan ukuran data yang juga meningkat seiring penggunaan komputer sebagai alat bantu penyelesaian masalah kehidupan manusia terutama di Indonesia. Peningkatan jumlah dan ukuran data setiap tahunnya mengakibatkan kebutuhan pengguna akan kemampuan komputer yang dapat memproses data dalam ukuran besar dengan jumlah banyak semakin meningkat.

Untuk mengatasi permasalahan tersebut, dikemukakan paradigma baru yaitu *parallel computing* (komputasi paralel) yang mengajarkan bahwa pekerjaan besar dapat dipecah-pecah menjadi pekerjaan kecil dengan jumlah pekerja yang banyak. Salah satu metode *parallel computing* ialah *cluster computing*, yaitu suatu metode *parallel computing* yang menggunakan jumlah prosesor lebih dari satu untuk menyelesaikan sebuah proses. Walaupun *cluster computing* terbukti mempercepat komputasi dibandingkan dengan komputasi konvensional, akan tetapi harga yang dibutuhkan untuk membangun sistem tersebut cukup mahal [1].

Penelitian ini memanfaatkan salah satu metode *parallel computing* lain yaitu GPU *computing* dan membandingkan hasil komputasinya terhadap *cluster computing*. GPU *computing* adalah komputasi yang memanfaatkan *Graphics Processing Unit* (GPU) untuk melakukan komputasi secara *parallel* [2] . Hasil penelitian menunjukkan bahwa pemanfaatan GPU *computing* selain memaksimalkan *processor*, juga terbukti memiliki kemampuan yang lebih dibandingkan dengan *cluster computing* untuk studi kasus perkalian matriks.

2 Metodologi

Pada penelitian ini, Metode penelitian yang digunakan dalam penelitian ini menggunakan beberapa tahap penelitian, yaitu proses studi literatur, persiapan pemasangan sistem, proses eksperimen, melakukan evaluasi hasil penelitian yang dilakukan, dan terakhir adalah melakukan dokumentasi.

2.1 Studi Literatur

Dilakukan dengan membaca beberapa penelitian terkait berupa jurnal, prosiding dan buku.

2.2 Persiapan Sistem

Menyiapkan lingkungan komputasi yang menggunakan arsitektur CUDA dan memasang software pendukung penelitian.

2.3 Eksperimen

Dilakukan proses komputasi perkalian matriks pada lingkungan GPU. Data perkalian matriks yang diinputkan berupa matriks nxn dengan sejumlah n data. Data tersebut didapat dengan menyesuaikan pada penelitian terdahulu yang dilakukan oleh Gani [1].

2.4 Analisa Hasil Penelitian

Dilakukan perbandingan hasil komputasi antara lingkungan GPU dan *cluster*.

2.5 Dokumentasi Hasil Penelitian

Dilakukan dokumentasi dalam bentuk hasil akhir dan kesimpulan penelitian.

3 Pembahasan

3.1 Persiapan Sistem

Sistem komputasi GPU yang dibangun menggunakan perangkat lunak CUDA Toolkit 4.1.2 dan menggunakan compiler visual studio 2008 untuk menjalankan program secara paralel. Sistem ini dibangun dengan sebuah unit komputer tunggal yang telah ter-integrasi dengan sebuah GPU/VGA sebagaimana sebuah unit komputer pada umumnya. Sistem ini dibuat dengan menggunakan bahasa pemrograman CUDA yang merupakan pengembangan dari bahasa C. CUDA inilah yang menjalankan seluruh proses eksekusi pada CPU maupun GPU.

Sistem yang telah dibuat menjalankan perhitungan secara bersamaan dalam melakukan proses komputasi perkalian matriks baik itu dalam CPU maupun pada GPU. Dan dilakukan pencatatan waktu komputasi secara otomatis, sehingga terlihat hasil yang paling optimal dilihat dari segi waktu.

3.2 Eksperimen

Pada tahap eksperimen, sistem komputasi GPU yang telah dibangun dan diintegrasikan dengan CUDA agar dapat menjalankan sistem komputasi parallel melalui layanan cuda toolkit dibandingkan dengan penelitian terdahulu yang menggunakan sistem komputasi *Cluster*.

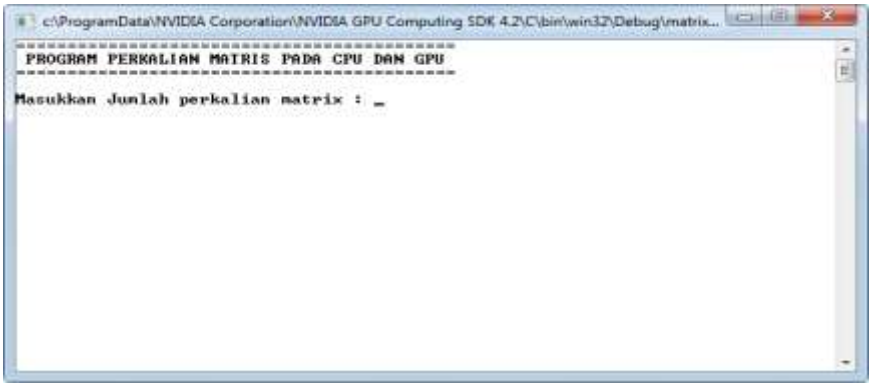
Aplikasi yang dijadikan perbandingan adalah aplikasi perkalian matriks. Pada penelitian terdahulu, aplikasi dibuat menggunakan bahas pemrograman C++. Namun, pada penelitian kali ini, aplikasi

dibuat dengan menggunakan bahasa CUDA yang tak lain adalah pengembangan bahasa C seperti halnya bahasa pemrograman C++. Penggunaan bahasa CUDA sendiri dikarenakan GPU yang digunakan adalah produk yang dikeluarkan oleh sebuah perusahaan grafis internasional NVIDIA, dan hanya didukung oleh bahasa pemrograman CUDA dalam mengakses mesin GPU tersebut.

Pada penelitan ini input data (n) yang digunakan menyesuaikan dengan penelitian terdahulu yang dilakukan oleh Gani [1] yaitu ordo matriks 500x500 hingga 3000x3000.

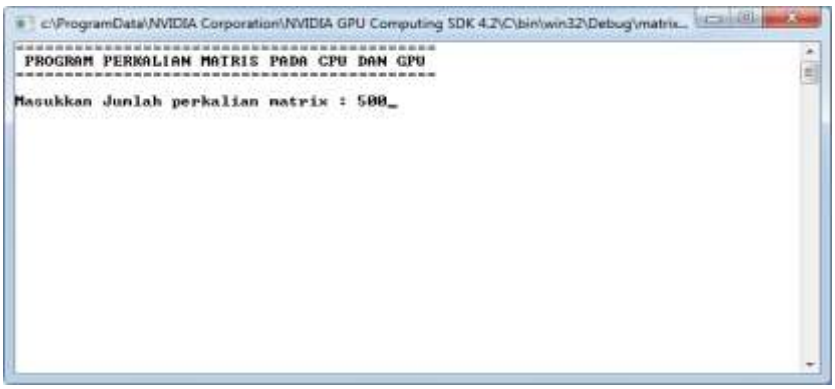
Dalam menjalankan program parallel menggunakan CUDA, langkah kerja dari aplikasi tersebut adalah sebagai berikut :

1. Lakukan *compiling* program perkalian matriks yang telah dibuat menggunakan Visual Studio 2008. Setelah di *compile* maka muncul jendela layar program perkalian matriks tersebut seperti pada Gambar 1.



Gambar 1 Tampilan Awal Hasil *Compiling*

2. Masukan data yang diproses. Pada penelitian kali ini, data yang dimasukkan yaitu ordo 500 hingga 3000. Sebagai contoh, dimasukkan data ordo matriks 500, maka ditampilkan seperti Gambar 2.



Gambar 2 Input Ordo Matriks n=500

3. Masukkan banyaknya jumlah blok dalam *grid* yang digunakan misalnya 10, seperti pada Gambar 3.

Gambar 3 Input Jumlah *Block*

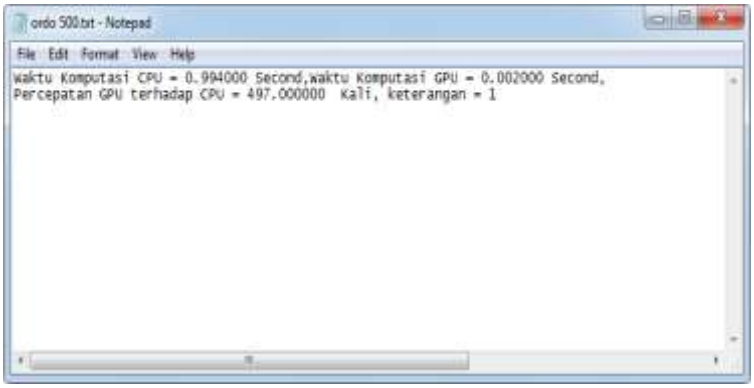
4. Masukkan banyaknya jumlah *thread* dalam blok yang digunakan misalnya 50, seperti pada Gambar 4.

Gambar 4 Input Jumlah *Thread*

5. Program langsung melakukan perhitungan secara otomatis seperti pada Gambar 5.

Gambar 5 Proses Perhitungan Komputasi Program

6. Tunggu hingga proses perhitungan selesai, maka jendela program seperti pada Gambar 6 tertutup dengan sendirinya, dan membuat sebuah teks baru yang memberikan info mengenai hasil yang telah didapat.



Gambar 4.6 Hasil Running Program dalam format File Teks

Langkah 1 sampai dengan 5 juga dilakukan pada matriks dengan ordo 1000, 1500, 2000, 2500 dan 3000.

3.3 Analisis Hasil Penelitian

Berdasarkan data yang telah didapatkan, maka dilakukan analisa perbandingan dengan penelitian terdahulu yang dilakukan oleh Gani [1] yang menggunakan sistem komputasi *Cluster*. Sistem *Cluster* yang dibangun menggabungkan dua buah komputer dengan spesifikasi identik, namun seolah-olah terlihat sebagai satu komputer tunggal. Pada penellitian komputasi *Cluster*, dilakukan proses sebanyak 4 kali pada setiap ordonya. Masing-masing proses dilakukan secara bertahap disertai dengan penambahan jumlah *processor* tiap prosesnya. Hasil penelitian terdahulu yang dilakukan oleh Gani [1] dengan menggunakan sitem komputasi *Cluster* dapat dilihat pada Tabel 1.

Tabel 1 Hasil Penelitian Komputasi *Cluster* [1]

Ordo matriks (nxn)	*Cluster* (detik)			
	Np 1	Np 2	Np3	Np4
500	4.523	1.525	1.187	1.108
1000	35.297	19.398	13.601	10.542
1500	104.275	82.157	53.169	39.410
2000	257.905	155.543	107.234	84.265
2500	431.841	328.265	227.645	170.059
3000	785.998	586.718	434.846	264.523

Keterangan :
Cluster = Hasil sistem komputasi paralel dengan menggunakan satuan waktu (detik)
Np1 = Jumlah Proses 1 dengan satuan waktu (detik)
Artinya komputasi dengan menggunakan 1 buah *processor*
Np2 = Jumlah Proses 2 dengan satuan waktu (detik)
Artinya komputasi dengan menggunakan 2 buah *processor*
Np3 = Jumlah Proses 3 dengan satuan waktu (detik)
Artinya komputasi dengan menggunakan 3 buah *processor*
Np4 = Jumlah Proses 4 dengan satuan waktu (detik)
Artinya komputasi dengan menggunakan 4 buah *processor*

Sedangkan, pada penelitian yang dilakukan dengan menggunakan sistem komputasi GPU. Hasil yang telah didapatkan untuk masing-masing ordo tertera pada Tabel 2.

Tabel 2 Hasil Penelitian Komputasi GPU

Ordo Matriks (nxn)	GPU (detik)
500	0.002
1000	0.002
1500	0.002
2000	0.004
2500	0.016
3000	0.167

Setelah didapatkan hasil perhitungan dari dua lingkungan komputasi yang berbeda, maka untuk melihat perbandingan hasiil-hasil tersebut dapat dilihat pada grafik Gambar 8.

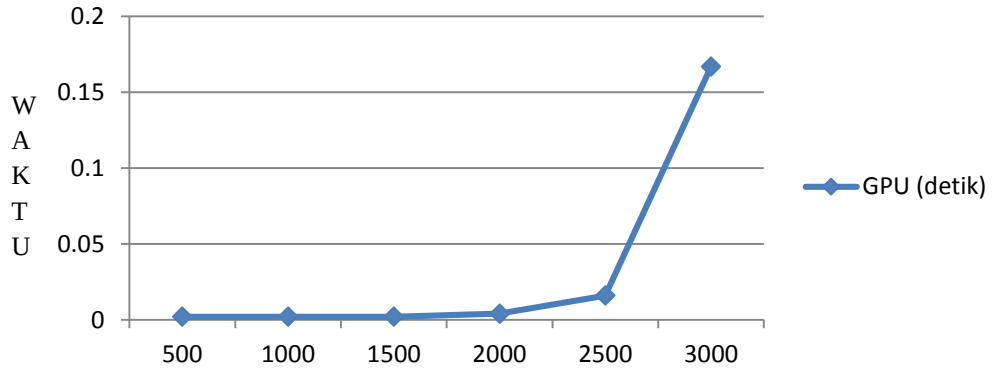
Analisa yang dapat diambil dari dua penelitian yang telah dilakukan, diantaranya :

1. Unit komputer yang digunakan pada lingkungan GPU dan *Cluster* [1] memiliki spesifikasi yang sama agar mendapat perlakuan yang adil diantara kedua belah pihak.
2. Data pada penelitian yang dimasukkan sama dengan penelitian terdahulu yang dilakukan oleh Gani [1], yaitu di mulai dari ordo 500, 1000, 1500, 2000, 2500 hingga 3000.
3. Pada lingkungan komputasi *Cluster* [1], dilakukan 4 kali proses. Np1, Np2, Np3 dan Np4. Proses-proses tersebut mengacu pada banyaknya jumlah prosesor yang digunakan untuk melakukan komputasi.
4. Pada tabel 3 dapat terlihat bahwa hasil perhitungan pada lingkungan GPU jauh lebih cepat dan optimal jika dibandingkan dengan lingkungan *Cluster*. Sebagai contoh, pada ordo matriks dengan jumlah n=500. GPU berhasil mencatatkan waktu eksekusi selama 0.002 detik, jauh lebih cepat jika dibandingkan wkatu eksekusi yang dicatatkan oleh *Cluster* yang hanya menghasilkan waktu 4.523 detik pada Np1, 1.525 detik pada Np2, 1.187 detik pada Np3 bahkan pada Np4 sekalipun yang berkisar diangka 1.108 detik. Begitu juga dengan *input* ordo yang lain. Waktu yang dihasilkan oleh GPU semakin jauh lebih cepat dibandingkan dengan oleh *Cluster*. Semakin

banyak data yang dimasukan, semakin banyak pula lah rentang waktu yang dihasilkan. Hal itu terlihat pada ordo matriks dengan jumlah n=3000. GPU menghasilkan catatan waktu hanya 0.167 detik sedangkan *Cluster* dengan jumlah proses Np4 cukup memakan waktu selama 264.523 detik.

5. Peningkatan kecepatan proses berbanding lurus dengan jumlah proses yang digunakan dan jumlah *processor* yang ada. Semakin banyak *processor* yang digunakan, semakin cepat waktu komputasi yang dihasilkan. Sistem komputasi GPU mampu jauh mengungguli sistem komputasi *Cluster* [1], karena penggunaan *processor* yang jauh lebih banyak. Pada sistem *Cluster*, *processor* yang digunakan hanya sebanyak 2 buah core ganda (*core 2 duo*) atau setara dengan 4 buah *processor*. Sedangkan pada sistem GPU menggunakan kartu grafis MSI GTX 550ti yang memiliki jumlah *processor* sebanyak 192 buah [3]. Hal itulah yang mendasari alasan kecepatan yang dihasilkan pada sistem komputasi GPU jauh lebih unggul dibandingkan dengan sistem komputasi *Cluster*.

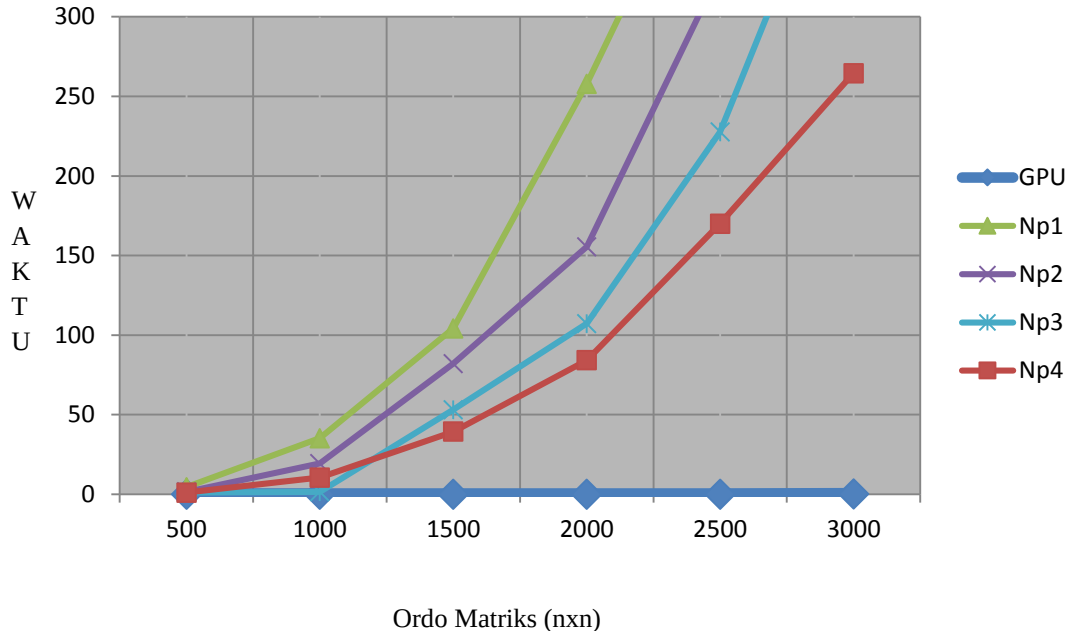
Untuk lebih memahami perbedaan hasil komputasi diantara kedua lingkungan yang berbeda, maka ditampilkan dalam bentuk grafik. Pada Gambar 7 ditampilkan grafik untuk komputasi GPU.



Gambar 7 Grafik Hasil Komputasi GPU

Berdasarkan Gambar 7, grafik menunjukkan waktu yang terus bertambah. Hal itu dirasa sangat wajar mengingat input data yang digunakan juga mengalami penambahan nilai. Namun, waktu yang dihasilkan pun masih sangat cepat, yaitu kurang dari 1 (satu) detik walaupun input yang dimasukan cukup besar. Hal itu menunjukan waktu komputasi yang cukup optimal di lingkungan komputasi GPU.

Untuk membandingkan hasil komputasi GPU dengan komputasi terdahulu yang menggunakan sistem *Cluster* [1], ditampilkan grafik perbandingan seperti pada Gambar 8.



Gambar 8 Grafik perbandingan waktu komputasi antara GPU dan *Cluster*

Terlihat dari Gambar 8 bahwa kecepatan pemrosesan *parallel* menggunakan sistem komputasi GPU jauh lebih cepat dibandingkan kecepatan dengan menggunakan sistem komputasi *Cluster*. Perbedaan kecepatannya sangat signifikan. Hal itu bisa dilihat dari Perbandingan Kecepatan yang dihasilkan diantara kedua buah lingkungan tersebut. Untuk melihat Perbandingan Kecepatan yang dihasilkan dilakukan perhitungan-perhitungan. Dari perhitungan, dapat diartikan sebagai, rata-rata perbedaan kecepatan GPU dengan Np1 adalah sebanyak 42055.096 kali. Rata-rata perbedaan kecepatan GPU dengan Np2 adalah sebanyak 28613.898 kali. Rata-rata perbedaan kecepatan GPU dengan Np3 adalah sebanyak 19404.670 kali. Dan rata-rata perbedaan kecepatan GPU dengan Np4 adalah sebanyak 14702.226 kali. Hasil perhitungan tersebut menunjukkan bahwa, kecepatan yang dihasilkan oleh komputasi GPU jauh lebih optimal jika dibandingkan dengan komputasi *Cluster*. Bahkan kecepatan GPU masih jauh di atas kecepatan komputasi *Cluster* yang paling cepat (Np4) sekalipun.

Berdasarkan hasil penelitian dapat diasumsikan perhitungan secara ekonomis dalam hal biaya pengadaan unit. Dari penelitian ini didapatkan bahwa dengan biaya pengadaan unit yang cenderung setara atau bahkan lebih murah. Asumsi biaya pengadaan yang diambil dari salah satu situs penjualan komputer lokal yaitu rakitan.com per tanggal 12 Desember 2013, bahwa satu sistem komputer spesifikasi yang dibutuhkan untuk sebuah sistem komputasi *Cluster* [1] membutuhkan 2 perangkat komputer dengan biaya masing-masing perangkat sebesar Rp 2.084.000,- ; artinya dibutuhkan nominal biaya sebesar Rp 4.164.000,-. Sedangkan untuk biaya pengadaan suatu sistem komputer dengan spesifikasi yang diperlukan untuk sebuah sistem komputasi GPU dibutuhkan 1 buah perangkat komputer dengan biaya Rp 2.084.000,- dan penambahan 1 unit GPU dengan estimasi Rp 1.220.000,-. Total biaya yang harus dikeluarkan adalah sebesar Rp 3.304.000,-, jauh lebih murah dibandingkan biaya yang harus dikeluarkan untuk pengadaan sistem komputasi *Cluster*. Walaupun dengan biaya yang dikeluarkan lebih sedikit, perbedaan kecepatan pemrosesan sangat besar. Dengan demikian, sistem komputasi GPU memberikan alternatif media komputasi yang kuat namun dengan harga yang relatif sama atau bahkan lebih murah.

4. Kesimpulan

Kesimpulan dari hasil penelitian yang telah dilakukan, adalah sebagai berikut:

1. Hasil penelitian yang diperoleh menunjukkan bahwa sistem komputasi GPU memiliki kemampuan yang jauh lebih unggul dari segi waktu dibandingkan dengan sistem komputasi *cluster*.
2. Penggunaan jumlah prosesor yang semakin banyak berdampak sangat besar terhadap waktu eksekusi suatu program sehingga menghasilkan kecepatan yang sangat tinggi.
3. Pemanfaatan sistem komputasi parallel dapat membantu menyelesaikan persoalan dengan jumlah data yang besar.
4. Sistem komputasi GPU merupakan alternatif penyelesaian persoalan yang dapat digunakan mengingat biaya pengadaan unit yang lebih murah dibandingkan sistem komputasi *cluster* dan hasil eksekusi yang diperoleh memiliki perbedaan yang sangat signifikan.

5. Refference

[1] Gani, R.A. 2011. Analisis Kinerja Sistem Komputasi Grid Menggunakan GLOBUS Toolkit dan MPI. IlmuKomputer : Universitas Lampung.

[2] Owens, J., Houston M., Laubke D., dan Green S., 2008.*Graphics Processing Units powerful, programmable, and highly parallel are increasingly targeting general-purpose co ing applications*. Prosiding IEEE.

[3] *NVIDIA* Corporation. 2007. *NVIDIA* CUDA *Programming Guide ver. 1.0*. Santa Clara.