Studi Literatur Pemanfaatan *High Performance Computing* dalam Bidang *Bioinformatics*

Maya Nirmala[1], Indriyani[2], Muhamad Reza Shahensha[3], Mutia Kentuah Nieate[4], Nova Eka Diana[5]

Fakultas Teknologi Informasi Jurusan Teknik Informatika
Universitas YARSI
Jakarta Pusat, Indonesia
[1]mayyaanirmala@gmail.com, [2]indriyani.f82@gmail.com, [3]muhammad7rezza@gmail.com,
[4]mkentuahn@gmail.com, [5]nova.diana@yarsi.ac.id

***Abstrak*— Komputasi tingkat tinggi atau *High Performance Computing* (HPC) dapat membantu memudahkan kinerja dalam berbagai aspek, salah satunya pada bidang bioinformatika. Penerapan HPC dalam bidang bioinformatika antara lain pembuatan basis data dan pengembangan algoritma untuk analisis sekuens biologis. Tujuan dari studi literatur ini untuk menganalisis perkembangan jumlah *paper* yang telah dipublikasikan mengenai penggunaan teknik HPC dalam bidang bioinformatika dalam kurun waktu lima tahun terakhir dari tahun 2012 sampai 2016. Teknik HPC yang digunakan untuk membantu menyelesaikan permasalahan bioinformatika, diantaranya adalah *Graphics Processing Unit* (GPU), *Multi-Core*, *Multi-processor* dan *Cluster Computing*. Metode yang digunakan dalam studi literatur ini adalah pengumpulan, filtrasi, klusterisasi, visualisasi dan analisis data. Dalam studi literatur ini, kami menunjukkan bahwa teknik GPU merupakan teknik yang paling cepat dan efisien dalam menghitung *big data*, khususnya data bioinformatika.**

Kata Kunci—studi literatur; bioinformatika; high performance computing; Graphics Processing Unit (GPU); Multi-Core; Multi-processor; Cluster Computing.

I. PENDAHULUAN

Komputasi tingkat tinggi atau *High Performance Computing* (HPC) dapat membantu memudahkan kinerja dalam berbagai aspek, salah satunya pada bidang *Bioinformatics* atau bioinformatika. Bioinformatika merupakan gabungan antara ilmu biologi dan teknik informasi. Penerapan HPC dalam bidang bioinformatika antara lain pembuatan basis data dan pengembangan algoritma untuk analisis sekuens biologis. Contoh topik utama pada bidang bioinformatika meliputi basis data untuk mengelola informasi biologis, *sequence alignment*, prediksi bentuk struktur protein maupun struktur sekunder RNA, analisis filogenetik, dan analisis gen.

HPC sangat penting untuk membantu menyelesaikan permasalahan pada bidang bioinformatika tersebut, karena data yang diperlukan sangat banyak dan apabila hanya menggunakan komputasi sederhana akan membutuhkan waktu yang lama untuk mengeksekusinya. Misalnya untuk melakukan penyelarasan DNA besar, memerlukan teknik GPU untuk mempercepat proses eksekusi program.

Ada berbagai macam teknik HPC untuk membantu menyelesaikan permasalahan bioinformatika, diantaranya adalah *Graphics Processing Unit* (GPU), *Multi-Core, Multi-processor* dan *Cluster Computing.* Kami akan menjelaskan perkembangan mengenai keempat teknik HPC tersebut yang digunakan untuk membantu proses penelitian pada bidang bioinformatika berdasarkan banyaknya *paper* yang telah dipublikasi pada IEEE dalam kurun waktu lima tahun terakhir dari 2012 hingga 2016.

Paper ini terdiri dari beberapa pembahasan, yaitu Pendahuluan, Studi Literatur, Metodologi serta Hasil dan Analisis. Pada Pendahuluan kami membahas sedikit mengenai kegunaan HPC untuk bioinformatika. Studi Literatur kami membahas beberapa *paper* yang terkait dengan HPC untuk bioinformatika, serta mengapa bioinformatika tersebut membutuhkan HPC dalam pengembangannya. Metodologi, kami membahas bagaimana cara kami melakukan penelitian ini dari awal pengumpulan data hingga analisis data. Hasil dan Analisis, kami membahas bagaimana perkembangan HPC untuk bioinformatika sesuai dengan data yang telah kami kumpulkan dan sesuai dengan kriteria yang diidentifikasi.

II. STUDI LITERATUR

Beberapa penelitian sebelumnya menggunakan teknik *Graphics Processing Unit* (GPU) untuk bioinformatika. Misalnya Konur dan Kiran pada penelitian mereka menggunakan teknik GPU. Platform yang digunakan adalah FLAME (*Flinders Large Scale Agent-based Modeling Environment*) untuk mensimulasikan model multi-agen pada arsitektur perangkat keras paralel, termasuk komputer dengan kinerja tinggi. Hasil dari penelitian menunjukkan bahwa FLAME GPU adalah kandidat yang baik untuk mensimulasikan dinamika koloni bakteri besar yang dibentuk oleh banyak sel individu. Oleh karena itu, ini adalah kerangka komputasi yang mudah digunakan untuk menganalisis sistem komputasi membran yang kompleks [1]. Penelitian yang dilakukan oleh Burkitt dan Walker juga menggunakan teknik GPU untuk mengekstrak informasi tentang struktur jaringan biologis dari data citra statis. Tujuan dari penelitian ini adalah untuk menghasilkan model komputasi *oviduct* 3D unik yang digunakan untuk menyelidiki dampak lingkungan 3D yang

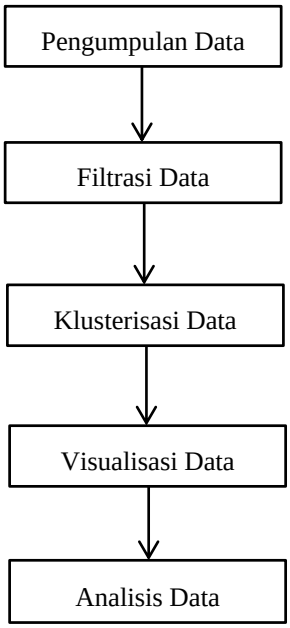
kompleks terhadap transportasi dan navigasi sperma pada mamalia. Hasil dari penelitian mereka adalah GPU menghasilkan kinerja baik untuk pengolahan citra maupun simulasi berbasis fisika partikel [2]. Penelitian yang dilakukan oleh Domanski dan Bednarz juga menggunakan teknik GPU pada penelitian ringkasan algoritma heterogen dan aplikasi CSIRO untuk memecahkan masalah sains praktis. Hasil penelitian menunjukan bahwa *cluster* GPU yang besar dapat digunakan untuk mempercepat berbagai aplikasi sains praktis [3]. Johnshon dan Shafer dalam penelitiannya memanfaatkan perhitungan paralel secara besar-besaran yang dimungkinkan oleh GPU pada diagnosis kanker untuk mempercepat penemuan panel biomarker yang optimal. Hasil penelitian ini menunjukan bahwa perhitungan masalah *parallelizable* dalam bioinformatika dapat ditingkatkan dengan perhitungan pada GPU [4].

Beberapa penelitian sebelumnya juga menggunakan teknik HPC yang lain yaitu *Multicore* untuk bioinformatika. Misalnya Hosny dan Hussein pada penelitian menggunakan teknik CPU *multicore* untuk mempercepat metode menyelaraskan urutan DNA dengan optimal. Hasil penelitian menunjukkan bahwa solusi yang diusulkan mencapai rekor tinggi dibandingkan solusi lain yang menargetkan sasaran yang sama dengan persyaratan perangkat keras yang lebih sedikit [5]. Penelitian yang dilakukan oleh Leite dan Melo menggunakan teknik *Multicore* untuk menghitung skor atau keselarasan terbaik antara dua urutan genomik dengan aplikasi Smith-Waterman (SW). Hasil yang diperoleh dengan arsitektur peneliti dan implementasi SW *MapReduce* di lima *Public Clouds* menunjukkan bahwa, hanya dengan menggunakan kuota gratis, peneliti dapat menjalankan aplikasi SW melalui basis data *big genome* pada waktu yang telah ditentukan[6]. Dan pada penelitian yang dilakukan Jiménez dan Pulido. Platform yang digunakan *Artificial Bee Colony* yaitu untuk menyimpulkan filogenies sesuai dengan kriteria parsimoni maksimum dan kriteria *likelihood* maksimum. Hasil penelitian menunjukkan bahwa teknik yang digunakan peneliti dapat memperbaiki pendekatan lain teknik komuputasi kinerja tinggi yang canggih pada kumpulan data yang besar. Tujuan peneliti adalah untuk membuktikan bahwa pendekatan untuk arsitektur *multicore* simetris dapat meningkatkan nilai percepatan dan efisiensi yang dilaporkan oleh komputasi berkinerja tinggi lainnya dalam literatur [7].

Beberapa peneliti lain sebelumnya juga menggunakan teknik *Clustering* untuk bioinformatika. Misalnya pada penelitian Seoud dan Eldin menggunakan platform BIG-BIO untuk menghitung jumlah kemunculan setiap kata dalam teks. Kemudian mengekstrak kata-kata unik dalam urutan molekul dan bisa menganalisa *big data MapReduce Hadoop Cluster* untuk aplikasi bioinformatika lebih cepat dan lebih efisien. Hasil penelitian menujukan bahwa BIG-BIO dapat membantu dalam mengidentifikasi jenis statistik yang harus digunakan untuk menggambarkan urutan, jenis organisme yang berasal dari urutan, urutan dan urutan tugas dan mengarah pada pengelolaan cluster yang kuat dan terbuka [8].

III. METODOLOGI

Metodologi yang dilakukan menggunakan beberapa pendekatan yaitu pengumpulan data, filtrasi data, klusterisasi data, visualisasi data, serta analisis data seperti yang digambarkan pada Gambar 1.



Gambar 1. Metodologi penelitian

A. *Pengumpulan Data*

Sumber data dari penelitian ini adalah artikel ilmiah yang dipublikasikan pada sejumlah jurnal ilmiah bereputasi antara lain *sciencedirect, biomed central, link springer* dan IEEE. Setelah dilakukan analisis terhadap distribusi publikasi ilmiah di setiap jurnal tersebut, diperoleh bahwa sebagian besar peneliti bidang bioinformatika mempublikasikan hasil risetnya di IEEE.

Selanjutnya kami melakukan kajian pustaka dengan mencari referensi artikel yang bertujuan untuk memahami penggunaan teknik HPC dalam bidang bioinformatika. Berdasarkan studi literatur yang telah dilakukan, kami mendapatkan beberapa *keyword* terkait dengan teknik HPC yang banyak digunakan di bidang bioinformatika. Tahap pencarian informasi ini menggunakan beberapa *keyword*, yaitu "*Multicore Bioinformatics", "Multiprocessor for Bioinformatics", "Clustering Bioinformatics",* dan *"GPU for Bioinformatics"*. Hasil pencarian menunjukkan bahwa terdapat 4831 artikel yang sudah dipublikasikan peneliti di bidang bioinformatika.

B. *Filtrasi Data*

Setelah pengumpulan data selesai dilakukan, data yang diperoleh dispesifikasi berdasarkan beberapa kriteria. Data akan difiltrasi berdasarkan tahun publikasi *paper* yaitu 2012 hingga 2016, judul, abstrak, dan *keyword* di artikel.

Proses filtrasi ini dilakukan melalui *focus group discussion*, yang terbagi menjadi dua kelompok. Tiap kelompok membahas dua teknik HPC per tahun dari tahun 2012 sampai 2016. Kelompok pertama membahas teknik GPU dan *Multicore* berdasarkan keyword "*Multicore Bioinformatics"* dan *"GPU for Bioinformatics"*. Kelompok kedua membahas teknik *Clustering* dan *Multiprocessor* berdasarkan keyword *"Multiprocessor for Bioinformatics"* dan *"Clustering Bioinformatics"*. Setiap kelompok melakukan proses yang sama, yaitu membaca sejumlah artikel dengan menggunakan teknik *skimming* yang berfokus pada bagian judul, *keyword* dan abstrak dari artikel. Hasil dari setiap kelompok selanjutnya dikombinasikan untuk mendapatkan hasil akhir.

C. *Klusterisasi Data*

Data yang telah difilter selanjutnya dikelompokkan berdasarkan tahun dan jenis teknik HPC. Terdapat beberapa teknik yang digunakan HPC dalam bioinformatika, yaitu

a. *Multicore*

Multicore adalah komponen komputasi tunggal yang memiliki dua atau lebih *core* independen atau unit pemrosesan. *Core* ini adalah yang membaca dan menjalankan instruksi program pada *Central Processing Unit* (CPU). Instruksi ini pada dasarnya adalah instruksi CPU biasa seperti menambahkan dan memindahkan data . *Multicore* dapat melakukan banyak instruksi secara bersamaan. Program berjalan dengan membagi instruksi menjadi bagian yang lebih kecil untuk dijalankan secara bersamaan pada *multicore* [9].

b. *Multiprocessor*

Multiprocessor memiliki dua atau lebih unit pemrosesan. CPU diintegrasikan ke dalam satu sistem komputer. Pada dasarnya *multiprocessore* memiliki dua atau lebih CPU pada sistem secara fisik. Namun, *processor* memerlukan dukungan sebuah sistem untuk mengeksekusi instruksi program, sementara *processor* lainnya melakukan instruksi program yang berbeda secara bersamaan [9].

c. *Clustering*

Clustering adalah unit logis tunggal yang terdiri dari beberapa komputer yang terhubung melalui LAN. Komputer berjejaring pada dasarnya bertindak sebagai mesin tunggal yang jauh lebih cepat, kapasitas penyimpanan yang lebih besar, dan integritas data yang lebih baik [10].

d. GPU

Graphics Processing Unit (GPU) adalah prosesor yang bertugas secara khusus untuk mengolah tampilan grafik. Model komputasi GPU adalah menggunakan *Central Processing Unit* (CPU) dan GPU bersama – sama dalam model komputasi *co-processing* yang heterogen. Bagian berurutan dari aplikasi berjalan pada CPU dan bagian komputasi intensif dipercepat oleh GPU. Aplikasi berjalan lebih cepat karena menggunakan kinerja GPU yang tinggi untuk meningkatkan kinerja [11].

D. *Visualisasi data*

Data yang sudah diklusterisasi kemudian divisualisasikan dalam bentuk grafik. Visualisasi data akan memudahkan dalam membaca dan menganalisis banyak data.

E. *Analisis Data*

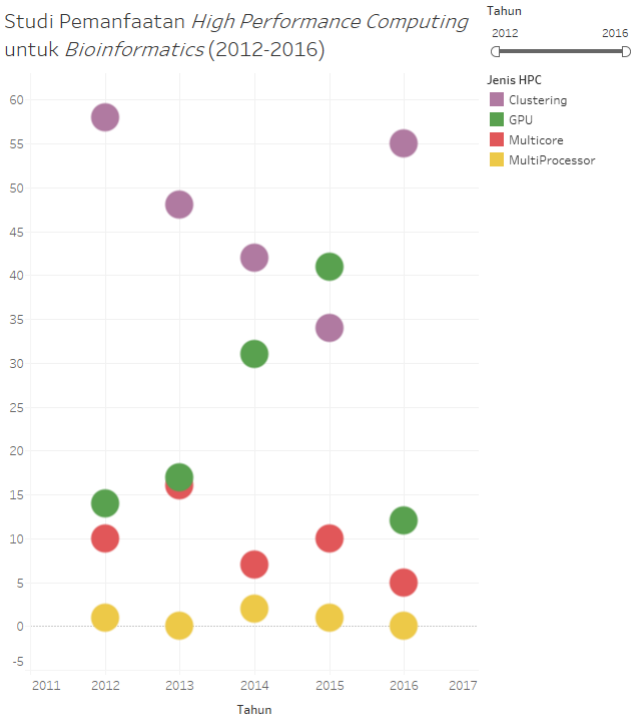
Pada tahap ini, fokus analisis kami adalah melihat teknik yang sering digunakan dalam bidang bioinformatika, berdasarkan penelitian tiap tahun ataupun dari keseluruhan tahun. Kemudian menganalisis *paper* mengenai penggunaan teknik HPC dalam bidang bioinformatika, serta menganalisis kelebihan dan kekurangan dari masing-masing teknik HPC.

IV. HASIL DAN PEMBAHASAN

Berdasarkan data yang diperoleh dari *sciencedirect, biomed central, link springer* dan IEEE pada tahap pengumpulan data, selanjutnya dilakukan proses filtrasi berdasarkan tahun publikasi yaitu dalam rentang tahun 2012 sampai dengan 2016. Distribusi artikel bidang bioinformatika dari setiap tahunnya dapat dilihat pada Tabel 1. Hasil distribusi artikel dibawah ini kami peroleh dari *focus group discussion.* Visualisasi pemanfaatan tiap teknik HPC dalam bidang bioinformatika setiap tahunnya dapat dilihat pada Gambar 2.

TABEL I. Jumlah artikel HPC untuk bioinformatika di IEEE

Tahun	GPU	*Multi-core*	*Multi-Processor*	*Clustering*
2012	14	10	1	58
2013	16	13	0	48
2014	29	6	2	42
2015	40	10	1	34
2016	12	4	0	55
Total	111	43	4	237



Gambar 2. Perkembangan HPC dalam bioinformatika

Berdasarkan pada Gambar 2, dapat diketahui bahwa teknik HPC yang paling banyak digunakan dalam penerapan bioinformatika dari tahun 2012 sampai 2016 adalah:

1. Teknik *Clustering*, pada tahun 2012 sampai 2016 paling banyak digunakan dibandingkan dengan tiga teknik lainnya, kecuali pada tahun 2015 berada diurutan kedua.
2. Teknik GPU, pada tahun 2012 sampai 2016 berada di bawah satu peringkat dengan teknik *Clustering*, kecuali pada tahun 2015 berada diurutan pertama.
3. Teknik *Multicore*, paling banyak digunakan pada tahun 2013. Tetapi untuk setiap tahunnya teknik ini tetap berada di peringkat ketiga.
4. Teknik *Multiprocessor*, merupakan teknik yang setiap tahunnya tidak mengalami perubahan yang signifikan.

Berdasarkan hasil yang diperoleh, teknik *Clustering* merupakan teknik yang paling banyak digunakan untuk bidang bioinformatika berdasarkan kurun waktu lima tahun terakhir. Sebaliknya, teknik *Multiprocessor* merupakan teknik yang paling sedikit digunakan untuk bidang bioinformatika. Teknik clustering banyak digunakan karena merupakan teknik yang hemat biaya, memiliki kecepatan yang tinggi, serta memiliki perangkat lunak yang terdistribusi dengan baik, sehingga bisa menjalankan tugas bioinformatika dengan efisien. Sebaliknya, proses multiprocessor memiliki beberapa hambatan misalnya lambatnya dalam mengeksekusi tugas padahal bidang bioinformatika memiliki data berukuran yang sangat besar sampai dengan 64GB. Penulis belum menemukan adanya penelitian atau referensi yang membandingkan tren penggunaan teknik HPC untuk bioinformatika setiap tahunnya.

Berdasarkan data yang dikumpulkan, proses bioinformatika yang paling banyak diteliti menggunakan HPC adalah filogenetik, DNA *sequence* , protein, kromatografi, *preteomics*, *genomics sequence*, RNA, oligonukleotida, neurophysiology, *molecular biophysics*, *cell biophysics*, dan *gene expression.* Perbandingan data lengkap tiap distribusinya dapat dilihat pada Tabel 2.

Penggunaan teknik HPC dalam bioinformatika memiliki kelebihan dan kekurangan dalam menyelesaikan tugas. Kekurangan dan kelebihan tiap teknik HPC telah dirangkum pada Tabel 3.

V. KESIMPULAN

Pada penelitian ini, beberapa kesimpulan yang dapat diambil adalah sebagai berikut.

1. Beberapa teknik HPC yang sering digunakan pada bidang bioinformatika adalah *Graphics Processing Unit* (GPU), *Multi-Core*, *Multi-processor* dan *Cluster Computing*.
2. Berdasarkan data yang diperoleh, teknik yang paling banyak digunakan untuk bioinformatika adalah teknik *Clustering*. Tetapi teknik yang paling efisien dalam menyelesaikan kasus bioinformatika adalah *Graphics Processing Unit* (GPU).
3. Proses dalam bidang bioinformatika yang paling banyak memanfaatkan teknik HPC adalah filogenetik, DNA *sequence* , protein, kromatografi, *preteomics*, *genomics sequence*, RNA, oligonukleotida, neurophysiology, *molecular biophysics*, *cell biophysics*, dan *gene expression.*

TABEL II. Distribusi proses bioinformatika – HPC (2012-2016)

Bioinformatika	*Clustering*	GPU	*Multicore*	*Multi Processor*
Genomic	43	23	17	2
DNA Sequence	23	25	6	0
Gen Sequence	6	7	3	0
Filogenetik	46	4	5	1
Protein	58	15	5	0
Amino Acid Sequence	0	0	2	0
Biological System Modeling	0	6	2	0
Proteomics	6	1	1	1
RNA	18	1	1	0
Thermostat	0	1	1	0
Neural Network	1	3	0	0
Gen Expression	25	1	0	0
Entropy	1	0	0	0
Molecular Biophysic	8	2	0	0
Kromatografi	2	0	0	0
Biomedical MRI	0	5	0	0
Biological Sequence	0	14	0	0
Hidden Markov Models	0	3	0	0
Total	237	111	43	4

TABEL III. Kelebihan dan Kekurangan teknik HPC untuk Bioinformatika

HPC	Kelebihan	Kekurangan
Clustering	Tidak memerlukan perubahan yang besar pada sumber *code* program CPU, kecuali perubahan untuk *message passing* [12].	Menggunakan banyak energi dan membutuhkan perawatan [12].
GPU	Memiliki jumlah unit program yang berkomputasi tinggi, sehingga memungkinkan bisa melakukan eksekusi ribuan simultan threads. Tersedia dari performansi tinggi memori lokal [12].	Berdasarkan paradigma komputasi, perubahan SIMD: konsional cabang menyiratkan serialisasi thread. Arsitektur GPU membutuhkan penulisan ulang *code* dan mendesain ulang algoritma[12].
Multi processor	Waktu eksekusi cepat pada multiple programs [9].	Memiliki banyak kendala, membutuhkan konfigurasi yang kompleks [9].
Multicore	Menggunakan sumber daya yang lebih kecil, desain *multicore* memiliki desain *error* yang lebih rendah [13].	Membutuhkan penyesuaian pada *software* untuk memaksimalkan kegunaan dari sumber daya komputasi *multicore.* Pengembangan *chip multicore* yang menurun karena semakin sulit mengatur suhu pada *chip* yang padat [13].

DAFTAR PUSTAKA

[1] S. Konur, M. Kiran, M. Gheorghe, M. Burkitt dan F. Ipate, "Agent-based High-Performance Simulation of Biological System on the GPU," IEEE 17th International Conference on High Performance Computing and Communications (HPCC), 2015.

[2] M. Burkitt, D. Walker, D. M. Romano dan A. Fazeli, "Constructing Complex 3D Biolgical Environments from Medical Imaging Using High Performance Computing," IEEE/Acm Transactions on Computational Biology and Bioinformatics, Vol. 9, NO. 3, May/June 2012, 2012.

[3] L. Domanski, T. Bednarz, T. E. Gureyev, L. Murray, E. Huang dan J. A. Taylor, "Application of Heterogeneous Computing in Computational and Simulation Science," Fourth IEEE International Conference on Utility and Cloud Computing, 2011.

[4] D. Johnson, B. Shafer, J. J. Lee dan J. Y. Chen, "Multi-Biomarker Panel Selection on a GPU," IEEE International Conference on Purdue University School of Engineering and Technology Indiana University Purdue University Indianapolis, IN, USA, 2012.

[5] A. M. Hosny, H.A. shedeed, A. S. Hussein dan M. F. Tolba, "An Efficient Solution for Aligning Huge DNA," Computer Engineering & Systems (ICCES), 2011 International Conference, 2011.

[6] A. F. Leite dan A. C. M. A. de Melo, " Executing a Biological Sequence Comparison Application on a Federated Cloud Environment," High Performance Computing (HiPC), 2012 19th International Conference, 2012.

[7] S. Santander-Jiménez, M. A. Vega-Rodríguez, J. A. Gómez-Pulido dan J. M. Sánchez-Pérez, "Evaluating the Performance of a Parallel Multiobjective Artificial Bee Colony Algorithm for Inferring Phylogenies on Multicore Architectures," 2012 10th IEEE International Symposium on Parallel and Distributed Processing with Applications, 2012.

[8] R. A. Seoud, M. A. Mahmoud dan E. Eldin, "BIG-BIO: - Big Data Hadoop-based Analytic Cluster Framework for Bioinformatics," Informatics, Health & Technology (ICIHT), International Conference, 2017.

[9] Theydiffer. (08 Juni 2017). *Difference between Multicore and Multiprocessor System.* Diakses melalui : http://theydiffer.com/difference-between-multicore-and-multiprocessor-systems/

[10] Techopedia. (08 Juni 2017). *Computer Cluster*. D iakses melalui: https://www.techopedia.com/definition/6581/computer-cluster

[11] Super Computing Applications and Innovation (SCAI). 08 Juni 2017. *General Purpose Graphics Processing Unit* (GPGPU).Diakses melalui: http://www.hpc.cineca.it/content/gpgpu-general-purpose-graphics-processing-unit

[12] M. S. Nobile, P. Cazzaniga, A. Tangherloni and D. Besozzi, "Graphics processing units in bioinformatics, computational biology and systems biology", Briefings in Bioinformatics, 2016.

[13] D. Parkesit. (08 Juni 2017). *Multiprocessor*. Diakses melalui: https://www.academia.edu/7547799/Multiprocessor