ZettaVEGA: ZettaScaler Verifying Environment for Genome Analysis

Accelerating genomic data analysis with PEZY-SC3

概要

ZettaVEGAは、PEZY-SC3を四基搭載したZettaScaler-3.0システム上で動作する、PEZY Computing社製の高速なゲノム解析ソフトウェアです。ZettaScaler-3.0は、ストレージ、ネットワーク、演算性能といったゲノム解析に必要なITリソースを、コンパクトでスケーラブルなソリューションに統合しています。
ZettaVEGAは、四基のPEZY-SC3を使用することで、33xカバレッジのヒト全ゲノムシークエンスのデータを最大96検体/日処理することができます。ZettaVEGAは、デファクトスタンダードとなっているGATK Best Practiceパイプラインと比較して100倍以上高速に動作しますが結果の互換性は99.99%以上保証しています。ZettaVEGAが使用しているソフトウェアはGATK Bset Practiceパイプラインを構成するソフトウェアと互換になるように設計・開発されています。ZettaVEGAのFastqを入力しVCFを得るためのパイプラインはシンプルなインターフェースを備えており、ユーザーはZettaVEGAの各ソフトウェアを従来のソフトウェアと同様の感覚で使用することが可能なため、従来のゲノム解析パイプラインからZettaVEGAへのスムーズな移行が可能です。またZettaVEGAでは、GATK4.2で導入されたFRDやBQDなどの高精度な確率モデルや、アライメントにおけるALT-contigからのLiftover機能や、チューニングされたパラメータのプリセットを利用することができ、高感度と低い偽陽性を両立した結果を得ることが可能になります。

本ドキュメントでは、ZettaVEGAのアーキテクチャとパフォーマンス・精度について解説します。

May 2023

目次

イントロダクション

2003年にヒトゲノムの解読が完了して以降、ヒトゲノムを解析しようとする試みは20年にわたり行われてきました。近年ではゲノムシーケンシングのコストがムーアの法則を凌ぐ勢いで下がり、個人の全ヒトゲノムシーケンシングにかかるコストが$1000以下となってきています。それに伴い、プレシジョン・メディシンという、遺伝情報、個々人の生活環境やライフスタイルの差異を考慮し、疾病予防や治療を行う、新しい医療の考え方も現実的なものとなってきました。また、次世代シーケンサによるゲノムシーケンシングの速度も高速化の一途を辿っており、高速なゲノム解析エンジンの重要性は非常に高まっています。

そこで、PEZY Computingでは、PEZY-SC3アクセラレータを使用し、高速なゲノム解析エンジンである ZettaVEGAを開発しました。

ZettaVEGA

ZettaVEGAは、ZettaScaler-3.0システム上で動作する、高速なゲノム解析ソフトウェアです。

ZettaScaler-3.0

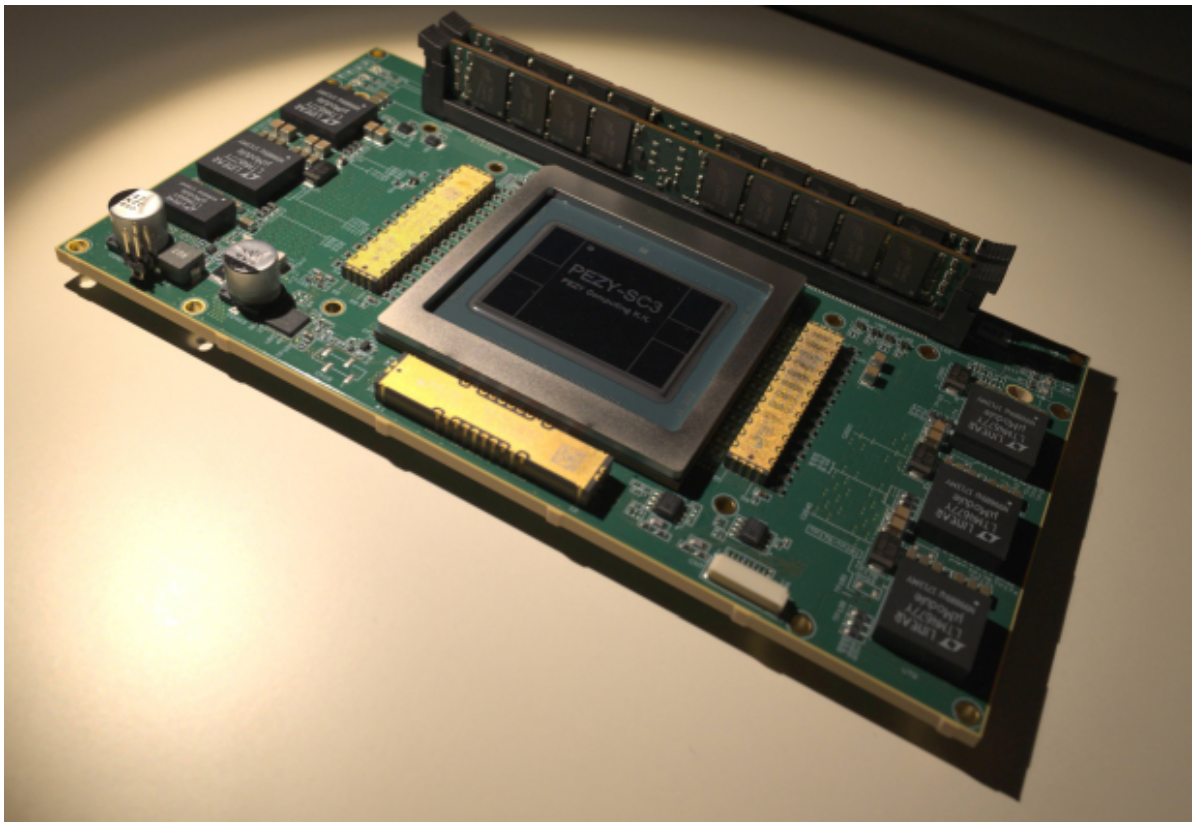
ZettaScaler-3.0は、PEZY Computingが開発したサーバーです。最大の特徴として、独自開発した汎用アクセラレータ、PEZY-SC3を四基搭載していることが挙げられます。PEZY-SC3の理論ピーク性能は倍精度で19.66TFlopsであり、システム全体としては78.64TFlopsとなります。また、ホストCPUにはEPYC-7713Pを採用しており、PEZY-SC3で高速化されていない従来のツールを使用する場合にも高速に処理を行うことができます。外部ネットワークインタフェースは10G EthernetまたはInfiniband EDRをサポートし、ネットワーク上のファイルサーバーなどとの高速なネットワーク通信も可能です。ZettaScaler-3.0では、ゲノム解析に必要な、リファレンスデータファイルやfastqファイル、またvcfデータなどの大容量ファイルを高速に読み書きするために、四基のPCI Express接続のNVMeストレージを搭載し、これらをRAID0またはRAID01で構成しています。これにより、ゲノム解析で必要になる大容量データを高速に読み書きすることが可能です。

以下にZettaScaler-3.0の諸元を示します。

	Spec
CPU	EPYC-7713P
RAM	1TB(or 2TB)
SSD	256GB(OS, SATA)
	8TB(Data, NVMe)
Network	10G Ether
	Infiniband EDR(100Gbps)
Accelerator	PEZY-SC3 * 4

PEZY-SC3

PEZY-SC3はPEZY Computingが開発した汎用アクセラレータカードです。PEZY-SCシリーズとしては第三世代の製品となります。高いピーク性能と面積効率・電力効率を両立するようにデザインされており、非常に高い演算性能と非常に高い電力効率を誇ります。理論ピーク性能は倍精度で19.66TFlops, 単精度で39.32TFlops, 半精度で78.64TFlopsとなります。また、メモリとしてHBM2を採用しており、1.2TB/sと非常に高速なメモリ帯域を持ちます。ホストとのインタフェースはPCIExpress Gen4 x16を採用し、双方向32GB/sの帯域を持ちます。消費電力はモジュールで600Wとなっています。

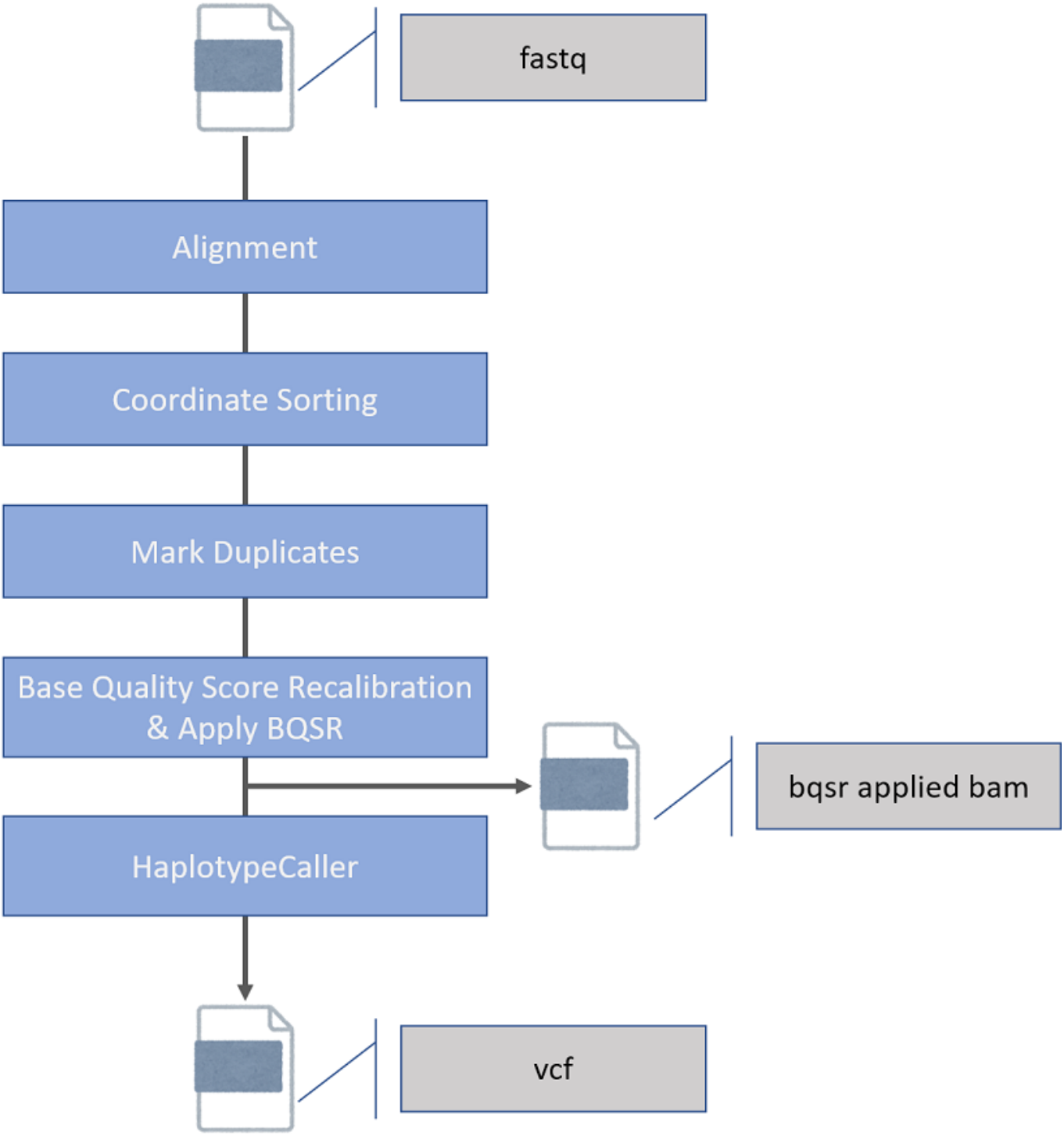


以下にPEZY-SC3の諸元を示します。

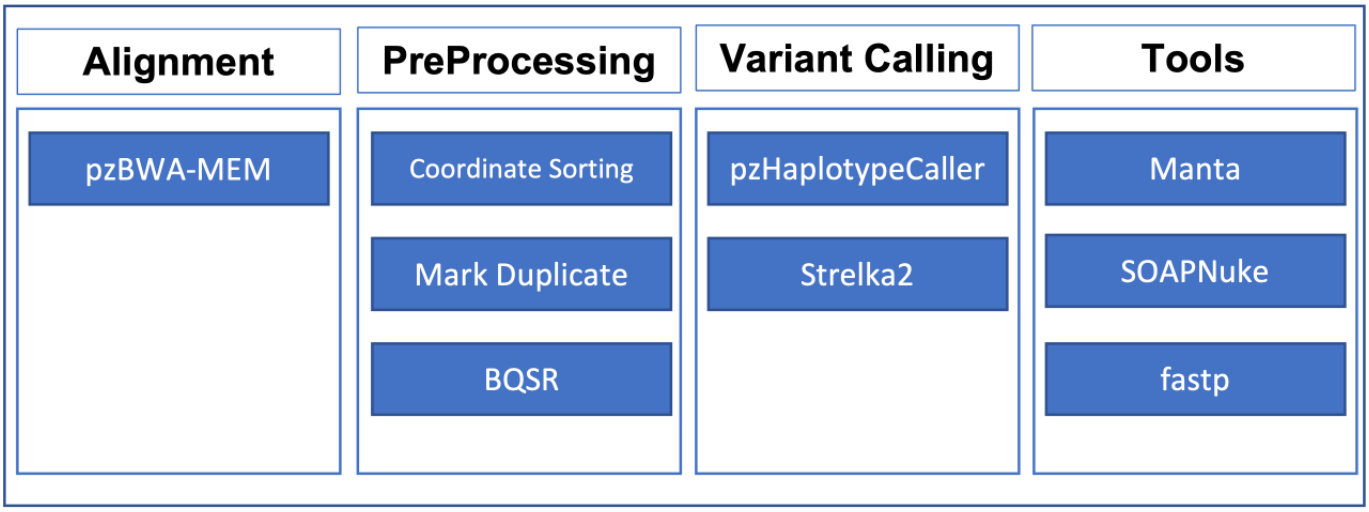
	PEZY-SC3
Number of PEs	4096
Clock(PE / GHz)	1.2
Clock(BUS / GHz)	1
Peak FP64 Flops(TFlops)	19.66
Peak FP32 Flops(TFlops)	39.32
Peak FP16 Flops (TFlops)	78.64
Memory Type	HBM2
Memory Size(GB)	32
Peak Memory bandwidth(GB/s)	1200
Power Consumption	600W

ZettaVEGA ソフトウェア概要

ZettaVEGAでは、以下のヒトゲノム解析パイプラインを提供しています。



また、ZettaVEGAを構成しているソフトウェアは以下の通りです。



以下に、主要なソフトウェアを解説します。

pzBWA-MEM

pzBWA-MEM は、BWA-MEM version 0.7.17 (r1198) をベースに、PEZY Computingによる改良を加えた、高速なアライメントソフトウェアです。改良点は以下のとおりです。

- PEZY-SC3 によるアライメント処理の高速化
- パイプライン段数とパイプライン構造の最適化による処理の高速化
- Fastq 読み込みの最適化による高速なクエリデータの読み込み
- スコアを調整するためのオプション等の追加
- Alternate contig へのアライメントのリフトオーバー機能
- ヒト集団に対するロングリードでのバリアントコール結果とリフトオーバー機能を活用したアライメント感度の向上

reshz

reshz は、pzBWA-memの出力したアライメントデータに対して、Coordinate Sorting, Picard MarkDuplicates, GATK BQSR, Applying BQSRを行うツールです。 ZettaScaler3.0の大容量メモリを活用することで高速なデータ処理を行います。

pzHaplotypeCaller

pzHaplotypeCallerは、GATK4.2.0.0のHaplotypeCallerをベースに、PEZY Computingで高速化・高精度化を行ったゲノム変異解析用ソフトウェアです。 改良点は以下の通りです。

- GATK4.2.0.0 HaplotypeCallerをもとにC++でフルスクラッチ
- PEZY-SC3を使用したSmith-WatermanアライメントとPairHMM処理の高速化
- CPU処理の最適化による高速化
- GATK4.2 で実装されたFRD, BQDなどの確率モデルがオプションで使用可能

速度と精度評価

本セクションでは、ZettaVEGAの速度と精度を評価します。

使用したデータは以下の通りです。

Run Name	Coverage	Total Gbp	Type	URL
CNR0028192 (HG001)	x46.1	138.329	PCR	https://db.cngb.org/search/run/CNR0028192/
CNR0028194 (HG001)	x42.1	126.174	PCR-Free	https://db.cngb.org/search/run/CNR0028194/
PrecisionFDA challenge V1 HG001	x53.6	160.700	PCR-Free	https://console.cloud.google.com/storage/browser/genomics-public-data/precision-fda/input
PrecisionFDA challenge V2 HG002	x41.8	125.356	PCR-Free	https://precision.fda.gov/challenges/10

これらのデータをseqkitのsampleコマンドを用いて100Gbpとなるようにsub-samplingを行い速度と精度の評価に用いました。

リファレンスは以下を用いました。

- hg19 gatk bundle + decoy (https://console.cloud.google.com/storage/browser/gcp-public-data--broad-references/hg19/v0 https://console.cloud.google.com/storage/browser/gatk-legacy-bundles/b37)
- hg38 gatk bundle (https://console.cloud.google.com/storage/browser/genomics-public-data/resources/broad/hg38/v0/) + population-contig

精度評価には、正解データとして v4.2.1 benchmarkを、精度評価ソフトウェアとしてrtg tools v3.12.1のvcfevalコマンドを使用しています。

hg19 + decoy

hg19 + decoyリファレンスを使用した際の速度を以下の表に示します。(単位は秒)

Run Name	BWA time	SHZ time	HaplotypeCaller time	ALL time
CNR0028192	532.73	123.28	242.06	781.65
CNR0028194	421.81	114.92	211.55	748.30
Precision FDA challenge V1 HG001	537.42	123.60	213.73	874.79
Precision FDA challenge V2 HG002	530.95	108.43	217.44	856.84

また、このときのそれぞれの精度は以下のようになります。

Summary

Name	precision	recall	f-measure
CNR0028192	0.9937	0.9923	0.9930
CNR0028194	0.9971	0.9924	0.9947
Precision FDA challenge V1 HG001	0.9897	0.9926	0.9911
Precision FDA challenge V2 HG002	0.9964	0.9910	0.9937

SNP

Name	precision	recall	f-measure
CNR0028192	0.9945	0.9939	0.9942
CNR0028194	0.9972	0.9924	0.9948
Precision FDA challenge V1 HG001	0.9900	0.9934	0.9917
Precision FDA challenge V2 HG002	0.9971	0.9913	0.9942

non-SNP

Name	precision	recall	f-measure
CNR0028192	0.9878	0.9806	0.9842
CNR0028194	0.9965	0.9925	0.9945
Precision FDA challenge V1 HG001	0.9874	0.9868	0.9871
Precision FDA challenge V2 HG002	0.9920	0.9893	0.9906

GRCh38 + population-contig

hg38 リファレンスをダウンロードし、アライメント困難な領域におけるヒト集団のロングリードでの解析結果をindexのビルド時に付与しています。 hg38 + population-contig リファレンスを使用した際の速度を以下の表に示します。

Run Name	BWA time	SHZ time	HaplotypeCaller time	ALL time
CNR0028192	978.3431	131.419	354.426	1464.2556
CNR0028194	1034.6179	120.1846	263.3924	1418.2293
Precision FDA challenge V1 HG001	1147.7322	114.6612	296.6762	1559.1012
Precision FDA challenge V2 HG002	1077.6147	117.178	271.4726	1466.2897

また、このときのそれぞれの精度は以下のようになります。

Summary

Name	precision	recall	f-measure
CNR0028192	0.9952	0.9953	0.9952
CNR0028194	0.9977	0.9963	0.9970
Precision FDA challenge V1 HG001	0.9912	0.9960	0.9936
Precision FDA challenge V2 HG002	0.9968	0.9946	0.9957

SNP

Name	precision	recall	f-measure
CNR0028192	0.9962	0.9972	0.9967
CNR0028194	0.9979	0.9965	0.9972
Precision FDA challenge V1 HG001	0.9915	0.9970	0.9943
Precision FDA challenge V2 HG002	0.9975	0.9951	0.9963

non-SNP

Name	precision	recall	f-measure
CNR0028192	0.9881	0.9825	0.9853
CNR0028194	0.9969	0.9943	0.9956
Precision FDA challenge V1 HG001	0.9884	0.9884	0.9884
Precision FDA challenge V2 HG002	0.9925	0.9908	0.9917

アライメント速度は低下してしまいますが、アライメントの精度と感度ともに向上します。

まとめ

本ドキュメントでは、高速なゲノム解析ソフトウェア ZettaVEGAを解説しました。PEZY-SC3を搭載したZettaScaler-3.0サーバーでは、33xのヒトゲノム解析を最大で96検体/日処理することが可能です。また、GATK best practice guideに準拠したソフトウェア群を提供しているため、精度はGATKに準ずるかまたは向上しており、ユーザーのワークロードも、GATK best practice guideからZettaVEGAにシームレスに移行が可能です。

Appendix

hg19 + decoyの作り方

hg19リファレンスを以下のURLよりダウンロードします。
https://console.cloud.google.com/storage/browser/gcp-public-data--broad-references/hg19/v0

以下のコマンドで 1 - 22, X, Y, MT の配列とそれ以外を分離して、indexのビルド時にデコイとしてそれ以外の配列を付与しています。この機能はZettaVEGA独自の機能です。

```
seqkit grep -n -r -p '^(X|Y|MT|[1-9])' Homo_sapiens_assembly19.fasta -o
Homo_sapiens_assembly19_add_hs37d5_decoy.fasta
seqkit grep -n -r -v -p '^(X|Y|MT|[1-9])' Homo_sapiens_assembly19.fasta -o
hs37d5_decoy.fasta
/opt/pezy/pzwgs/bin/bwa-index -i 8 -d hs37d5_decoy.fa
Homo_sapiens_assembly19_add_hs37d5_decoy.fasta
```