AMD EPYC

WHITE PAPER

# 4TH GEN AMD EPYC™ PROCESSORS ACCELERATE GENOMICS DATA PROCESSING WITH SENTIEON®

AMD
together we advance_data center computing

*First Edition*
*March, 2024*

*AMD: Harini Malik & Michael Seniziaz*
*Sentieon: Don Freed & Brendan Gallagher*

# 4TH GEN AMD EPYC™ PROCESSORS ACCELERATE GENOMICS DATA PROCESSING WITH SENTIEON®

## CONTENTS


INTRODUCTION - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 3

TEST SETUP - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 4

- Benchmark Datasets ..... 4
- Hardware Configuration ..... 4
- Software Configuration ..... 4

GERMLINE ALIGNMENT AND VARIANT CALLING FOR SHORT READS 5

GERMLINE ALIGNMENT AND VARIANT CALLING FOR LONG READS 6

CONCLUSION - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 7

FOOTNOTES - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 7

# INTRODUCTION

A growing number of applications use next-generation sequencing, such as liquid biopsy, genetic tests for hereditary cancer or clinical diagnostics, and academic and industrial research projects. Lower costs and new applications are spurring a dramatic increase in the amount of sequence data produced each year, with an estimated 2-40 exabytes of storage needed by 2025*. This abundance of sequence data is driving demand for fast and efficient genomic data processing solutions. This white paper showcases processing a variety of sequence datasets using Sentieon®* running both on premises and in the cloud on systems powered by 4th Gen AMD EPYC™ processors.

Element Biosciences and Ultima Genomics are joining Illumina and MGI/Complete Genomics in marketing high-throughput short read instruments. The PacBio and Recent ONT instruments have substantially improved throughput and data quality over the past few years. These factors make high-accuracy long read sequencing a compelling alternative to short read sequencing for some applications*, especially where structural variant calling, or phasing information is beneficial. This competition* in the NGS market is pushing down the cost of NGS data generation, reducing the cost of a human genome sequenced to 30x coverage (a common level of coverage for a clinical whole genome) to well under $1000 per sample.

Sentieon creates a fast, efficient, easily deployable, and robust suite of software for processing genomic data using x86 or Arm processors. Sentieon provides a suite of tools for a broad range of genomics data processing needs, including tools for DNA and RNA alignment, preprocessing, germline and somatic variant calling, handling of molecular barcodes, and joint calling of cohorts. The tools are highly optimized for x86 and Arm processors and are multi-threaded by default, allowing researchers and clinical teams to process their data without needing to implement complex workflow logic or acquire specialized hardware.

Numerous open-source tools also exist for processing genomic data with x86 CPUs. Many of these tools are either single-threaded or limited to only a few threads effectively, which may impact users' ability to effectively leverage the computing power available on x86 systems with high processor core counts. For example, the popular BWA-GATK pipeline uses GATK HaplotypeCaller for germline variant calling* but cannot effective use more than four threads, which reduces efficiency on large machines*.

This situation spawned the misplaced notion that achieving reasonable genomic data processing runtimes requires hardware acceleration despite the fact that many genomic algorithms use complex, multi-step data manipulation well suited for modern x86 CPUs.

By contrast, Sentieon software exhibits good scalability at high thread counts*. This white paper showcases processing a variety of sequence datasets using Sentieon running both on premises and in the cloud on systems powered by 4th Gen AMD EPYC processors.

The test results presented in this white paper provide detailed benchmark results across key genomic analysis stages: alignment, preprocessing, and germline variant calling. This testing reveals insights into the pipeline's performance, efficiency, and cost-effectiveness. Both configurations process a 30x whole genome in less than 15 minutes at a cost of under $1.50 when using cloud instances –12x cheaper than the comparable NVIDIA Parabricks pipeline on an AWS instance with NVIDIA H100 GPUs. The AMD cloud instance was also ~1.5 minutes faster than the NVIIDA Parabricks outline. (See Figures 1 and 2.)

.

# TEST SETUP

This section describes the on-premises and cloud instance configurations used for the tests showcased in this white paper.

## BENCHMARK DATASETS

The benchmarks described in this white paper use the GIAB HG002 sample sequenced on the Illumina NovaSeq, PacBio Sequel II (HiFi), and Oxford Nanopore instruments in the Fastq file format:

- **Illumina NovaSeq:** These Fastq files were obtained from sequencing the Google Genome in a Bottle* samples and then downsampled to 30x coverage (93 billion bases).
- **PacBio Sequel II:** These Fastq files the PacBio Sequel II instrument were obtained from the Genome in a Bottle data repository* and then downsampled to 30x coverage (93 billion bases).
- **Oxford Nanopore:** These tests used Fastq files 1-11 from the Human Pangenome Reference Consortium data repository*. These files were selected to obtain a coverage of 29.6x (91.8 billion bases.

## HARDWARE CONFIGURATION

The Sentieon software can process data across a variety of hardware configurations. This section describes the on-premise and cloud configurations used for testing.

- **Amazon EC2:** The Amazon EC2 Hpc7a instance family* provides compute infrastructure suitable for high-performance computing applications at a cost-effective price point. These instances include gp3 EBS storage with 5000 IOPS and 500 MiB/s. These tests also used a ramdisk in the instance memory to hold intermediate files. Sentieon anticipates that the Amazon EC3 C7a instance family* would achieve similar results. Sentieon performed the cloud testing described in this white paper.
- **Oracle Cloud:** The Oracle Cloud BM.GPU.H100.8* shape is powered by dual 56-core 4th Gen Intel® Xeon® processors, 8 x NVIDIA H100 80GB Tensor Core GPUs, 2 TB of DDR5 memory, 16 x 3.84 TB NVMe local storage, and 8 x 400 Gb/sec cluster networking.
- **On-premises:** Testing used a server configured as shown in Table 1. AMD performed the on-premises testing described in this white paper.

| Item | Description |
|---|---|
| Processors | 2 x AMD EPYC 9654 |
| Cores \| Threads (per Socket) | 96 \| 192 |
| Memory | 1.5 TB (24 x 64 GB)<br>Dual Rank DDR5-4800, 1 DPC |
| Storage OS \| Data | 1 x 256 GB SATA \|<br>1 x 2 TB NVMe |
| BIOS Version | 1007D |
| BIOS Settings | SMT=off, Determinism=Power, NPS=1 |
| OS | RHEL 8.7 |
| OS Settings | amd_iommu=on, iommu=pt, mitigations=off |

*Table 1: On-premises system configuration*

## SOFTWARE CONFIGURATION

- Sentieon software:
  - **Cloud:** v202308.01
  - **On-premises:** v202308
  - **CLI:** The sentieon-cli at commit 0a79561 was used for germline variant calling of the long-read datasets.
- DNAscope model bundles used during data processing:
  - **Illumina WGS:** v2.0
  - **PacBio HiFi:** v2.0
  - **ONT:** v2.0

# GERMLINE ALIGNMENT AND VARIANT CALLING FOR SHORT READS

Sentieon software processed the entire Illumina NovaSeq whole-genome 30x dataset in 12.7 minutes[1] during on-premise testing and in 12.3 minutes during cloud testing.[2] These runtimes are about 12% faster and 12x cheaper[3] than the previously published NVIDIA Parabricks pipeline runtimes* running with eight NVIDIA H100 GPUs.

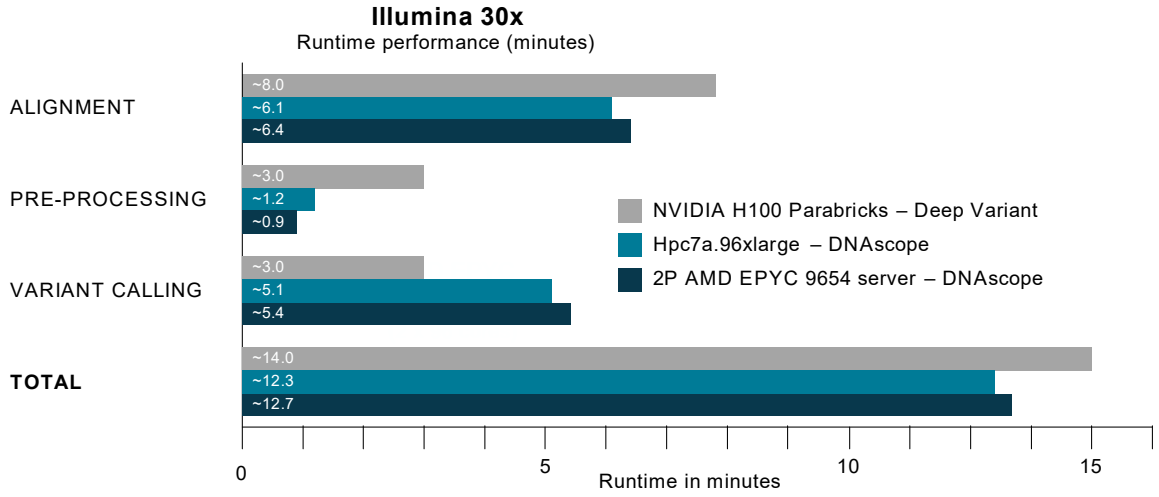


*Figure 1: Sentieon DNAscope and NVIDIA runtime performance*

Figure 1 shows the on-premises and cloud runtime performance of the Sentieon DNAscope and NVIDIA Parabricks pipelines.NVIDIA reports the NVIDIA Parabricks numbers using a configuration with 8 x NVIDIA H100 GPUs. The faster cloud runtime may be caused by the different Sentieon software versions used in the cloud and on-premise configurations.

Users requiring sporadic or small-scale genomic data processing can use cloud instances as a flexible solution with no significant upfront hardware investment. As for cloud costs, Sentieon estimated the cost savings of using non-accelerated Amazon EC2 hpc7a.96xlarge instances versus the NVIDIA Parabricks pipeline to run the 30x Illumina whole genome using Sentieon software.

To do this, Sentieon used the reported runtime of the NVIDIA Parabricks pipeline with 8 x NVIDIA H100 GPUs, and the GPU cost of an Oracle Cloud BM.GPU.H100.8* shape. These calculations reveal that the Sentieon DNAscope pipeline running on Amazon EC2 Hpc7a instances is more than 12x cheaper per genome compared to the GPU-accelerated cloud instances. Essentially, a 2P server powered by 4th Gen AMD 9654 processors running Sentieon can process 11 more human genomes per day versus an Oracle Cloud BM.GPU.H100.8 instance with 8 Nvidia H100 GPUs running the HG002_NovaSeq_30x dataset.[4]

Figure 2 shows the cost per 30x Illumina whole genome for Sentieon on the hpc7a.96xlarge instance and NVIDIA Parabricks. The cost of the Nvidia Parabricks pipeline is estimated using the reported runtime of the Nvidia Parabricks pipeline with 8x H100 GPUs, and the current On-Demand cost of a p5.48xlarge instance or the GPU cost of the BM.GPU.H100 shape containing 8 H100 GPUs. The cost of the Sentieon DNAscope pipeline with the AMD hpc7a.96xlarge instance is calculated using the pipeline runtime and the current On-Demand price of the instance in us-east-2.

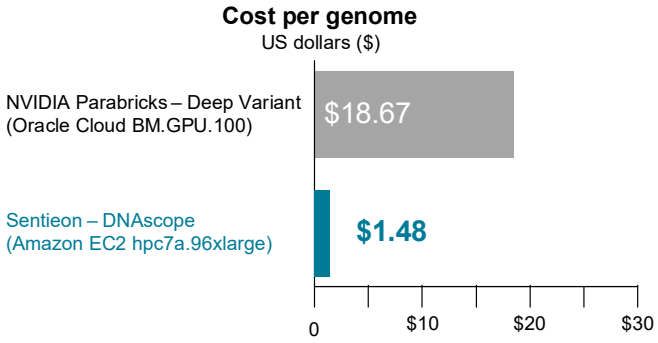


*Figure 2: Hpc7a.96xlarge costs for Sentieon vs. NVIDIA Parabricks*

## GERMLINE ALIGNMENT AND VARIANT CALLING FOR LONG READS

The Illumina NovaSeq instrument is commonly used for short-read data, but high quality long-read data is increasingly used in whole-genome sequencing or other applications that can benefit from the additional long-range information.

Sentieon processed high-quality PacBio HiFi and Oxford Nanopore Technologies (ONT) duplex sequencing datasets at approximately 30x coverage using the Sentieon DNAscope LongRead pipeline running on an Amazon EC2 hpc7a.96xlarge instance. Processing of the PacBio HiFi completed in 16.9 minutes[5], and the ONT processing completed in 39.7 minutes.[6] The cost per genome was $2.02 and $4.77 for the PacBio HiFi and ONT datasets, respectively.

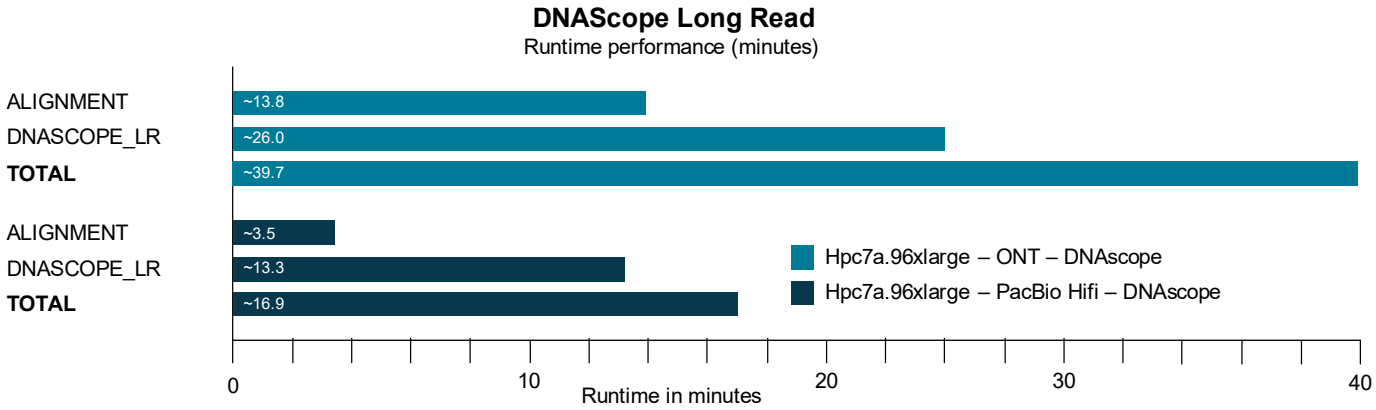


*Figure 3: DNAscope LongRead pipelines runtime at ~30x coverage*

Figure 4 shows the cost per genome for the DNAscope LongRead pipeline on the Amazon EC2 hpc7a.96xlarge instance for the PacBio HiFi and ONT datasets.

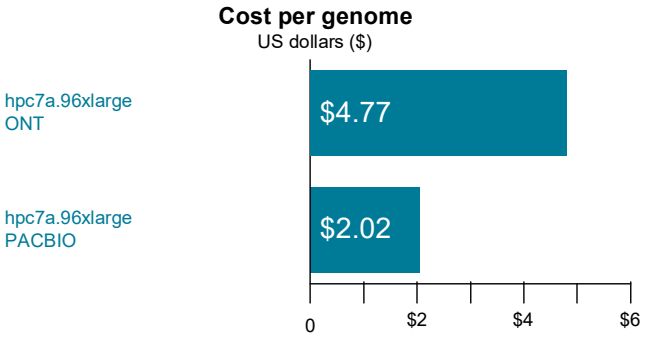


*Figure 4: Per-genome hpc7a.96xlarge DNAscope LongRead cost*

## CONCLUSION

Sentieon software running on servers or cloud instances powered by 4th Gen AMD EPYC processors is a highly performant, and cost-effective solution for efficiently processing genomic data processing. Sentieon software running on either on-premise servers or cloud instances powered by 4th Gen AMD EPYC processors delivers an estimated 12x cheaper cost per genome while offering 12% faster performance compared to Oracle Cloud BM.GPU.H100.8 instances. This powerful combination delivers the performance and cost effectiveness researchers and bioinformaticians need to effectively process data while containing costs. Please email info@sentieon.com* or visit Sentieon today to get started with Sentieon software on AMD. For questions about AMD, please contact your AMD sales representative.

## FOOTNOTES

1. SP5-168B: In AMD testing as of 11/13/2023, Sentieon® Release 202308 for read sorting on a system configured with 2P AMD EPYC 9654 (96 cores/socket, 192 cores/node); 1.5 TB (24x) Dual-Rank DDR5-4800 64GB DIMMs, 1DIMM per channel; 1 x 256 GB SATA (OS) | 1 x 2 TB NVMe (data); BIOS Version 1007D, SMT=off, Determinism=power, NPS=1; RHEL 8.7; OS settings: amd_iommu=on, iommu=pt, mitigations=off. Dataset used was GIAB HG002.novaseq.pcr-free.30x. Results may vary based on factors such as software version, hardware configurations and BIOS version and settings.
2. SP5C-025: Testing conducted by Sentieon as of 1/31/2024. Sentieon® Release 202308.01. 2P 96-core EPYC 9R14. System configurations: AWS hpc7a.96xlarge – 2P AMD EPYC 9R14 (96 cores/socket, 192 cores/node); 500 GB AWS gp3 (500 MB/s, 5000 IOPS). Amazon Linux® 2023 version 2023.3. Cloud performance results presented are based on the test date in the configuration. AMD has not independently verified the results. Results may vary due to changes to the underlying configuration, and other conditions such as the placement of the VM and its resources, optimizations by the cloud service provider, accessed cloud regions, co-tenants, and the types of other workloads exercised at the same time on the system.
3. SP5C-026: Testing conducted by Sentieon as of 1/31/2024. Sentieon® Release 202308.01. Nvidia ParaBricks processes an Illumina 30x genome in 14 minutes using a system with 8xH100 GPUs on Oracle Cloud (from Figure 2 of this white paper).

   Sentieon DNAscope pipeline processed an Illumina 30x genome in an average of 12.3 minutes using the Amazon hpc7a.96xlarge instance. 2P 96-core EPYC 9R14 System Configurations: AWS hpc7a.96xlarge - 2P AMD EPYC 9R14 (96 cores/socket, 192 cores/node); 500 GB AWS gp3 (500 MB/s, 5000 IOPS). Amazon Linux 2023 version 2023.3.

| | Cost/hr | Cost/genome | Multiplier |
|---|---|---|---|
| AWS 2P EPYC 9R14 hpc7a.96xlarge | $7.20 | $1.48 | |
| Oracle Nvidia 8xH100 BM.GPU.H100.8 | $80.00 | $18.67 | 12.64679 |

   Price comparison is based on AWS n-demand us-east-1 region pricing and Oracle pricing. Sources: https://aws.amazon.com/ec2/pricing/on-demand/ and https://www.oracle.com/cloud/price-list/.

   Cloud performance results presented are based on the test date in the configuration. AMD has not independently verified the results. Results may vary due to changes to the underlying configuration, and other conditions such as the placement of the VM and its resources, optimizations by the cloud service provider, accessed cloud regions, co-tenants, and the types of other workloads exercised at the same time on the system.
4. SP5-167B: AMD testing as of 11/13/2023, Sentieon® Release 202308 for read sorting. A 2P 96-core EPYC 9654 on an AMD reference platform vs. Oracle cloud tested NVIDIA Parabricks 4.2 with eight Nvidia H100 GPUS in a single server run. Published Oracle/Nvidia testing claims 14 minutes (102 Genomes/Day) to run a full sequence of the dataset HG002.novaseq.pcr-free.30x (end-to-end Parabricks germline workflow) on BM.GPU.H100.8 instances. https://developer.nvidia.com/blog/accelerate-genomic-analysis-for-any-sequencer-with-parabricks-v4-2/?ncid=so-nvsh-821697&dysig_tid=ebd0be8791b5432b9b254574e4d515cf.

   AMD + Sentieon run the same sequencing data in an average of 12.42 minutes (114 Genomes/Day avg.). AMD System Configuration: 2P AMD EPYC 9654 (96 cores/socket, 192 cores/node); 1.5 TB (24x) Dual-Rank DDR5-4800 64GB DIMMs, 1DIMM per channel; 1 x 256 GB SATA (OS) | 1 x 2 TB NVMe (data); BIOS Version 1007D, SMT=off, Determinism=power, NPS=1 RHEL 8.7; OS settings: amd_iommu=on, iommu=pt, mitigations=off. Results may vary based on factors such as software version, hardware configurations, and BIOS version and settings.
5. SP5C-027: Testing conducted by Sentieon as of 1/31/2024. Instance Name: hpc7a.96xlarge; OS Type: Amazon Linux 2023; OS Version: 2023.3; Kernel Version: 6.1.66; Sentieon® Version: 202308.01; Cost per genome = 7.2 *17/60 = $2.02; AWS 2P EPYC 9R14 hpc7a.96xlarge $7.20/hr. Cloud performance results presented are based on the test date in the configuration. AMD has not independently verified the results. Results may vary due to changes to the underlying configuration, and other conditions such as the placement of the VM and its resources, optimizations by the cloud service provider, accessed cloud regions, co-tenants, and the types of other workloads exercised at the same time on the system.
6. SP5C-028: Testing conducted by Sentieon as of 1/31/2024. Instance Name: hpc7a.96xlarge; OS Type: Amazon Linux 2023; OS Version: 2023.3; Kernel Version: 6.1.66; Sentieon® Version: 202308.01; Cost per genome = 7.2 *40/60 = $4.8; AWS 2P EPYC 9R14 hpc7a.96xlarge = $7.20/hr. Cloud performance results presented are based on the test date in the configuration. AMD has not independently verified the results. Results may vary due to changes to the underlying configuration, and other conditions such as the placement of the VM and its resources, optimizations by the cloud service provider, accessed cloud regions, co-tenants, and the types of other workloads exercised at the same time on the system.

*Harini Malik is a Senior Director of Business Development at AMD.*

*Tony Nunes is a Senior Product Manager at AMD.*

*Don Freed is a Sr. Bioinformatics Scientist at Sentieon.*

*Brian Gallagher is the Head of Business Development at Sentieon.*

## DISCLAIMERS

*Links to third party sites are provided for convenience and unless explicitly stated, AMD is not responsible for the contents of such linked sites and no endorsement is implied.

The information contained herein is for informational purposes only and is subject to change without notice. While every precaution has been taken in the preparation of this document, it may contain technical inaccuracies, omissions and typographical errors, and AMD is under no obligation to update or otherwise correct this information. Advanced Micro Devices, Inc. makes no representations or warranties with respect to the accuracy or completeness of the contents of this document, and assumes no liability of any kind, including the implied warranties of noninfringement, merchantability or fitness for purposes, with respect to the operation or use of AMD hardware, software or other products described herein. No license, including implied or arising by estoppel, to any intellectual property rights is granted by this document. Terms and limitations applicable to the purchase or use of AMD's products are as set forth in a signed agreement between the parties or in AMD's Standard Terms and Conditions of Sale. GD-18