NVIDIA DGX BasePOD Solutions for Genomic Sequencing

White Paper

WP-11016-001 | October 2023

Document History

WP-11016-001

Version	Date	Authors	Description of Change
01		Gary Burnett, Jason Fenwick, John Sanson, and Robert Sohigian	Initial Release
02		Harry Clifford	Updated ONT benchmarks to Dorado
03		Harry Clifford	Updated Parabricks benchmarks to v4.1.1

Table of Contents

Abstract

The field of genomics is growing exponentially, transforming the healthcare, agriculture, and life-sciences industries, as well as being one of our greatest weapons in the fight against SARS-CoV-2 and COVID-19. Since the introduction of next-generation sequencing (NGS) in 2005, the field has experienced a data explosion, creating new industries built around the human genome, from genealogy research to early-stage detection of cancer. Today, commercially available sequencing platforms can sequence hundreds of whole genomes in one to two days, and some labs are sequencing 20,000 exomes per week. Instrument throughputs have exploded, and the cost of sequencing is approaching as little as $100 for a whole human genome.

Because of the advances in sequencing, the data analytics components of bioinformatics pipelines now represent a proportionally larger share of the cost and time for sequencing a human genome. This computational bottleneck can be a barrier for clinical labs, where extended turn-around times can have significant impacts on patient outcomes. It can also limit large scale sequencing programs and population projects, as the cost and time to analyze hundreds of thousands of genomes quickly adds up.

In this paper, we present the NVDIA BasePOD™ Solution for Genomic Sequencing as a scalable solution for bioinformatics pipelines, delivering accelerated, high-throughput genomics analysis solutions across the entire genomics workflow.

Introduction

With genomics data doubling every seven months, today's researchers need data-center-grade performance to handle the computationally intensive steps that transform sequencing read data into useful biological information. To continue making advancements in genomic sequencing, researchers will need a flexible and robust application ecosystem to facilitate broader research with more refined outcomes. Also, researchers will need world-class infrastructure that is easy to manage with predictable performance. To address the growing accelerated computing needs of genomic sequencing, NVIDIA has designed the NVIDIA DGX BasePOD for Genomic Sequencing. The solution is a reference infrastructure design created for genomics research and built around the industry leading NVIDIA DGX™ A100 system.

This white paper outlines the computational steps in sequencing for genomics and the key applications involved, then details how the NVIDIA Parabricks solution can simplify the sequencing process and improve data accuracy and access. Additionally, this paper provides an overview of the reference design and how it helps IT organizations simplify deployment, management, and scalability of accelerated compute infrastructure.

State-of-the-Art Genomic Sequencing

Today's massively parallel sequencing technology provides ultra-high throughput, scalability, and speed. This technology is used to determine the order of nucleotides in entire genomes or targeted regions of DNA or RNA. Sequencing delivers rapid answers to clinical questions for many patients, as well as speed-to-market for new product development.

NVIDIA accelerated solutions—including hardware and software—are uniquely positioned to address the major steps and associated bottlenecks of the entire genomics analysis workflow and are used by leading bioinformatics researchers, academic centers, cancer clinics, sequencing centers, and pharmaceutical companies around the world.

From biological samples through digital processing, the genomics workflow is complicated and produces a large amount of data. In the primary analysis step, the sample is consumed by a sequencing platform. Using signal processing, the platform converts sensor outputs into individual nucleic bases, generating millions of individual sequencing reads. Depending upon the application and platform, those sequencing reads are either assembled into a genome or aligned to a reference genome. The secondary analysis step generates a list of genetic variants within the sequence of a given sample. The tertiary analysis step centers on interpreting these variants and understanding their impact. This can involve determining the clinical relevance of a variant in the case of an individual sample, or the identification of genes associated with a particular disease in a population study.

Primary Analysis

NVIDIA is a key partner for many sequencing instrument providers, providing GPUs and software to increase the throughput of data generated, improve data quality, and reduce instrument runtime. GPUs can be used to accelerate computationally intensive steps in Primary Analysis such as signal processing and basecalling.

Basecalling is the process of classifying raw instrument signals into the A, C, G, and T bases of the genome. It is a computationally critical step in ensuring accurate sequence data that in turn impacts all downstream analysis tasks.

Oxford Nanopore Technology

Oxford Nanopore Technologies is an Oxford-based company focused on delivering disruptive sequencing technologies to the market. Oxford Nanopore's new generation of DNA and RNA sequencing technology offers real-time analysis in a variety of formats, ranging from pocket-size to bench-top devices. These sequencers can analyze native DNA or RNA and sequence fragment lengths as long as 4 Mb to achieve short to ultra-long read lengths.

Oxford Nanopore's PromethION P48 sequencer is the highest throughput platform offered and can generate as much as 14 terabases in a 72-hour run. The rapid classification task required for this already benefits from deep learning (DL) innovation and GPU acceleration. The core data processing toolkit for this purpose, Dorado, uses a recurrent neural network (RNN) for basecalling, with the option of three different architectures of smaller (faster) or larger (higher accuracy) recurrent layer sizes.

The main computational bottleneck in basecalling is the RNN layer. This bottleneck can be relieved using GPU integration with ONT sequencers: for example, the benchtop GridION Mk1 includes a single NVIDIA V100 Tensor Core GPU, and the handheld MinION Mk1C includes a Jetson Edge platform. To keep pace with the data generation of a PromethION P48, a more powerful computational platform is required. The DGX A100 system provides the necessary performance and scale-out capability to match the high throughput of the PromethION P48. Leveraging the DGX BasePOD, researchers can accelerate the number of samples processed by orders of magnitude for higher throughput and increased accuracy.

Table 1. Basecalling throughput for Dorado in number of samples processed per second.

Model	Throughput (samples per second, in millions)		
	1 DGX System	4 DGX Systems	8 DGX Systems
Fast	1,196 M	4,783 M	9,566 M
High Accuracy	425 M	1,701 M	3,403 M
Super High Accuracy	87 M	349 M	698 M

Secondary Analysis

Secondary Analysis focuses on identifying the genetic variants in a sample, and immediately follows the primary analysis steps. First, the reads sequenced during the primary analysis are aligned or mapped to a reference genome, most commonly GRCh38 (the latest stable human reference genome). Alignment is the process of taking the resultant basecalled fragments of DNA, now in the form of character strings of As, Cs, Gs, and Ts, and determining the genome location where those fragments originated, assembling a full genome from the massively parallelized sequencing process. The final step in secondary analysis is typically variant calling, which identifies the differences between the sequenced sample and the reference genome. The identification of these variants can uncover genetic causes for diseases, provide targets for drug development, and enhance patient care.

NVIDIA Parabricks—Alignment

NVIDIA Parabricks includes tools to align and process FASTQ files (the output of primary analysis) for both DNA and RNA sequencing. The software allows users to run many of the steps individually, but also provides fq2bam and rna_fq2bam pipelines for a simplified, accelerated workflow using best practices for identifying duplicate reads and base quality score recalibration (BQSR). This paper will focus on the popular fq2bam pipeline for DNA sequencing.

The fq2bam pipeline accelerates the GATK4 Best Practices workflow, providing equivalent results with speeds up to 45X faster. Fq2bam takes one or more pairs of FASTQ files as inputs, and outputs a BAM/CRAM with an optional BQSR report. The pipeline runs BWA-Mem alignment, Coordinate Sorting, Mark Duplicates, and BQSR (if the *–knownSites input* and *–out-recal-file output* options are provided).

Figure 1. The NVIDIA Parabricks fq2bam workflow, delivering equivalent output to BWA-MEM and GATK4 best practices.

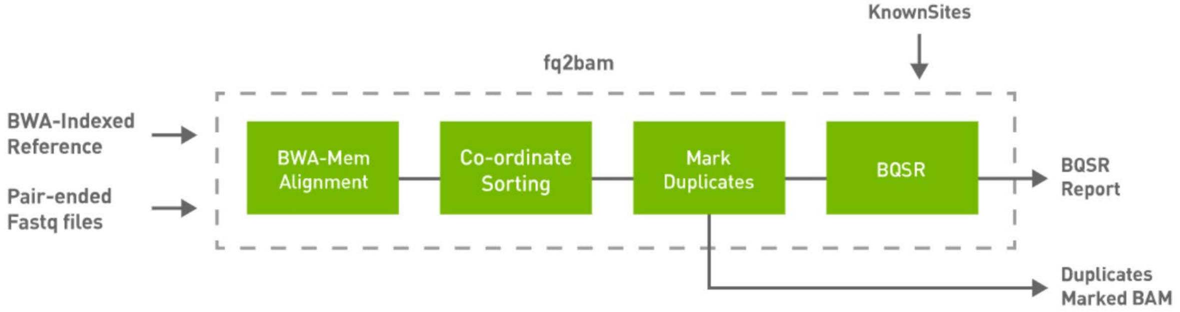


Table 2. Performance of Parabricks v4.1.1 fq2bam on DGX A100 systems for whole genome analysis, displaying the impact of the DGX BasePOD on throughput of genomics pipelines. Benchmarked on HG002 Genome-in-a-bottle sample, whole genome sequenced at 5x, 30x, and 50x depth of coverage.

Whole Genome Coverage	No. of DGX Systems	Genomes per Day	Genomes per Week (7 days)	Genomes per Month (30 days)	Genomes per Year (365 days)
5x	1	360	2,520	10,800	131,400
	4	1,440	10,080	43,200	525,600
	8	2,880	20,160	86,400	1,051,200
30x	1	90	630	2,700	32,850
	4	360	2,520	10,800	131,400
	8	720	5,040	21,600	262,800
50x	1	60	420	1,800	21,900
	4	240	1,680	7,200	87,600
	8	480	3,360	14,400	175,200

NVIDIA Parabricks—Variant Calling

Variant calling is the portion of the workflow designed to identify all points in the newly assembled genome that differ from a reference genome. This involves scanning the full breadth of the genome to look for different types of variation. This can include anything from small single-base-pair variants all the way to large structural variants covering thousands of base-pairs. Parabricks provides tools for variant calling spanning germline, somatic, and RNA applications.

Germline analysis is used to detect inherited variants, and can be used for clinical genetic testing, population studies, and disease research. Parabricks includes GPU-accelerated versions of several gold-standard germline variant callers, including GATK's

HaplotypeCaller and Google's DeepVariant. The Parabricks version of HaplotypeCaller can be run on a 30X whole genome sample in as little as 4 minutes, and DeepVariant can take as little as 11 minutes, compared to tens of hours on a CPU-only system.

Somatic mutation analysis is used to detect the genetic alterations that are not inherited, but acquired during one's lifespan, such as those present in a tumor. This analysis is important in diagnosing cancers and identifying courses of treatment for cancer patients based on their specific mutations. Parabricks uses Mutect2 from GATK as its GPU-accelerated variant caller, running in under 45 minutes on 50X SEQC2 data, compared to 31 hours for the CPU version.

Table 3. Performance of Parabricks v4.1.1 GPU-accelerated HaplotypeCaller on DGX A100 systems for whole genome germline variant calling. Benchmarked on HG002 Genome-in-a-bottle sample, whole genome sequenced at 5x, 30x, and 50x depth of coverage.

Whole Genome Coverage	No. of DGX Systems	Genomes per Day	Genomes per Week (7 days)	Genomes per Month (30 days)	Genomes per Year (365 days)
5x	1	720	5,040	21,600	262,800
	4	2,880	20,160	86,400	1,051,200
	8	5,760	40,320	172,800	2,102,400
30x	1	360	2,520	10,800	131,400
	4	1,440	10,080	43,200	525,600
	8	2,880	20,160	86,400	1,051,200
50x	1	288	2,016	8,640	105,120
	4	1,152	8,064	34,560	420,480
	8	2,304	16,128	69,120	840,960

NVIDIA Parabricks End-to-End Pipelines

Parabricks also includes intuitive end-to-end pipelines for simple, accelerated deployment of germline, somatic, and RNA workflows. These pipelines replicate GATK Best Practices for germline and somatic variant calling, and there is another pipeline that utilizes DeepVariant as the germline variant caller.

Figure 2 and Table 4 illustrate the end-to-end germline workflow (run using a single command with Parabricks) and the performance for users running on Genomics POD. With Parabricks, the germline pipeline takes as little as 19 minutes to complete on a 30X genome, compared to over 20 hours for the non-accelerated version.

Figure 2. The NVIDIA Parabricks end-to-end germline workflow, delivering equivalent output to BWA-MEM, HaplotypeCaller, and GATK4 best practices.

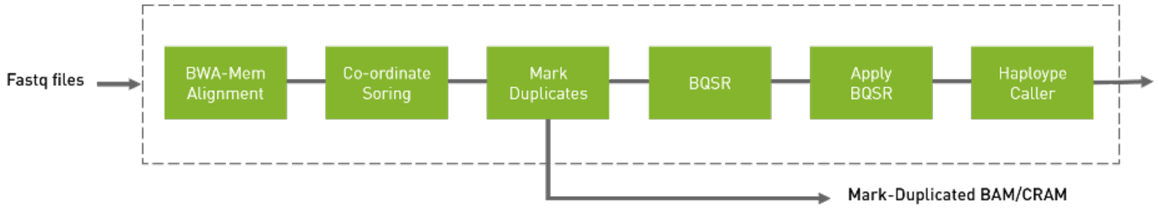


Table 4. Performance of Parabricks v4.1.1 GPU-accelerated GATK germline workflow on DGX A100 systems for whole genomes. Benchmarked on HG002 Genome-in-a-bottle sample, whole genome sequenced at 5x, 30x, and 50x depth of coverage.

Whole Genome Coverage	No. of DGX Systems	Genomes per Day	Genomes per Week (7 days)	Genomes per Month (30 days)	Genomes per Year (365 days)
5x	1	240	1,680	7,200	87,600
	4	960	6,720	28,800	350,400
	10	1,920	13,440	57,600	700,800
30x	1	75	530	2,273	27,663
	4	303	2,122	9,094	110,652
	10	606	4,244	18,189	221,305
50x	1	51	360	1,542	18,771
	4	205	1,440	6,171	75,085
	10	411	2,880	12,342	150,171

Tertiary Analysis

The tertiary analysis step ties the specific genetic variants to observable characteristics or traits (i.e. phenotypes). In clinical applications, a single individual can be analyzed in isolation, as only known, clinically actionable genetic variants are used; other studies try to annotate the significance of any given variant using a cohort of samples. Common tasks in tertiary analysis applications include array manipulation and computation, regressions, clustering, machine learning, and visualization.

NVIDIA RAPIDS

RAPIDS is a suite of open-source libraries that can speed up data science workflows through the power of GPU acceleration. RAPIDS was designed to accelerate the whole realm of data science, making it well suited for accelerating tertiary analysis of bioinformatics data. RAPIDS makes it possible to perform interactive data analysis on large datasets using Python APIs that closely resemble NumPy, Pandas, and scikit-learn. With minimal code changes, and no new tools to learn, developers can speed up operations at an order of magnitude. Common data science tasks—including data preparation and manipulation with cuDF and ML with cuML–can be further scaled out with GPU-accelerated SPARK and DASK. It is possible to achieve data science workflows that are 70X faster, and 20X as cost effective--when compared to a similar CPU configuration.

Pre-Processing Steps can be accelerated with cuDF, a Python GPU DataFrame library for loading, joining, aggregating, filtering, and otherwise manipulating data--all in a pandas-like API familiar to data scientists. Users can create GPU DataFrames from Numpy arrays, pandas DataFrames, and PyArrow Tables, and, once in cuDF format, other GPU-accelerated libraries can be used to easily conduct machine learning and analytics processes.

Rapids Single Cell Analysis

Using the flexibility of RAPIDS, NVIDIA researchers have been able to accelerate entire single-cell genome analysis pipelines without writing custom CUDA code. Preprocessing, dimension reduction, clustering, and visualization are all done with calls to RAPIDS libraries. These libraries also transparently handle memory transactions and spilling from

GPU to system memory for huge datasets, allowing this pipeline to handle over 1 million cells.

Consider a typical workflow to perform single cell analysis, beginning with a matrix that maps the counts of each gene encountered in each cell. Preprocessing steps are performed to filter out noise, then the data is normalized to obtain the activity of every human gene in every individual cell of the dataset. Machine learning is also commonly used in this step to correct artifacts from data collection. Next, dimensionality reduction is performed before clustering and visualization to identify clusters of cells with similar genetic activity. Finally, the genetic activity of these cell clusters is compared to understand why different types of cells behave and respond differently.

Table 5 demonstrates the use of RAPIDS to accelerate the analysis of single-cell RNA-seq data from 1.3 million cells on CPU systems versus DGX systems, including preprocessing, dimension reduction, clustering, and visualization.

Table 5. Analysis of single-cell RNA-seq data on CPU versus DGX systems. For more details on the analysis and benchmarking data refer to https://github.com/clara-parabricks/rapids-single-cell-examples.

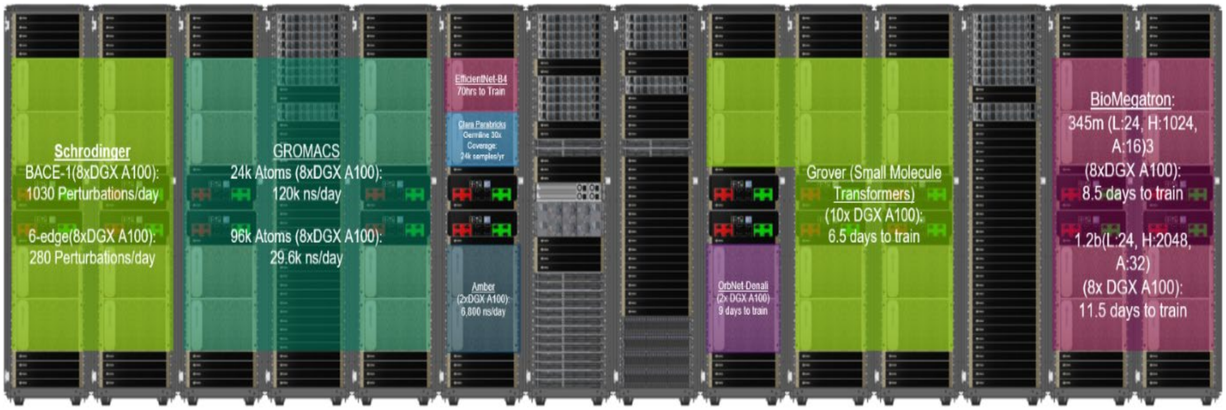
Task	CPU (sec)	A100 GPU (sec)	Speedup
Data load + Preprocessing	1120	475	2.4X
PCA	44	17.8	2.5X
t-SNE	6509	37	175.9X
k-means (single iteration)	148	2	74X
KNN	154	62	2.5X
UMAP	2571	21	122X
Louvain clustering	1153	2.4	480X
Leiden clustering	6345	1.7	3732X
Re-analysis of subgroup	255	17.9	14X
End-to-end notebook run	18338	686	26.7X

Next Generation Genomic Sequencing Requires Leadership-Class Computing Infrastructure

Designing and building scaled computing infrastructure for AI requires an understanding of the computing goals of genomic researchers to build fast, capable, and cost-efficient systems. Developing infrastructure requirements can often be difficult because the needs of research are often an ever-moving target. Additionally, crafting robust benchmarks that represent the overall needs of an organization is a time-consuming process. This dilemma requires organizations to leverage a standardized approach to building and scaling computational infrastructure.

The NVIDIA DGX BasePOD™ is designed to help genomic research organizations overcome the difficulties in designing, deploying, operating, and maintaining a best-in-class operating environment for the most intensive AI and HPC workloads. As the leader in accelerated computing and data science, NVIDIA has leveraged experience from deploying solutions with the leading research organizations to build the reference design. This solution takes the focus off IT and allows researchers to focus on their most important health initiatives.

Figure 1. Sample DGX BasePOD for Genomics Sequencing configuration



Leading HLS engineers at NVIDIA have benchmarked the sequencing applications highlighted in this document using the design of this validated infrastructure platform. Using the world class DGX A100 system and following the DGX BasePOD architecture, an IT department can follow a prescriptive process to scale its environment to meet the most intensive performance needs in the world. To make deployment and management seamless and efficient, the solution comes with the NVIDIA Bright Cluster Manager™ platform, which allows IT administrators to streamline deployment, management, and monitoring, removing the risks and delays involved in building out enterprise GPU-enabled clusters.

As with any DGX investment, the DGX BasePOD features NVIDIA AI Enterprise Software, which includes the entire suite of Parabricks software licenses without additional cost. Users also gain access and support to a plethora of NVIDIA developed optimized applications and toolkits. NVIDIA AI Enterprise Software and NGC (NVIDIA GPU Cloud) offer a curated set of NVIDIA applications and integrations that simplify building, customizing, and integration of GPU-optimized software into workflows, accelerating the time to productivity for users. Applications such as NVIDIA Parabricks and RAPIDS are deployed and accelerated with toolkits in the NGC catalog and NVIDIA AI Enterprise. With DGX BasePOD, IT departments, data scientists and researchers can build an AI Center of Excellence with a NVIDIA validated technology stack to drive all GPU enabled workloads for genomics.

With the NVIDIA DGX BasePOD Solution for Genomic Sequencing, organizations can expect the following:

- A fully validated operating environment for infrastructure management, data science, and research
- An NVIDIA-engineered solution architected for streamlined scalability and predictable performance
- Powerful scale-up nodes, a large memory footprint, and fast connections between the GPUs for computing to support the variety of DLD models and HPC applications
- A low-latency, high-bandwidth network interconnect designed with the capacity and topology to minimize bottlenecks
- Support and access from leading NVIDIA engineers to drive initiatives
- DGX workshops for startup and operations for IT administrators

Notice

This document is provided for information purposes only and shall not be regarded as a warranty of a certain functionality, condition, or quality of a product. NVIDIA Corporation (“NVIDIA”) makes no representations or warranties, expressed or implied, as to the accuracy or completeness of the information contained in this document and assumes no responsibility for any errors contained herein. NVIDIA shall have no liability for the consequences or use of such information or for any infringement of patents or other rights of third parties that may result from its use. This document is not a commitment to develop, release, or deliver any Material (defined below), code, or functionality.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and any other changes to this document, at any time without notice.

Customer should obtain the latest relevant information before placing orders and should verify that such information is current and complete.

NVIDIA products are sold subject to the NVIDIA standard terms and conditions of sale supplied at the time of order acknowledgement, unless otherwise agreed in an individual sales agreement signed by authorized representatives of NVIDIA and customer (“Terms of Sale”). NVIDIA hereby expressly objects to applying any customer general terms and conditions with regards to the purchase of the NVIDIA product referenced in this document. No contractual obligations are formed either directly or indirectly by this document.

NVIDIA products are not designed, authorized, or warranted to be suitable for use in medical, military, aircraft, space, or life support equipment, nor in applications where failure or malfunction of the NVIDIA product can reasonably be expected to result in personal injury, death, or property or environmental damage. NVIDIA accepts no liability for inclusion and/or use of NVIDIA products in such equipment or applications and therefore such inclusion and/or use is at customer’s own risk.

NVIDIA makes no representation or warranty that products based on this document will be suitable for any specified use. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer’s sole responsibility to evaluate and determine the applicability of any information contained in this document, ensure the product is suitable and fit for the application planned by customer, and perform the necessary testing for the application in order to avoid a default of the application or the product. Weaknesses in customer’s product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this document. NVIDIA accepts no liability related to any default, damage, costs, or problem which may be based on or attributable to: (i) the use of the NVIDIA product in any manner that is contrary to this document or (ii) customer product designs.

No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this document. Information published by NVIDIA regarding third-party products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property rights of the third party, or a license from NVIDIA under the patents or other intellectual property rights of NVIDIA.

Reproduction of information in this document is permissible only if approved in advance by NVIDIA in writing, reproduced without alteration and in full compliance with all applicable export laws and regulations, and accompanied by all associated conditions, limitations, and notices.

THIS DOCUMENT AND ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, “MATERIALS”) ARE BEING PROVIDED “AS IS.” NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. TO THE EXTENT NOT PROHIBITED BY LAW, IN NO EVENT WILL NVIDIA BE LIABLE FOR ANY DAMAGES, INCLUDING WITHOUT LIMITATION ANY DIRECT, INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE, OR CONSEQUENTIAL DAMAGES, HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF ANY USE OF THIS DOCUMENT, EVEN IF NVIDIA HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA’s aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the Terms of Sale for the product.

Trademarks

NVIDIA, the NVIDIA logo, DGX BasePOD, and Bright Cluster Manager are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

VESA DisplayPort

DisplayPort and DisplayPort Compliance Logo, DisplayPort Compliance Logo for Dual-mode Sources, and DisplayPort Compliance Logo for Active Cables are trademarks owned by the Video Electronics Standards Association in the United States and other countries.

HDMI

HDMI, the HDMI logo, and High-Definition Multimedia Interface are trademarks or registered trademarks of HDMI Licensing LLC.

Arm

Arm, AMBA, and Arm Powered are registered trademarks of Arm Limited. Cortex, MPCore, and Mali are trademarks of Arm Limited. All other brands or product names are the property of their respective holders. "Arm" is used to represent Arm Holdings plc; its operating company Arm Limited; and the regional subsidiaries Arm Inc.; Arm KK; Arm Korea Limited.; Arm Taiwan Limited; Arm France SAS; Arm Consulting (Shanghai) Co. Ltd.; Arm Germany GmbH; Arm Embedded Technologies Pvt. Ltd.; Arm Norway, AS, and Arm Sweden AB.

OpenCL

OpenCL is a trademark of Apple Inc. used under license to the Khronos Group Inc.

Copyright

NVIDIA Corporation | 2788 San Tomas Expressway, Santa Clara, CA 95051
http://www.nvidia.com