
AWS
re:Invent
NOV. 28 – DEC. 2, 2022 | LAS VEGAS, NV

PRT220

SPONSORED BY NVIDIA

# How to maximize HPC productivity, performance, and portability

Dr. John Linford (he/him)

Principal Technical Manager, CPU Software
NVIDIA

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Agenda

Introduction to the NVIDIA platform

Tour of the NVIDIA HPC SDK

- Multi-platform standards-compliant compilers
- Programming models for performance and portability
- High performance math libraries

Demonstration of NVIDIA HPC SDK on Amazon EC2

aws

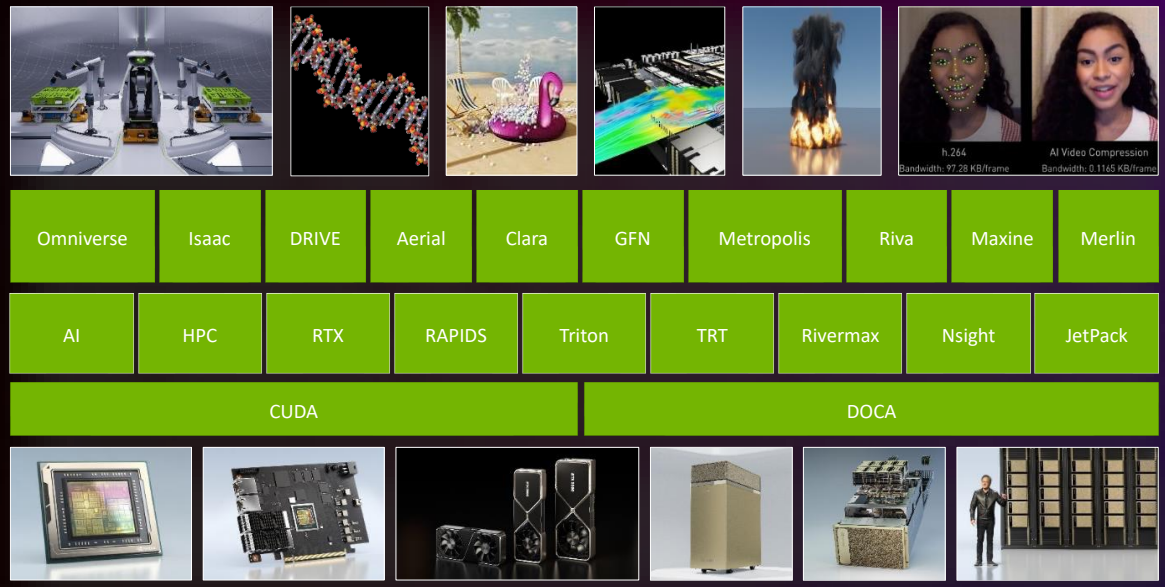
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# The NVIDIA platform

## Best-in-class hardware platform solutions



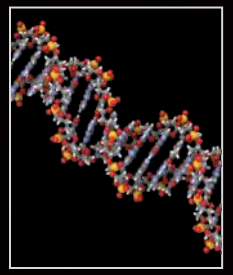

## Frameworks, libraries, tools, and applications

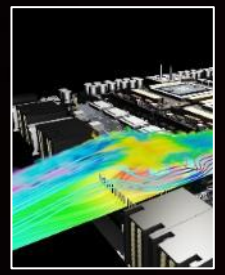


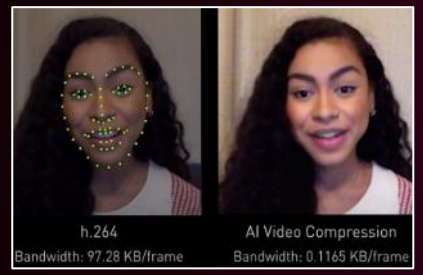

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# The NVIDIA developer ecosystem

**2.5M DEVELOPERS | 7,600 STARTUPS**






| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Omniverse | Isaac | DRIVE | Aerial | Clara | GFN | Metropolis | Riva | Maxine | Merlin |
| AI | HPC | RTX | RAPIDS | Triton | TRT | Rivermax | Nsight | JetPack | |
| CUDA | | | | | DOCA | | | | |








**500+**
GitHub projects

**60+**
GPU accelerated SDKs

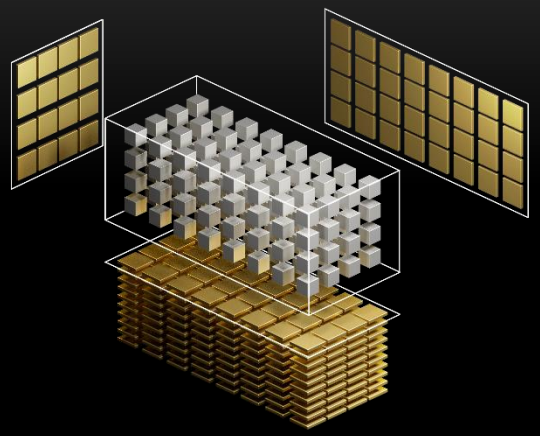
**50+**
developer libraries

**20+**
developer tools and utilities

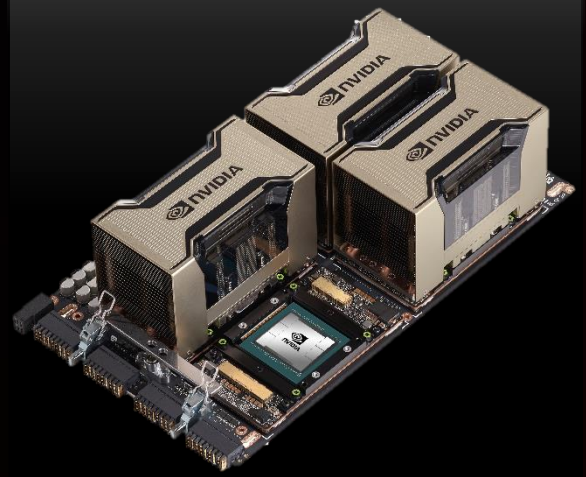
aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# NVIDIA HPC software initiatives

## Seamless acceleration
Standard languages, tensor cores, GH C2C



## Scaling up
Multi-GPU and multi-node libraries




## Domain libraries
Quantum

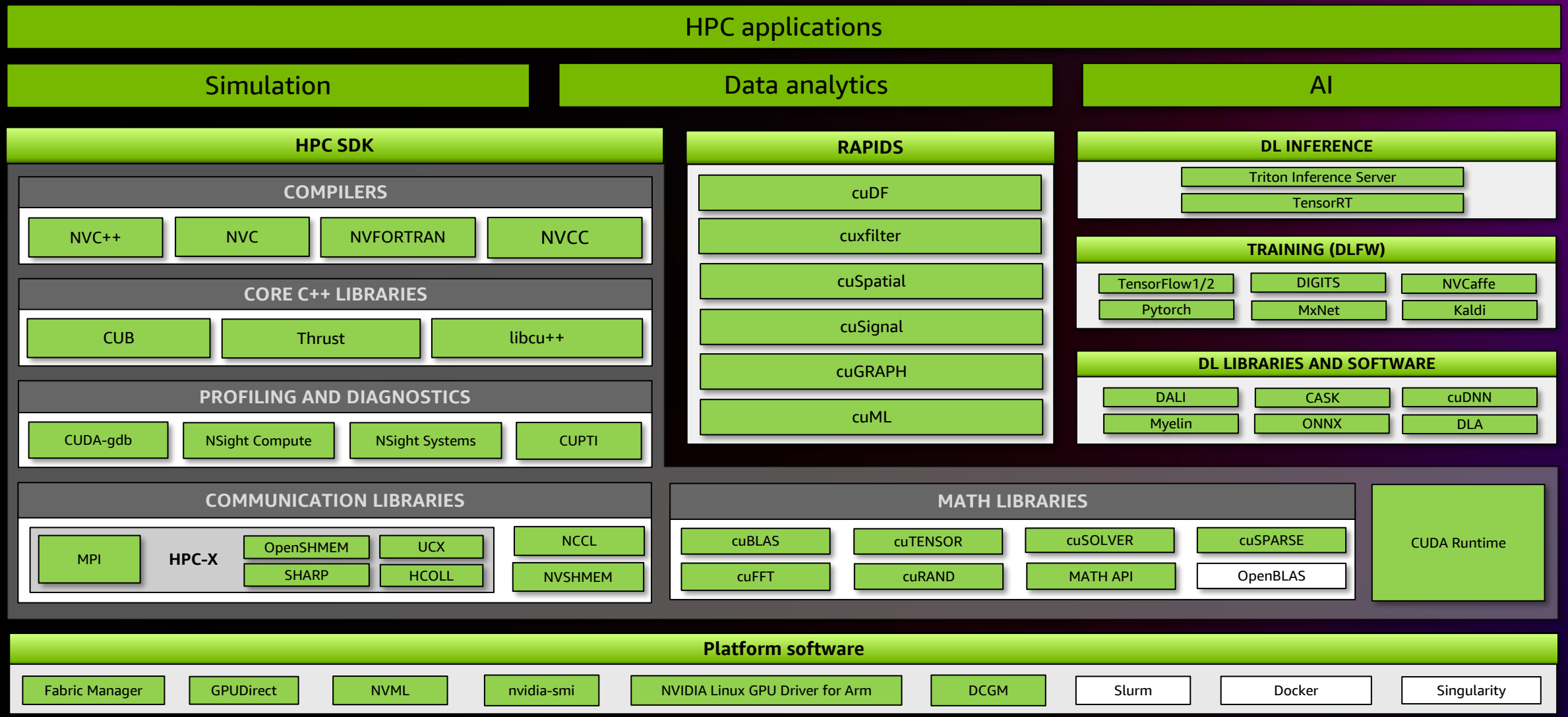


## Arm CPU software
Compilers, libraries, ecosystem




aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# NVIDIA HPC software

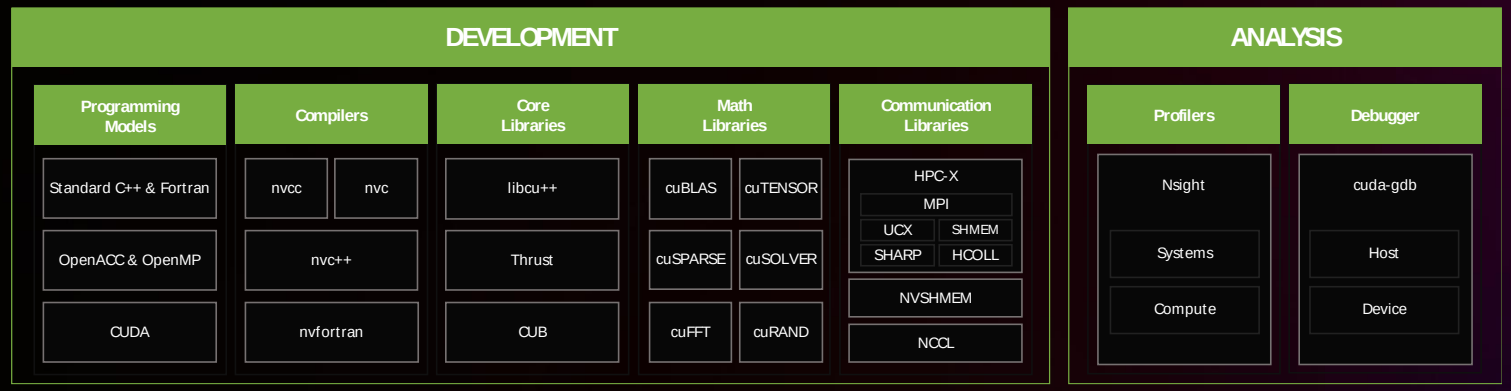



aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# NVIDIA HPC SDK

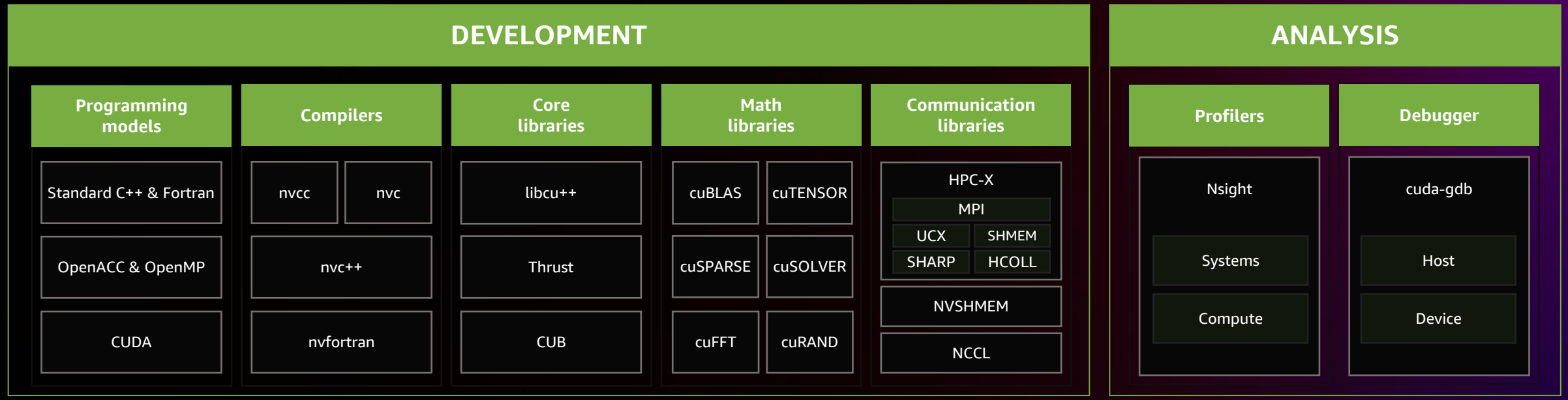
## A COMPREHENSIVE SUITE OF COMPILERS, LIBRARIES, AND TOOLS FOR HPC




© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# NVIDIA HPC SDK

AVAILABLE AT DEVELOPER.NVIDIA.COM/HPC-SDK, ON NGC, VIA SPACK, AND IN THE CLOUD




Developed for the NVIDIA platform: GPU, CPU, and Interconnect
Libraries | Accelerated C++ and Fortran | Directives | CUDA
x86_64 | Arm | OpenPOWER
7–8 releases per year | Freely available

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# NVIDIA HPC SDK is ready for Arm

## Auto vectorization

## Vector intrinsics

```
float32x2_t vadd_f32 (float32x2_t a, float32x2_t b);
float32x4_t vdupq_n_f32 (float32_t value);
float32x2_t vget_high_f32 (float32x4_t a);
```

## Host parallelism

Standard Parallelism

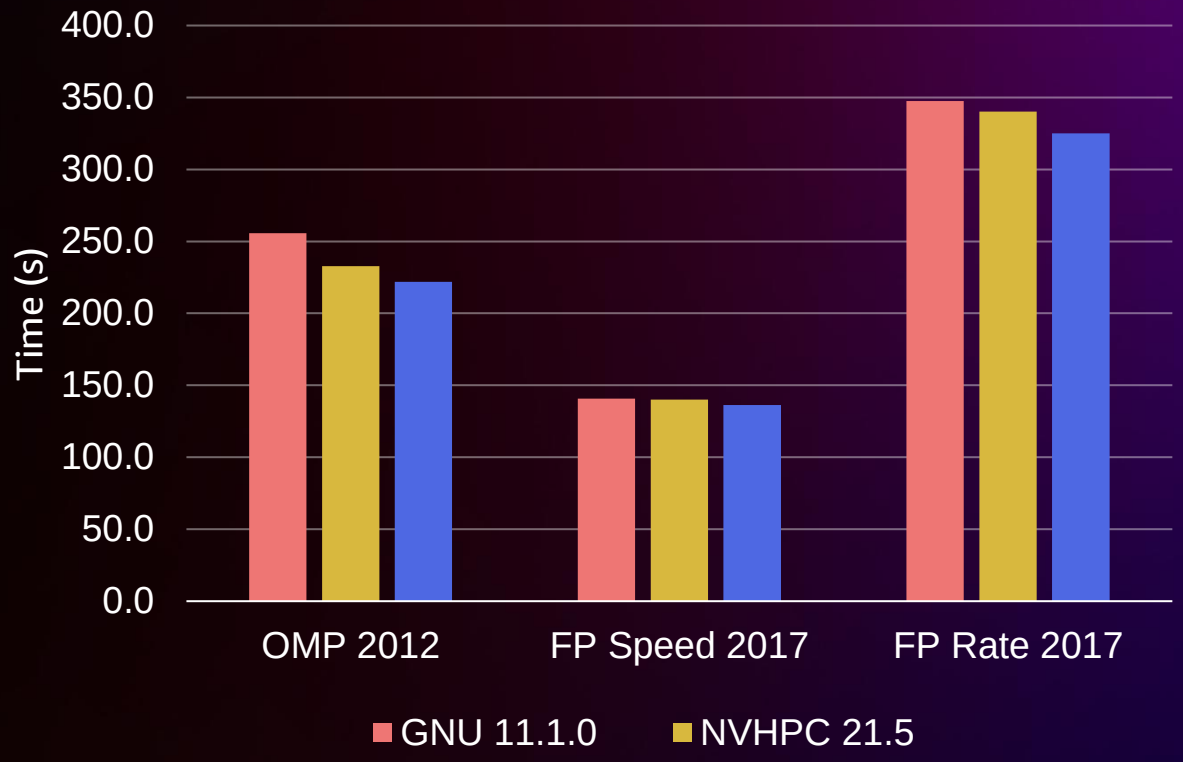
OpenACC

OpenMP

## FP models

- Precise
- Fast
- Relaxed

Estimated SPEC OMP 2012, SPEC CPU 2017
Floating point speed and rate GEOMEAN on
80-core Ampere Altra




aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

SPEC OMP and SPEC CPU are registered trademarks of the Standard Performance Evaluation Corporation (www.spec.org)

# NVIDIA HPC SDK supports AWS Graviton

- NVIDIA HPC SDK fully supports the AWS Graviton3 CPU
- Auto-vectorization for SVE and NEON
- See NVIDIA technical blog for details

https://bit.ly/3FWGwYB



Technical Blog

Subscribe

News

Sep 19, 2022 English

## Accelerating NVIDIA HPC Software with SVE on AWS Graviton3

By John Linford, Scott Manley and Graham Lopez

Discuss (1) 0 Like

Tags: AWS, featured, HPC / Supercomputing, HPC SDK, News

The latest NVIDIA HPC SDK update expands portability and now supports the Arm-based AWS Graviton3 processor. In this post, you learn how to enable Scalable Vector Extension (SVE) auto-vectorization with the NVIDIA compilers to maximize the performance of HPC applications running on the

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# NVIDIA HPC compilers

PORTABILITY ACROSS ALL HPC-CLASS AMAZON EC2 INSTANCE TYPES




## Accelerated
GPU
Automatic

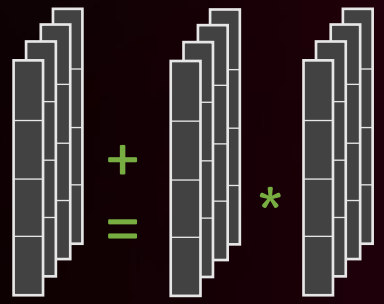
Fortran

C++

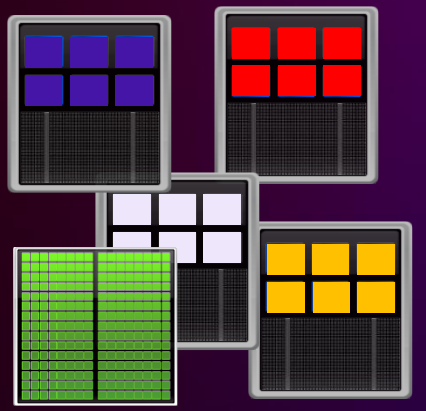
OpenACC
More Science, Less Programming

OpenMP

## Programmable
Standard languages
Directives
CUDA



## CPU optimized
Directives
Vectorization



## Multi-platform
x86_64
AArch64
OpenPOWER

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# NVIDIA HPC SDK programming models

PRODUCTIVITY AND PERFORMANCE FOR CPU, GPU, AND NETWORK PROGRAMMING

## ACCELERATED STANDARD LANGUAGES
ISO C++, ISO Fortran

```
std::transform(par, x, x+n, y, y,
    [=](float x, float y){ return y +
a*x; }
);
```

---

```
do concurrent (i = 1:n)
   y(i) = y(i) + a*x(i)
enddo
```

---

```
import cunumeric as np
…
def saxpy(a, x, y):
    y[:] += a*x
```

## INCREMENTAL PORTABLE OPTIMIZATION
OpenACC, OpenMP

```
#pragma acc data copy(x,y) {
...
std::transform(par, x, x+n, y, y,
    [=](float x, float y){
        return y + a*x;
});
...
}

#pragma omp target data map(x,y) {
...
std::transform(par, x, x+n, y, y,
    [=](float x, float y){
        return y + a*x;
});
...
}
```

## PLATFORM SPECIALIZATION
CUDA

```
__global__
void saxpy(int n, float a,
           float *x, float *y) {
  int i = blockIdx.x*blockDim.x +
          threadIdx.x;
  if (i < n) y[i] += a*x[i];
}

int main(void) {
  ...
  cudaMemcpy(d_x, x, ...);
  cudaMemcpy(d_y, y, ...);

  saxpy<<<(N+255)/256,256>>>(...);

  cudaMemcpy(y, d_y, ...);
```

## ACCELERATION LIBRARIES

Core | Math | Communication | Data analytics | AI | Quantum

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Accelerated standard languages

PARALLEL PERFORMANCE FOR WHEREVER YOUR CODE RUNS

## ISO C++

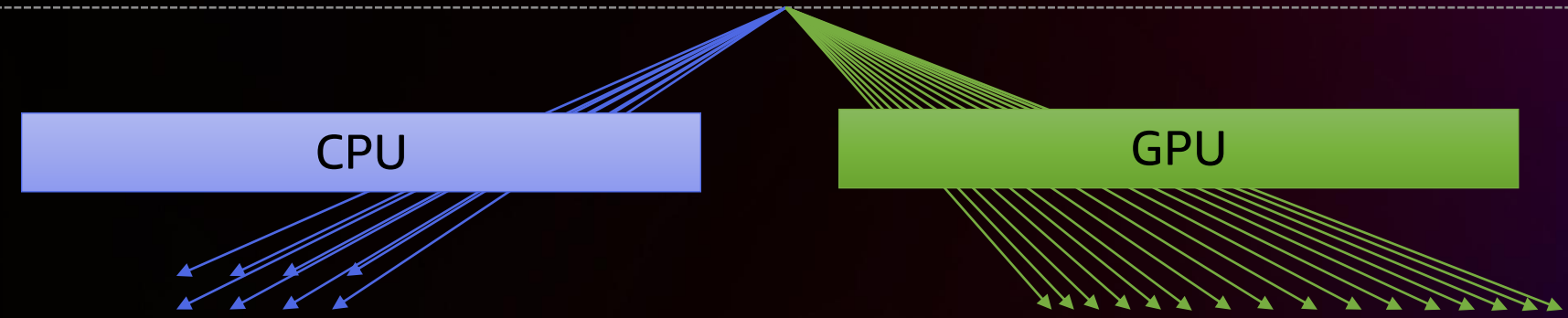
```
std::transform(par, x, x+n, y,
   y,[=](float x, float y){
      return y + a*x;
    }
);
```

## ISO Fortran

```
do concurrent (i = 1:n)
   y(i) = y(i) + a*x(i)
enddo
```

## Python

```
import cunumeric as np
…
def saxpy(a, x, y):
    y[:] += a*x
```




nvc++ -stdpar=**multicore**
nvfortran –stdpar=**multicore**
**legate –cpus 16** saxpy.py

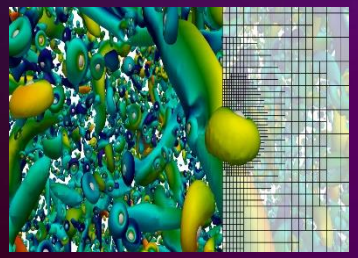
nvc++ -stdpar=**gpu**
nvfortran –stdpar=**gpu**
**legate –gpus 1** saxpy.py

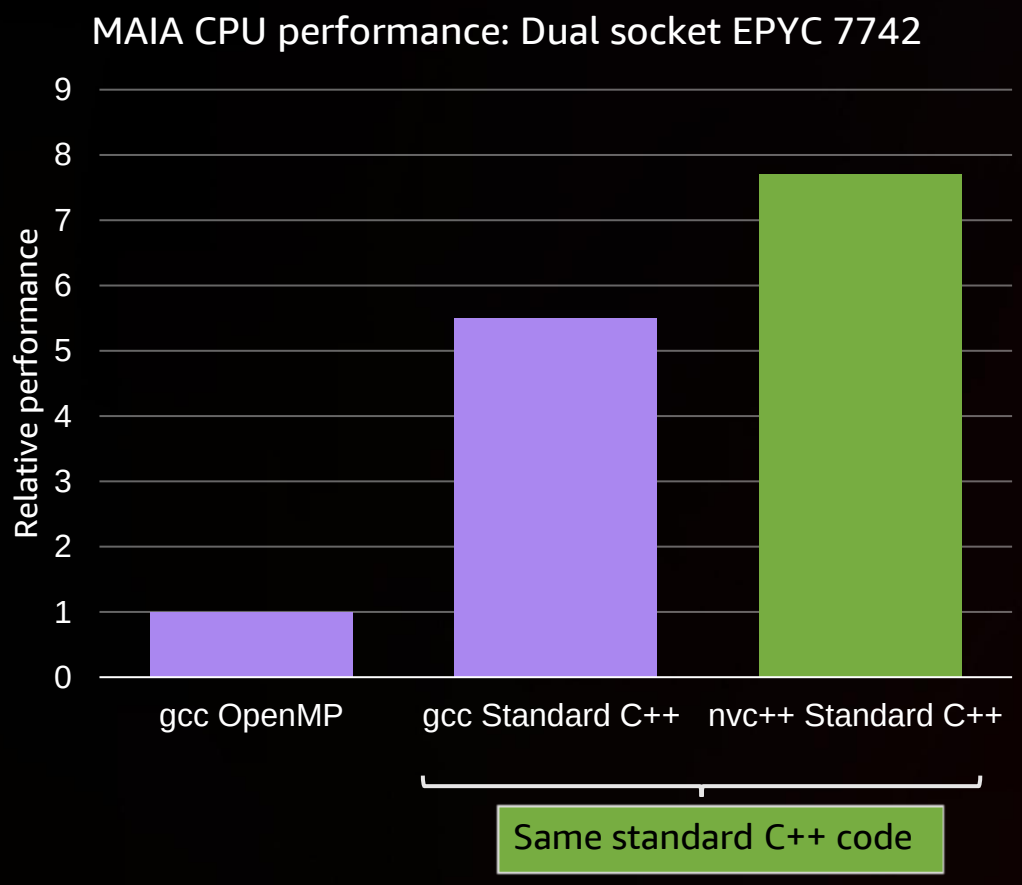
aws

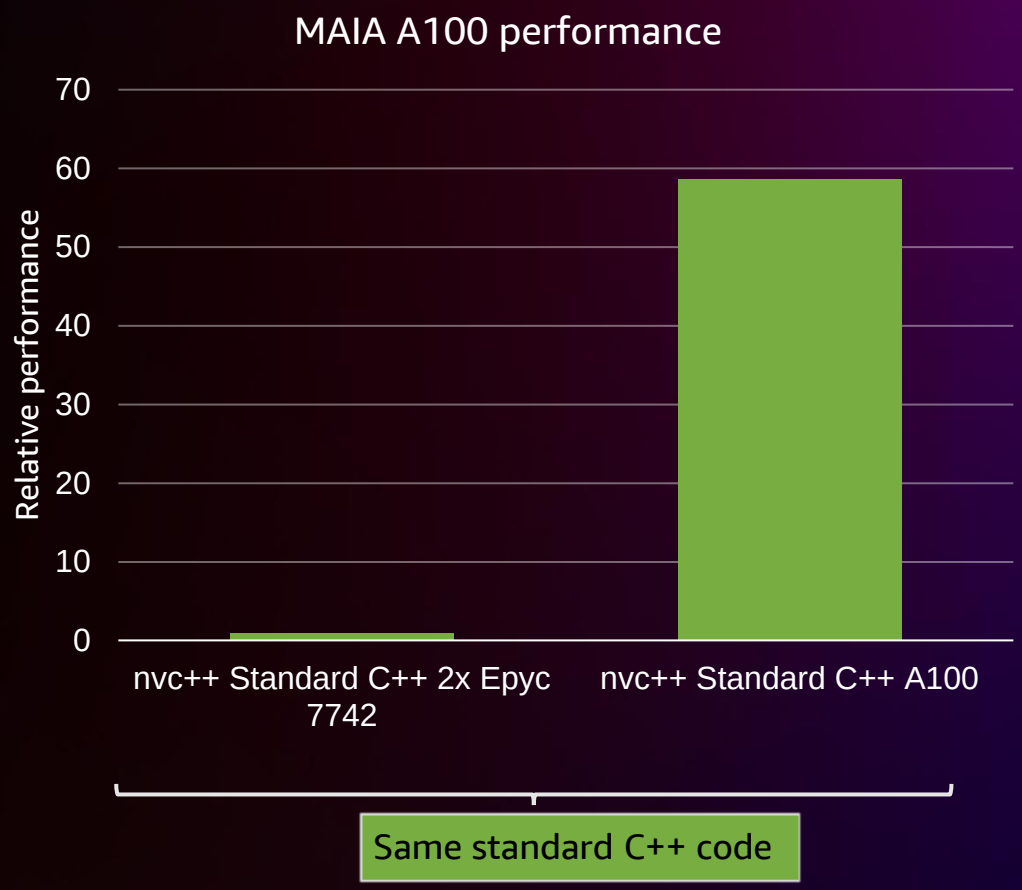
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# C++ standard parallelism

M-AIA ON CPUS AND GPUS



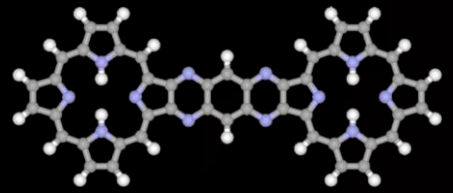


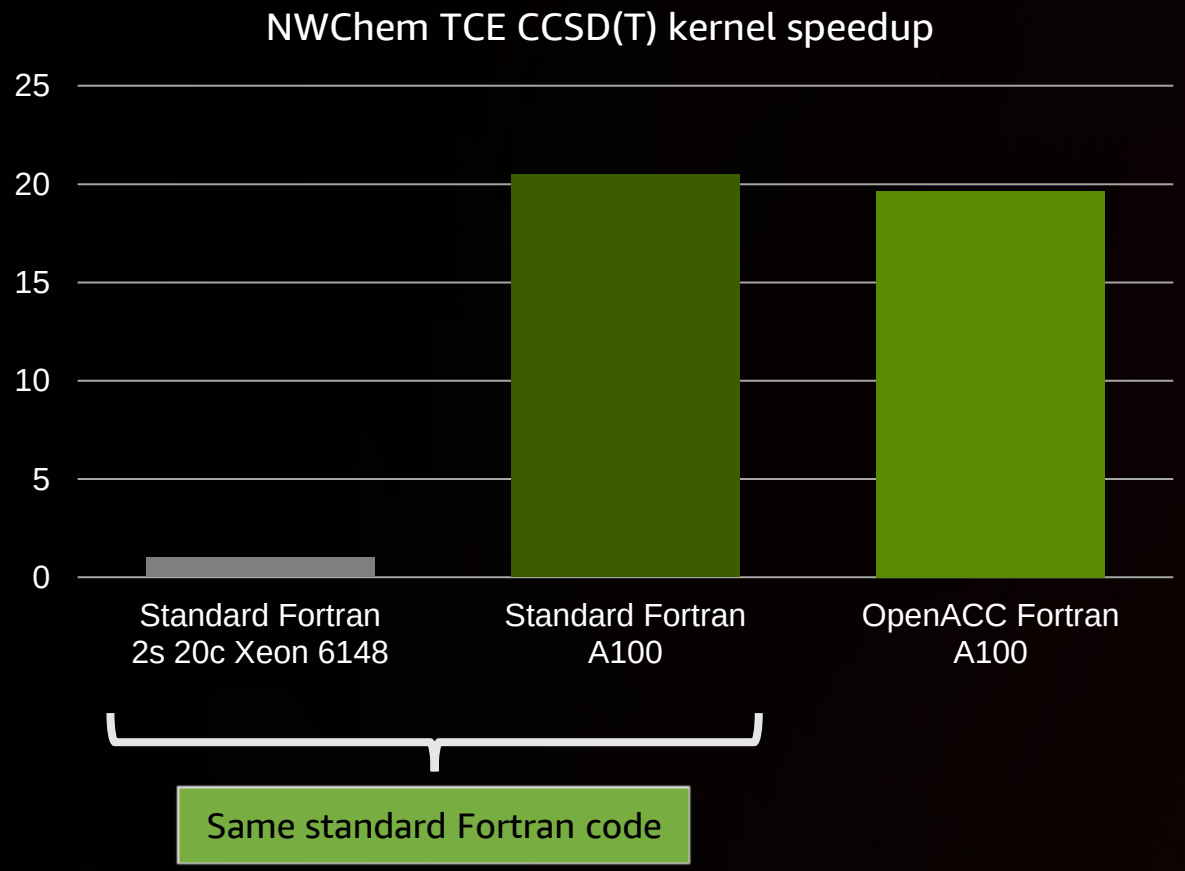

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

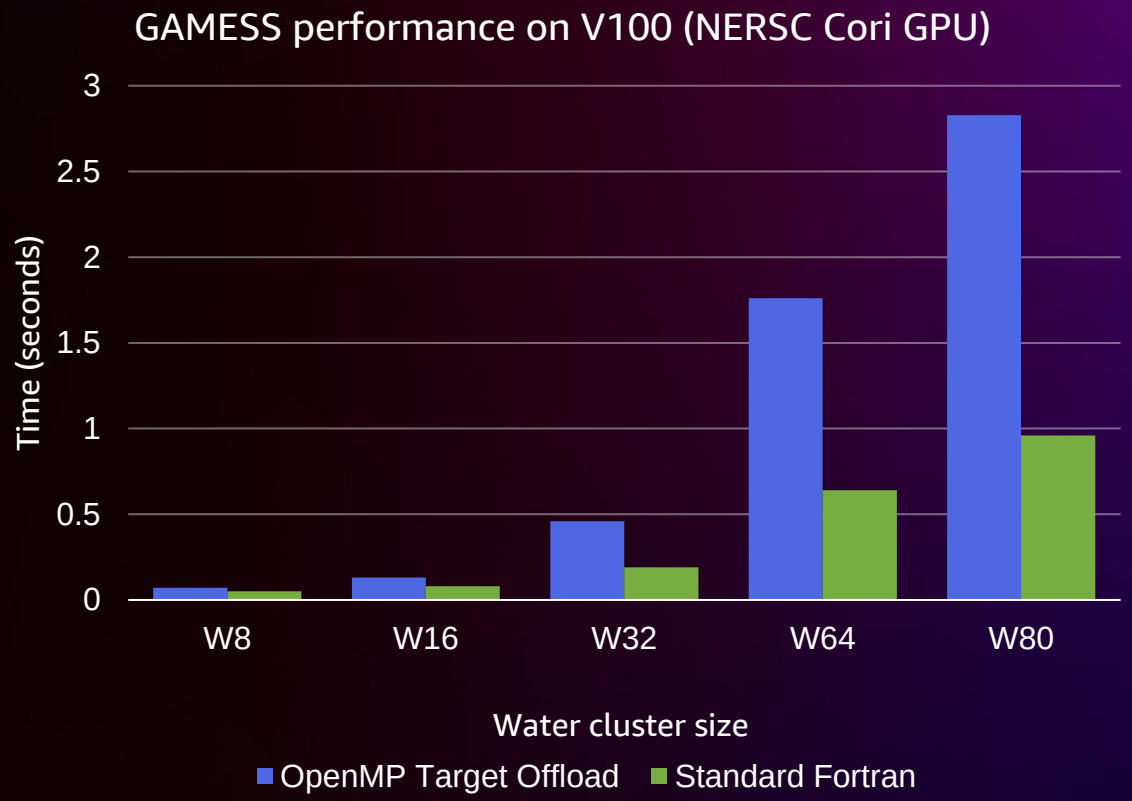
# Fortran standard parallelism

NWCHEM AND GAMESS WITH DO CONCURRENT



GAMESS



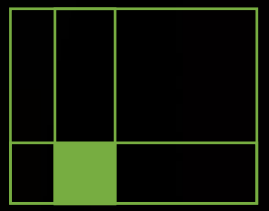

GAMESS results from Melisa Alkan and Gordon Group, Iowa State

aws

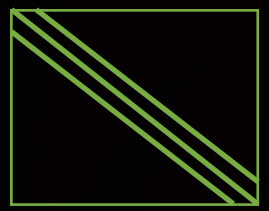
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# NVIDIA math libraries

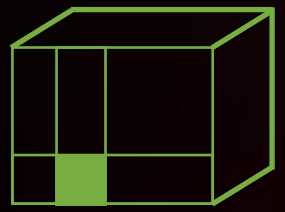
LINEAR ALGEBRA, FFT, RNG, AND BASIC MATH



cuBLAS



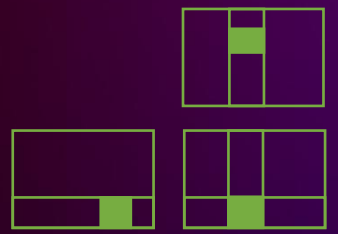
cuSPARSE



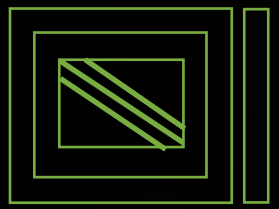
cuTENSOR



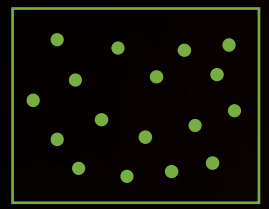
cuSOLVER



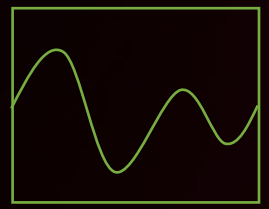
CUTLASS



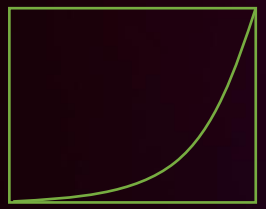
AMGX



cuRAND



cuFFT



CUDA Math API

aws

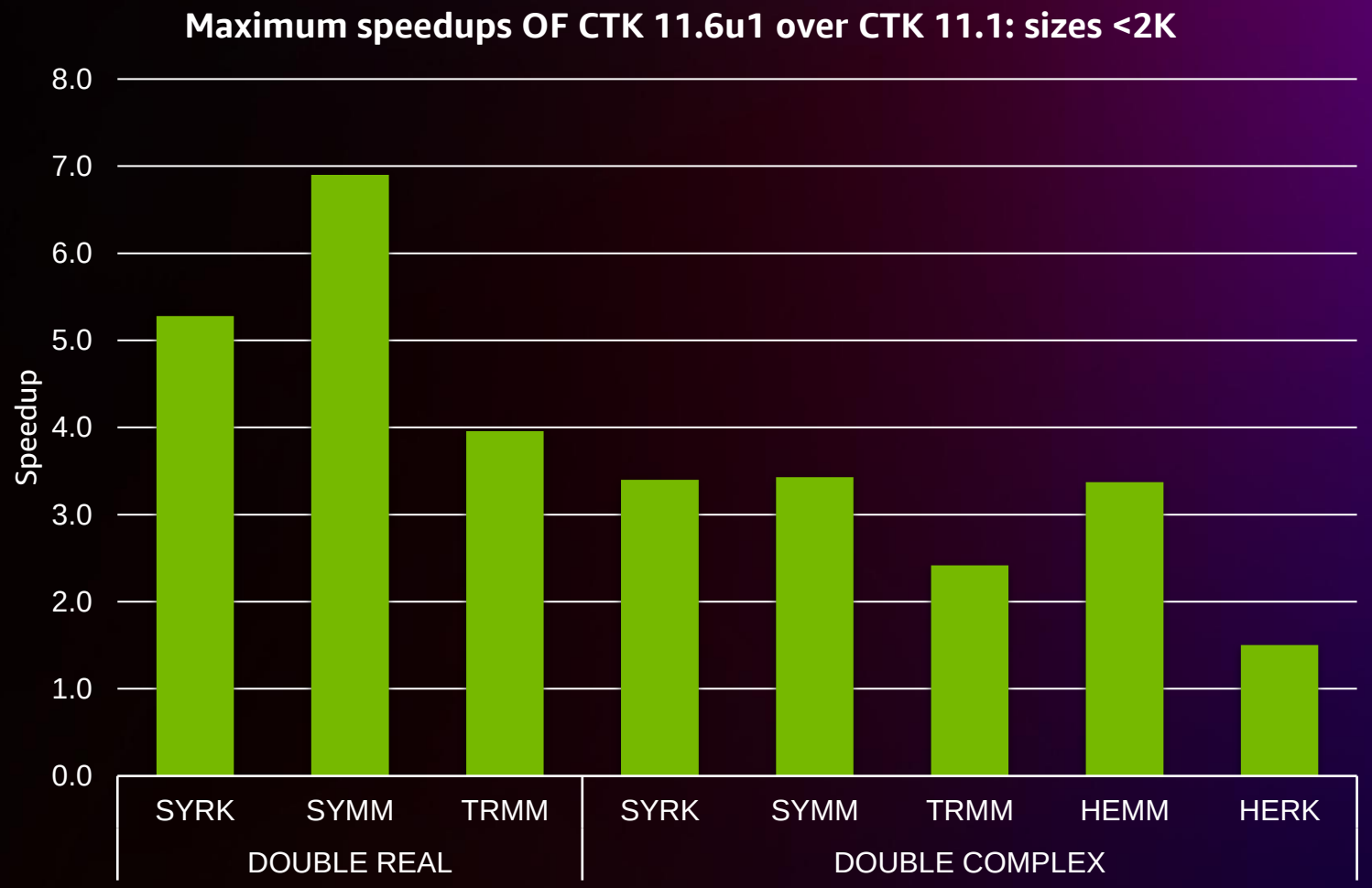
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# cuBLAS

GPU-OPTIMIZED BLAS IMPLEMENTATION

## Full BLAS implementation with extensions

- Vector vector/matrix vector/matrix matrix
- Mixed precision/multiple GPUs/batched APIs




* A100 80 GB @ 1095 MHz: CTK 11.1 vs. CTK 11.6U1

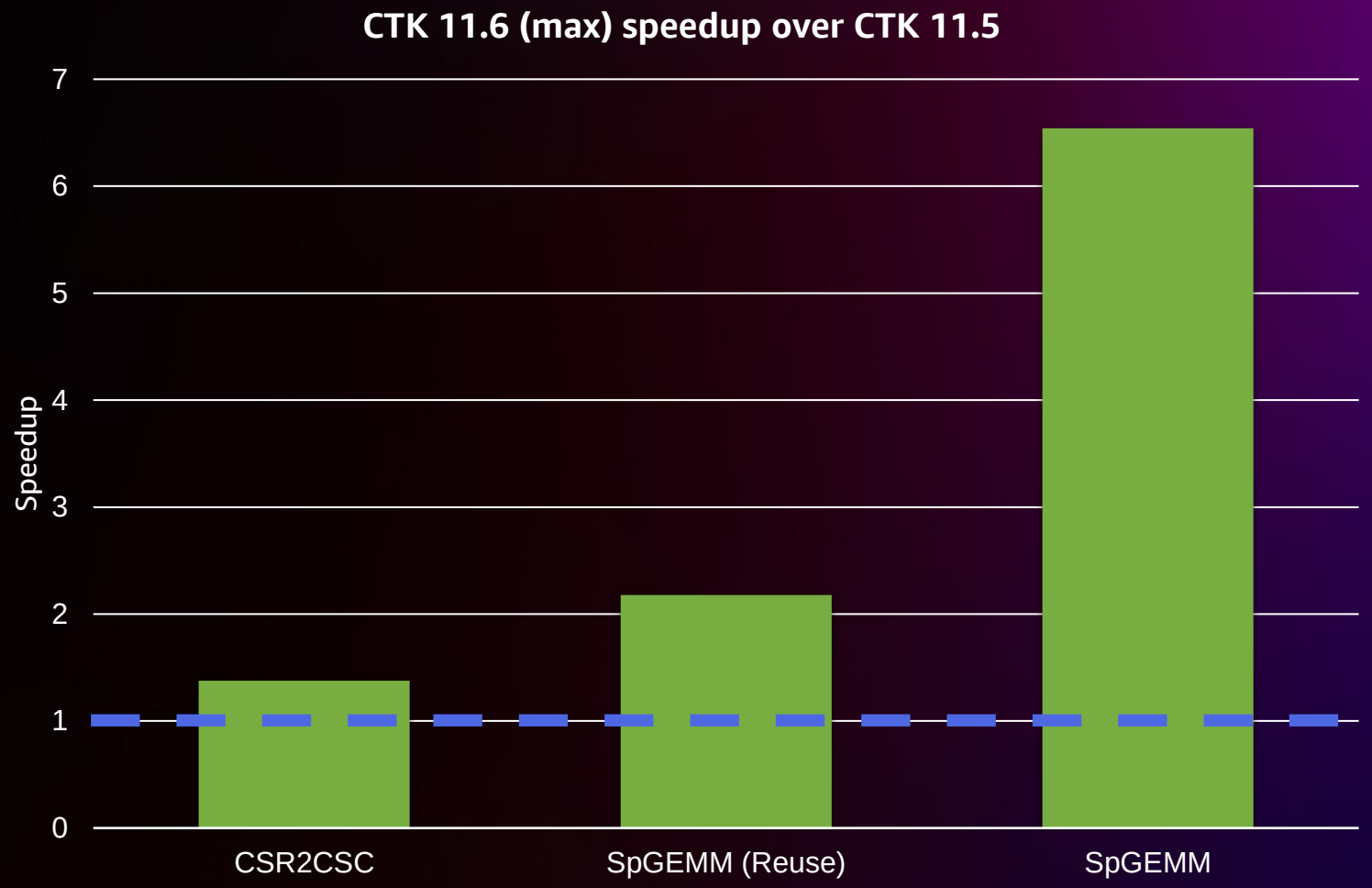
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# cuSPARSE

GPU-OPTIMIZED SPARSE BLAS

## Full sparse BLAS implementation optimized for GPU

- Sparse vector dense vector
- Sparse matrix dense vector
- Sparse matrix dense matrix
- Sparse matrix sparse matrix



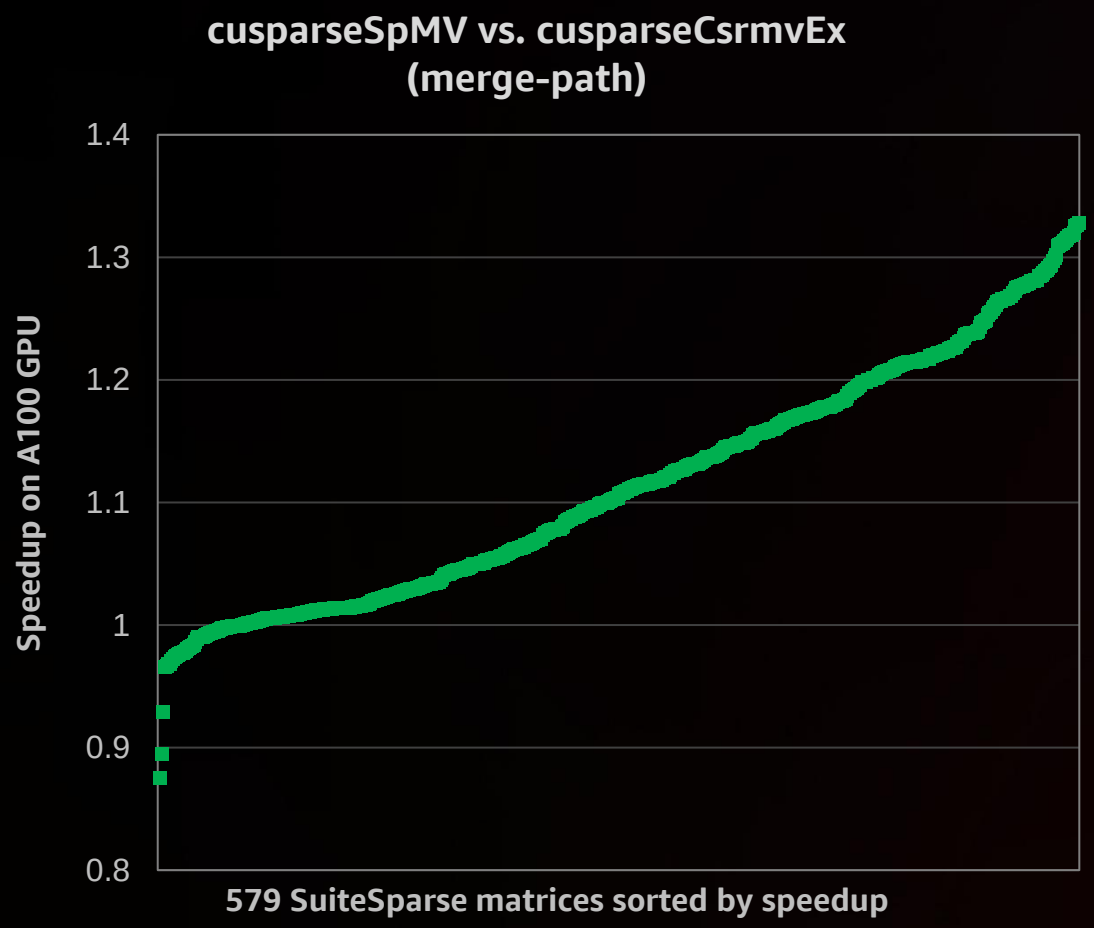

* A100 80 GB @ 825 MHz
* Based on SuiteSparse Matrix Collection

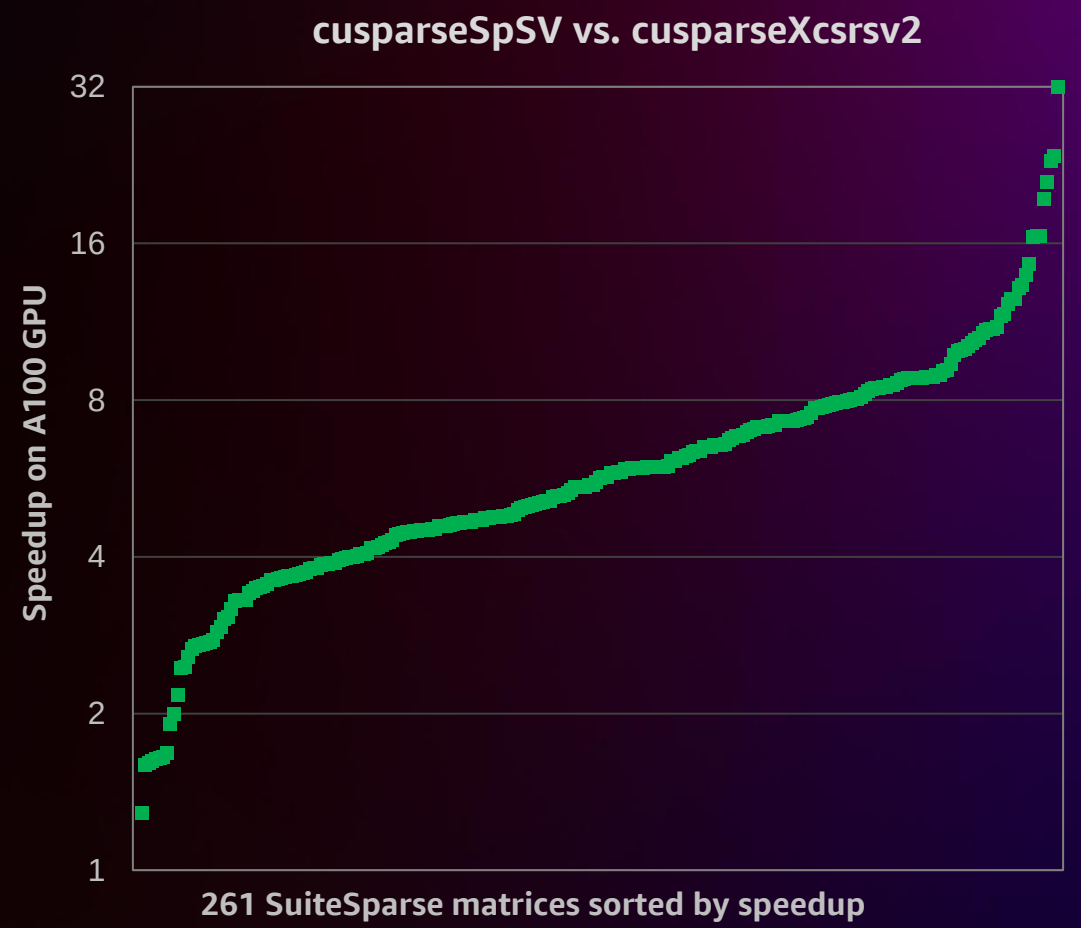
aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# cuSPARSE improvements in key HPC kernels

SPARSE MATRIX X VECTOR AND SPARSE TRIANGULAR SOLVER




aws

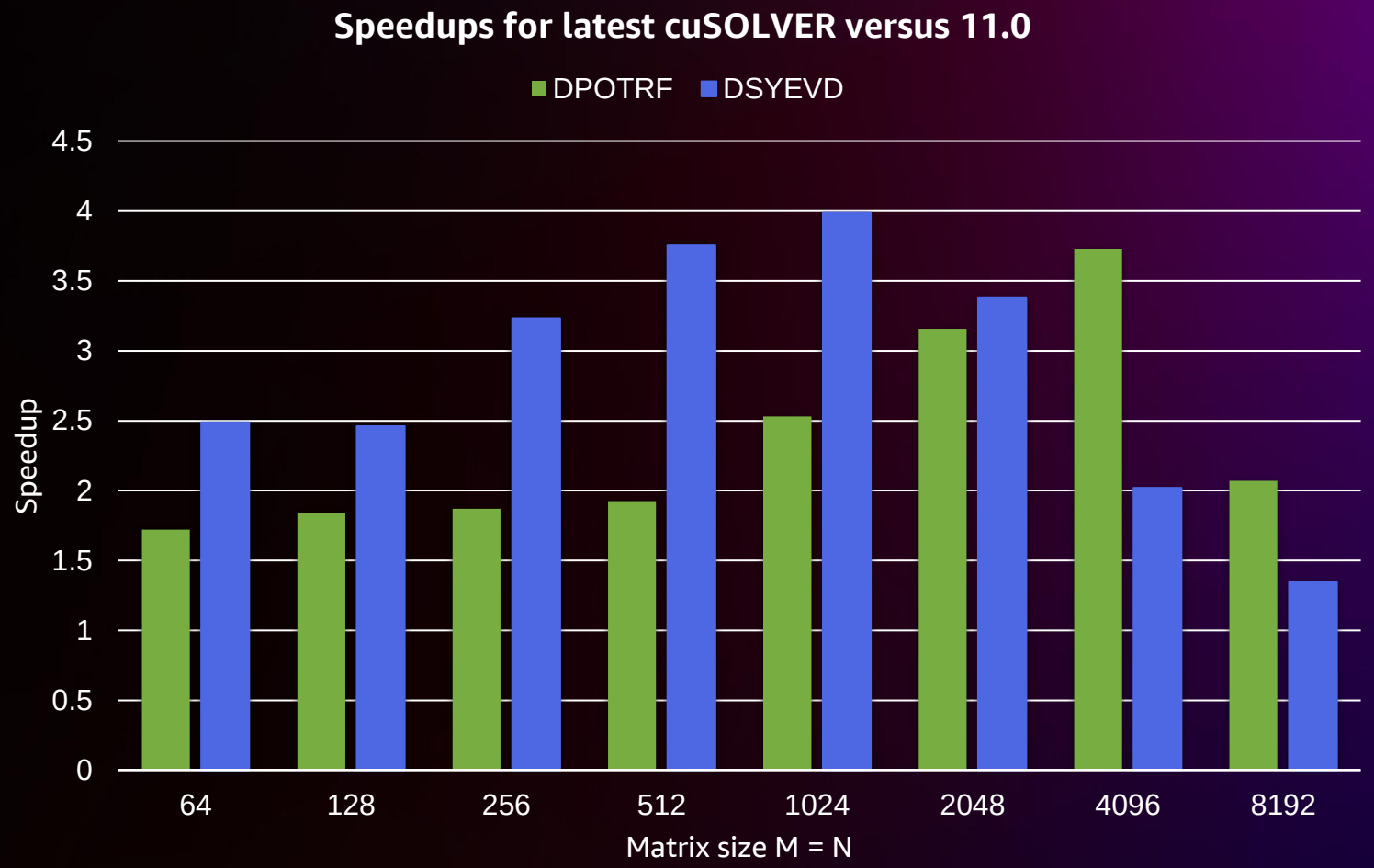
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# cuSOLVER

GPU-OPTIMIZED FACTORIZATIONS AND SOLVERS

## Dense and sparse factorizations and solvers

- LU, Cholesky, QR
- Symmetric and generalized eigensolvers
- Tensor core accelerated iterative refinement solvers
- Multi-GPU and multi-node support




aws

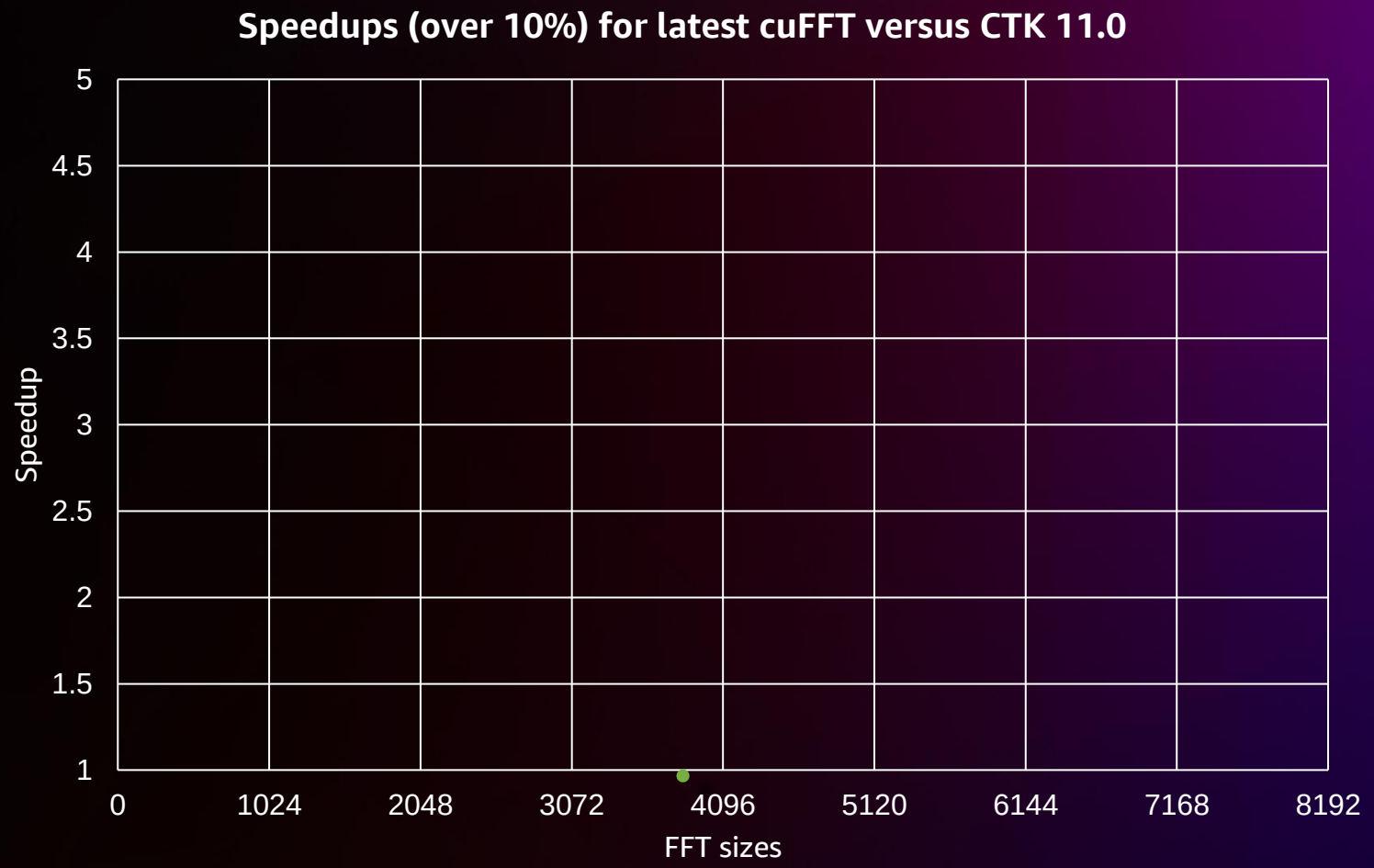
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# cuFFT

GPU-OPTIMIZED FAST FOURIER TRANSFORMS

## FFT optimized for GPU

- 1D, 2D, and 3D FFT
- Single-process multi-GPU support




* A100 80GB Default clocks: CTK 11.0 vs. CTK 11.7U1

aws

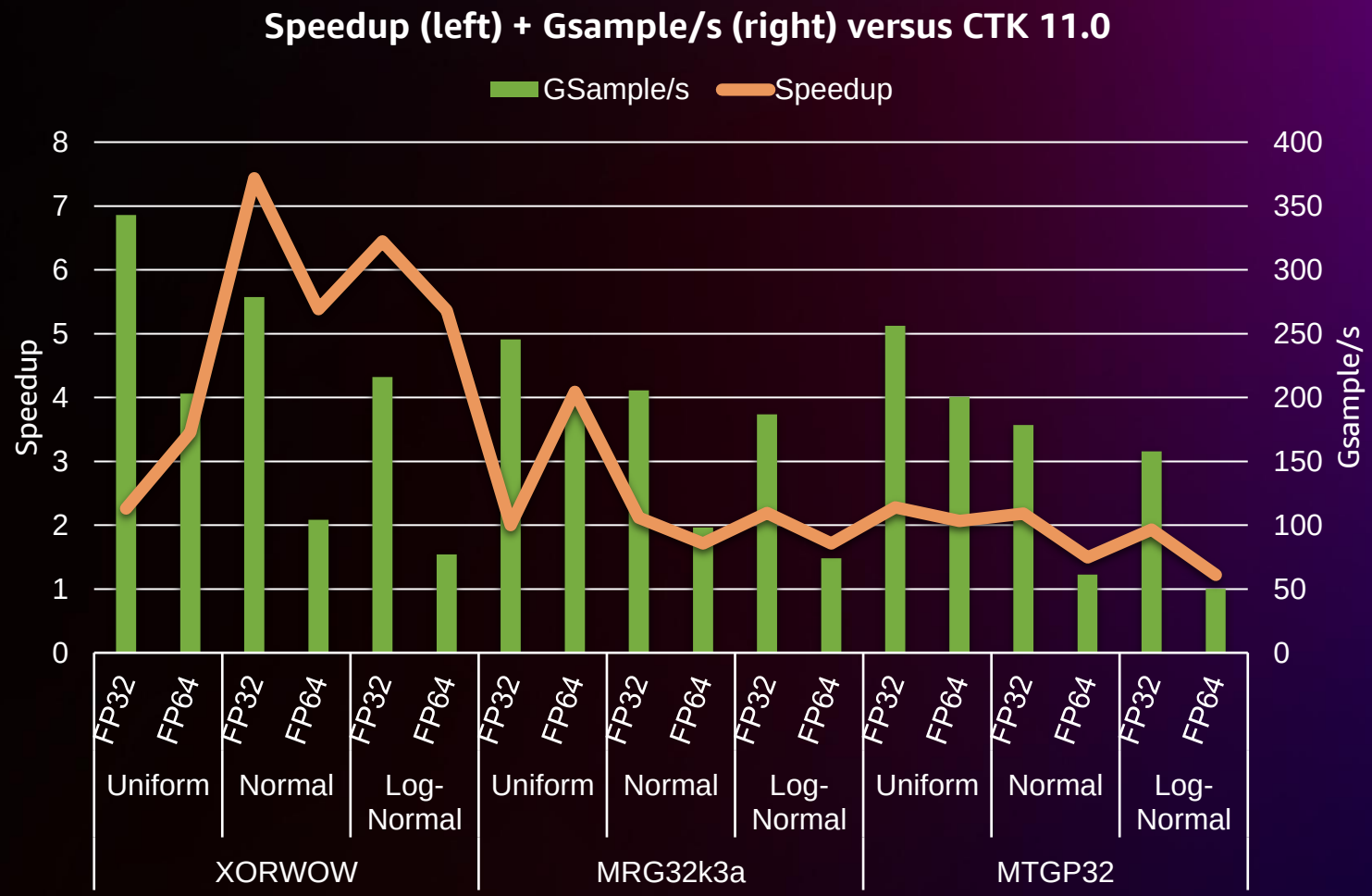
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# cuRAND

RANDOM NUMBER GENERATOR–OPTIMIZED FOR LARGE GPUS

## Random number generators optimized for GPU

- XORWOW/MRG32k3a/MTGP32/ Philox/Sobol
- Host and device APIs
- Uniform/normal/ Log-normal/Poisson




* A100 80 GB @ 1410 MHz: CTK 11.0 vs. CTK 11.6U1

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# cuTENSOR

A HIGH-PERFORMANCE CUDA LIBRARY FOR TENSOR PRIMITIVES

## Tensor primitives optimized for GPU

- Tensor contractions and reductions
- Elementwise operations and elementwise fusion
- Mixed-precision support and tensor core acceleration

## Impact

- DL frameworks aggressively adopting elementwise operations
- Up to 23x application E2E speedup over previously CPU-only quantum chemistry simulations from drop-in contraction API



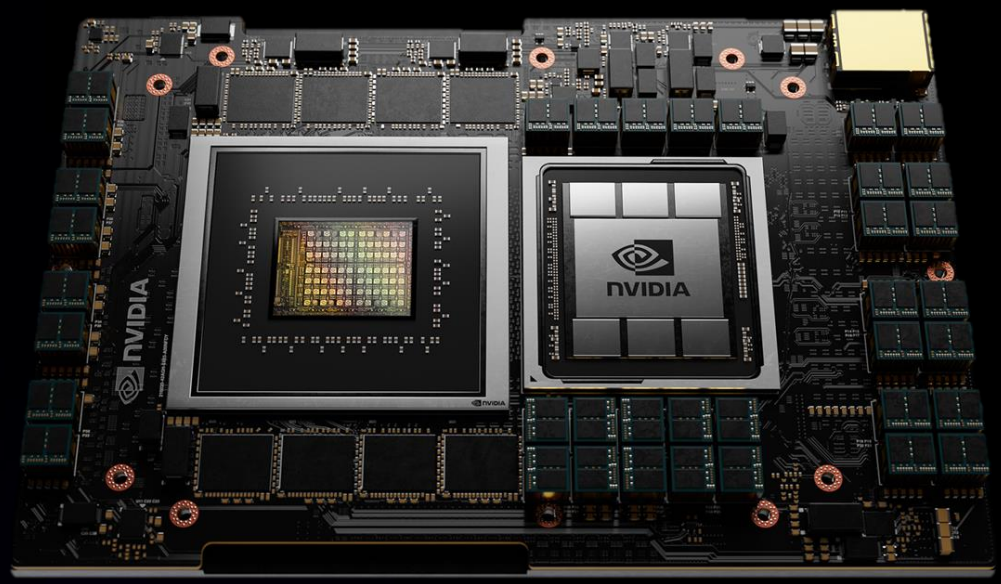

- DGX A100 80GB
- 300 random cases

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Arm CPU math libraries

Coming to Arm CPUs in 2023

- BLAS
- FFT
- LAPACK and ScaLAPACK
- Sparse BLAS
- Tensors
- RNGs

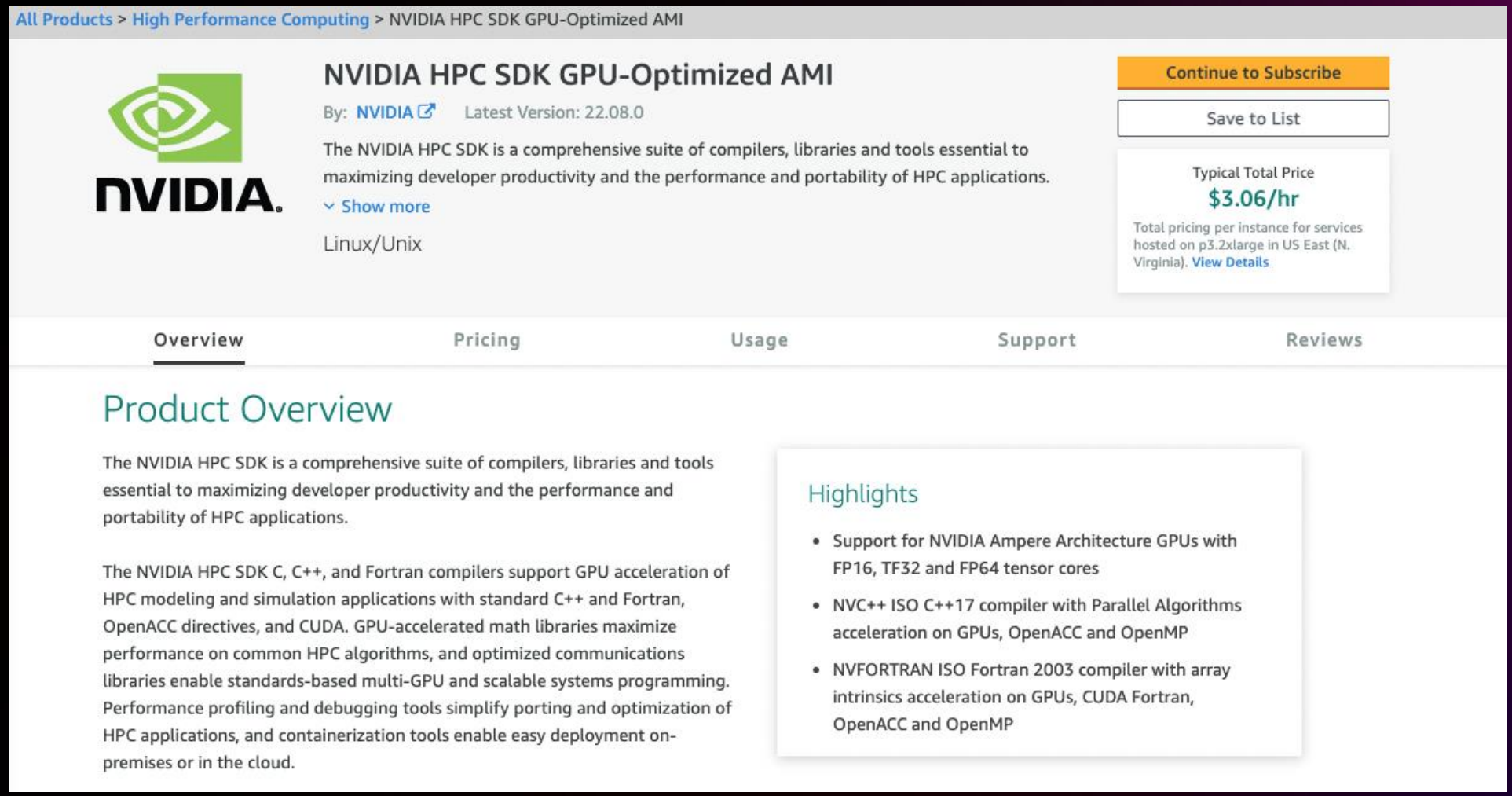



aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# NVIDIA HPC SDK on Amazon EC2

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

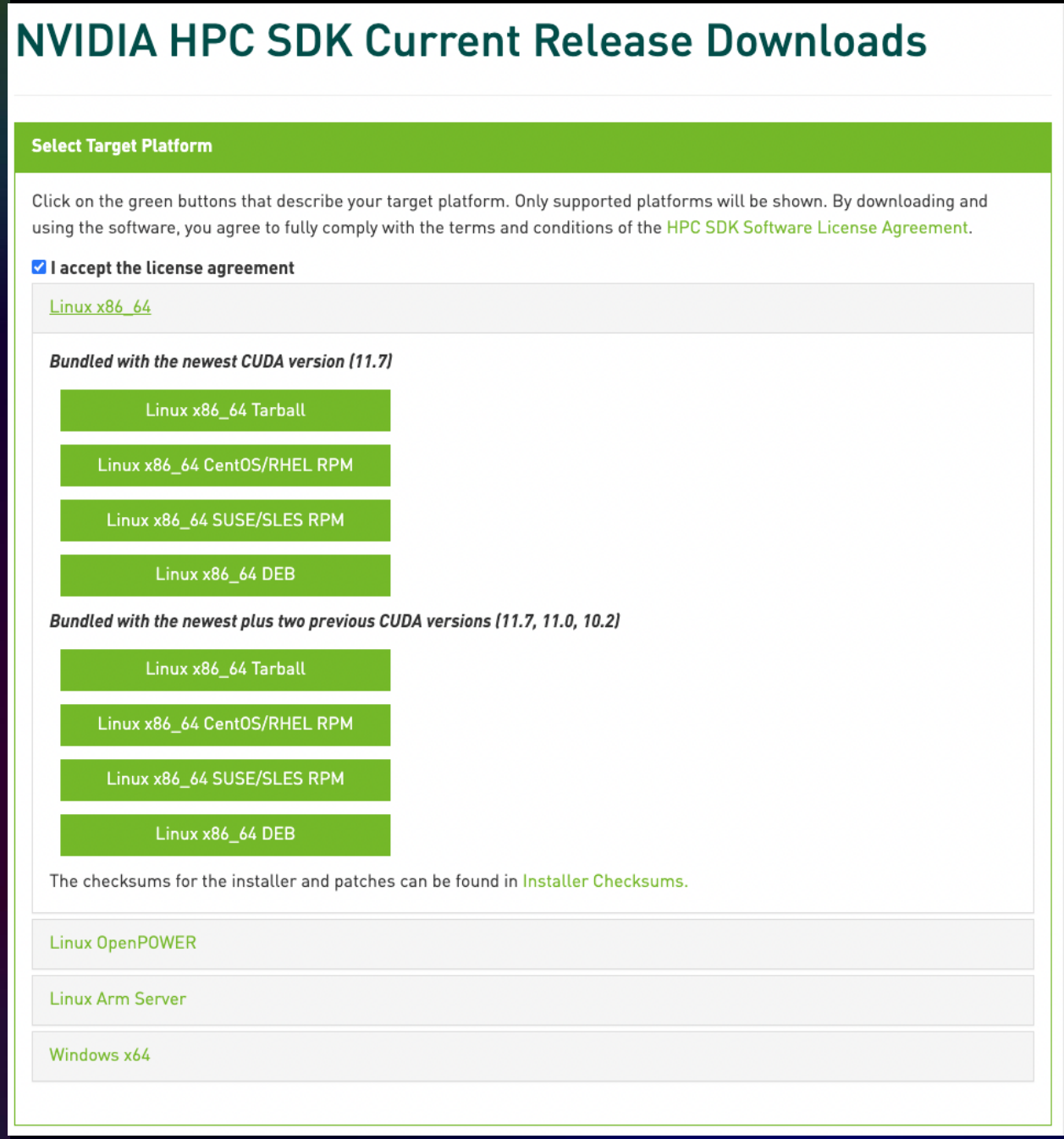
# NVIDIA HPC SDK in the AWS Marketplace




© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Easy to download and install on Amazon EC2 instances

developer.nvidia.com/hpc-sdk




aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Installation on Amazon Linux 2

```
sudo yum-config-manager --add-repo
https://developer.download.nvidia.co
m/hpc-sdk/rhel/nvhpc.repo

sudo yum install -y nvhpc-22.7

# Note: also install GPU drivers if
a GPU is available
```

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

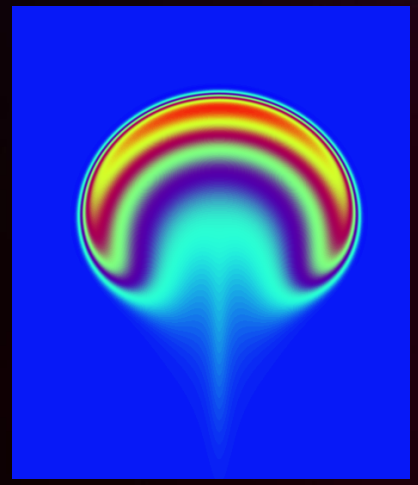
NVIDIA HPC SDK on AWS Graviton 3

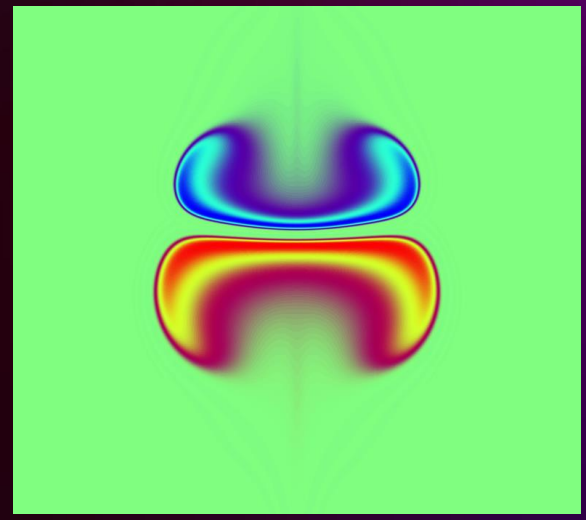
```
The default interactive shell is now zsh.
To update your account to use zsh, please run `chsh -s /bin/zsh`.
For more details, please visit https://support.apple.com/kb/HT208050.
jlinford-mlt:x jlinford$
```

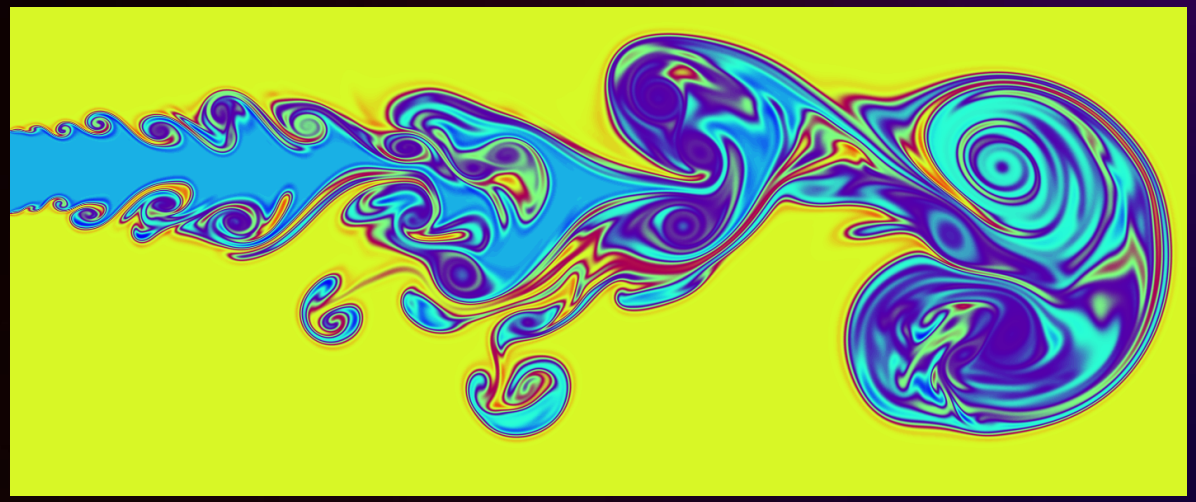
# MiniWeather

STANDARD LANGUAGE PARALLELISM IN CLIMATE/WEATHER APPLICATIONS

- C++ and Fortran
- Simulates weather-like fluid flows using the finite volume and Runge-Kutta methods
- Used in the SPEChpc benchmark suite
- Open source and commonly used in training events







aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# MiniWeather

STANDARD LANGUAGE PARALLELISM IN CLIMATE/WEATHER APPLICATIONS

- Parallelized via MPI, OpenMP, OpenACC, and StdPar
- OpenMP 4.5 target offload
- Fortran DO CONCURRENT

```
do concurrent (ll=1:NUM_VARS, k=1:nz, i=1:nx)
      local(x,z,x0,z0,xrad,zrad,amp,dist,wpert)

  if (data_spec_int == DATA_SPEC_GRAVITY_WAVES) then
    x = (i_beg-1 + i-0.5_rp) * dx
    z = (k_beg-1 + k-0.5_rp) * dz
    x0 = xlen/8
    z0 = 1000
    xrad = 500
    zrad = 500
    amp = 0.01_rp
    dist = sqrt( ((x-x0)/xrad)**2 + ((z-z0)/zrad)**2 )
         * pi / 2._rp
    if (dist <= pi / 2._rp) then
      wpert = amp * cos(dist)**2
    else
      wpert = 0._rp
    endif
    tend(i,k,ID_WMOM) = tend(i,k,ID_WMOM)
                      + wpert*hy_dens_cell(k)
  endif
  state_out(i,k,ll) = state_init(i,k,ll) + dt * tend(i,k,ll)

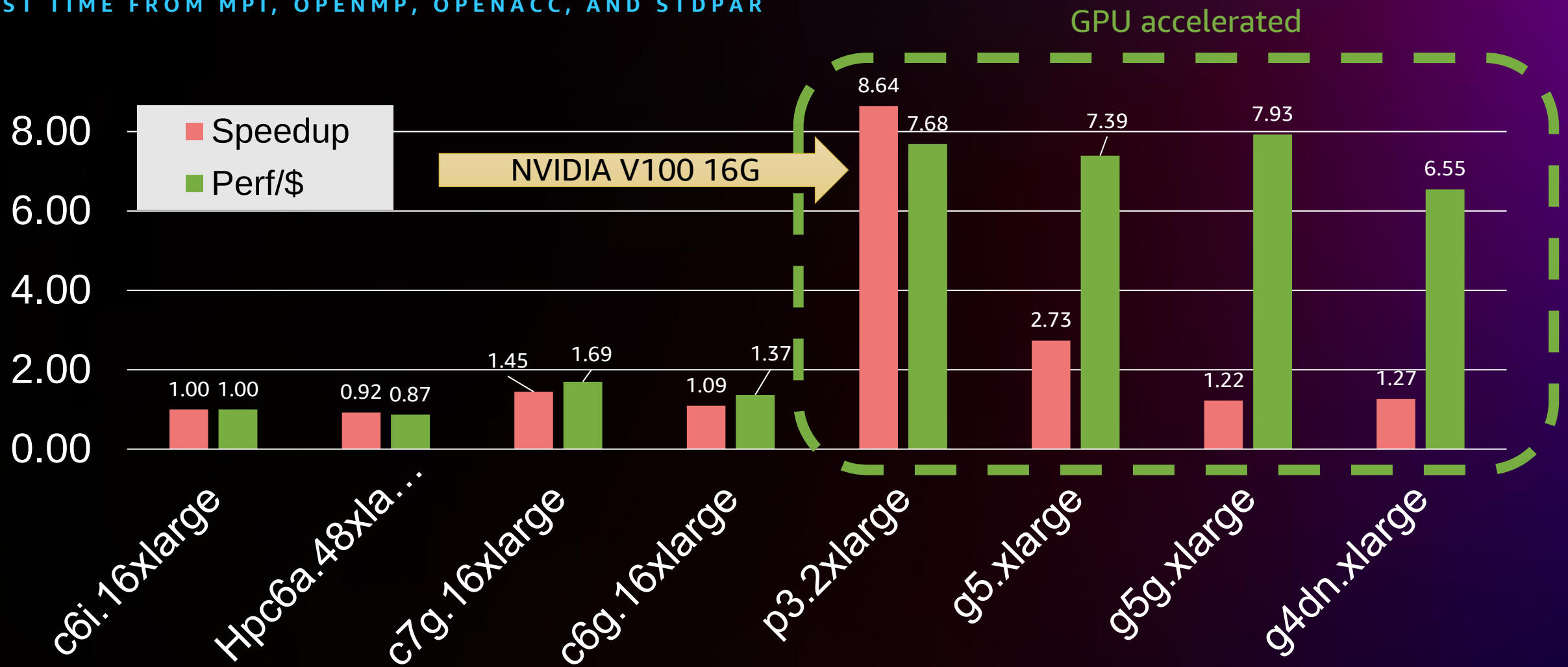
enddo
```

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# MiniWeather with NVIDIA HPC SDK

BEST TIME FROM MPI, OPENMP, OPENACC, AND STDPAR

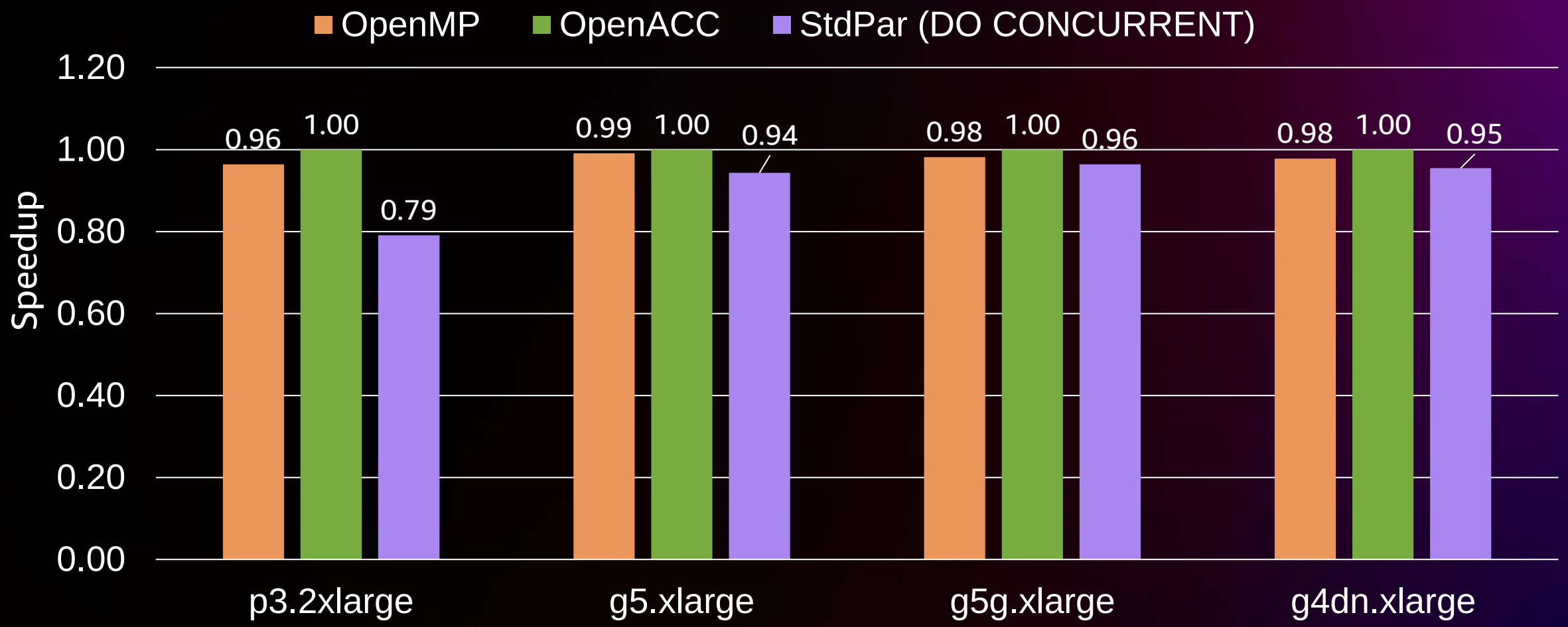



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

HPC SDK 22.7, MiniWeather: NX=2000, NZ=1000, SIM_TIME=5

# ISO Fortran MiniWeather with NVIDIA HPC SDK

ISO FORTRAN CODE CAN RUN ON A CPU OR A GPU




© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

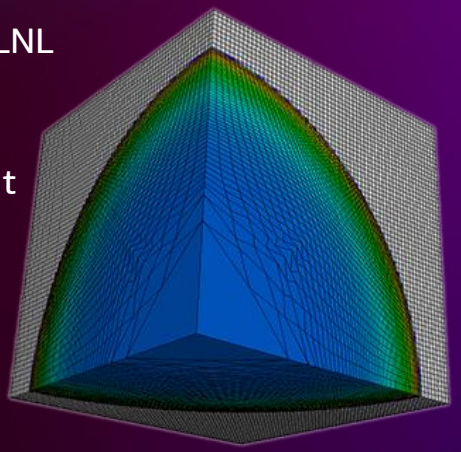
*HPC SDK 22.7, MiniWeather: NX=2000, NZ=1000, SIM_TIME=5*

# LULESH with standard C++

C++ hydrodynamics mini-application from LLNL

Rewritten from OpenMP to ISO C++

- More composable, compact, and elegant
- Easier to read and maintain
- ISO standard
- Portable – nvc++, g++, icpc, MSVC, . . .



```
static inline
void CalcHydroConstraintForElems(Domain &domain, Index_t length,
    Index_t *regElemlist, Real_t dvovmax, Real_t& dthydro)
{
#if _OPENMP
  const Index_t threads = omp_get_max_threads();
  Index_t hydro_elem_per_thread[threads];
  Real_t dthydro_per_thread[threads];
#else
  Index_t threads = 1;
  Index_t hydro_elem_per_thread[1];
  Real_t dthydro_per_thread[1];
#endif
#pragma omp parallel firstprivate(length, dvovmax)
  {
    Real_t dthydro_tmp = dthydro ;
    Index_t hydro_elem = -1 ;
#if _OPENMP
    Index_t thread_num = omp_get_thread_num();
#else
    Index_t thread_num = 0;
#endif
#pragma omp for
    for (Index_t i = 0 ; i < length ; ++i) {
      Index_t indx = regElemlist[i] ;

      if (domain.vdov(indx) != Real_t(0.)) {
        Real_t dtdvov = dvovmax / (FABS(domain.vdov(indx))+Real_t(1.e-20)) ;

        if ( dthydro_tmp > dtdvov ) {
          dthydro_tmp = dtdvov ;
          hydro_elem = indx ;
        }
      }
    }
    dthydro_per_thread[thread_num] = dthydro_tmp ;
    hydro_elem_per_thread[thread_num] = hydro_elem ;
  }
  for (Index_t i = 1; i < threads; ++i) {
    if(dthydro_per_thread[i] < dthydro_per_thread[0]) {
      dthydro_per_thread[0] = dthydro_per_thread[i];
      hydro_elem_per_thread[0] = hydro_elem_per_thread[i];
    }
  }
  if (hydro_elem_per_thread[0] != -1) {
    dthydro = dthydro_per_thread[0] ;
  }
  return ;
}
```

C++ with OpenMP

```
static inline void CalcHydroConstraintForElems(Domain &domain, Index_t l
ength,
                Index_t *regElemlist,
                Real_t dvovmax,
                Real_t &dthydro)
{
  dthydro = std::transform_reduce(
    std::execution::par, counting_iterator(0), counting_iterator(length)
,
    dthydro, [](Real_t a, Real_t b) { return a < b ? a : b; },
    [=, &domain](Index_t i)
  {
      Index_t indx = regElemlist[i];
      if (domain.vdov(indx) == Real_t(0.0)) {
        return std::numeric_limits<Real_t>::max();
      } else {
        return dvovmax / (std::abs(domain.vdov(indx)) + Real_t(1.e-20));
      }
  });
}
```

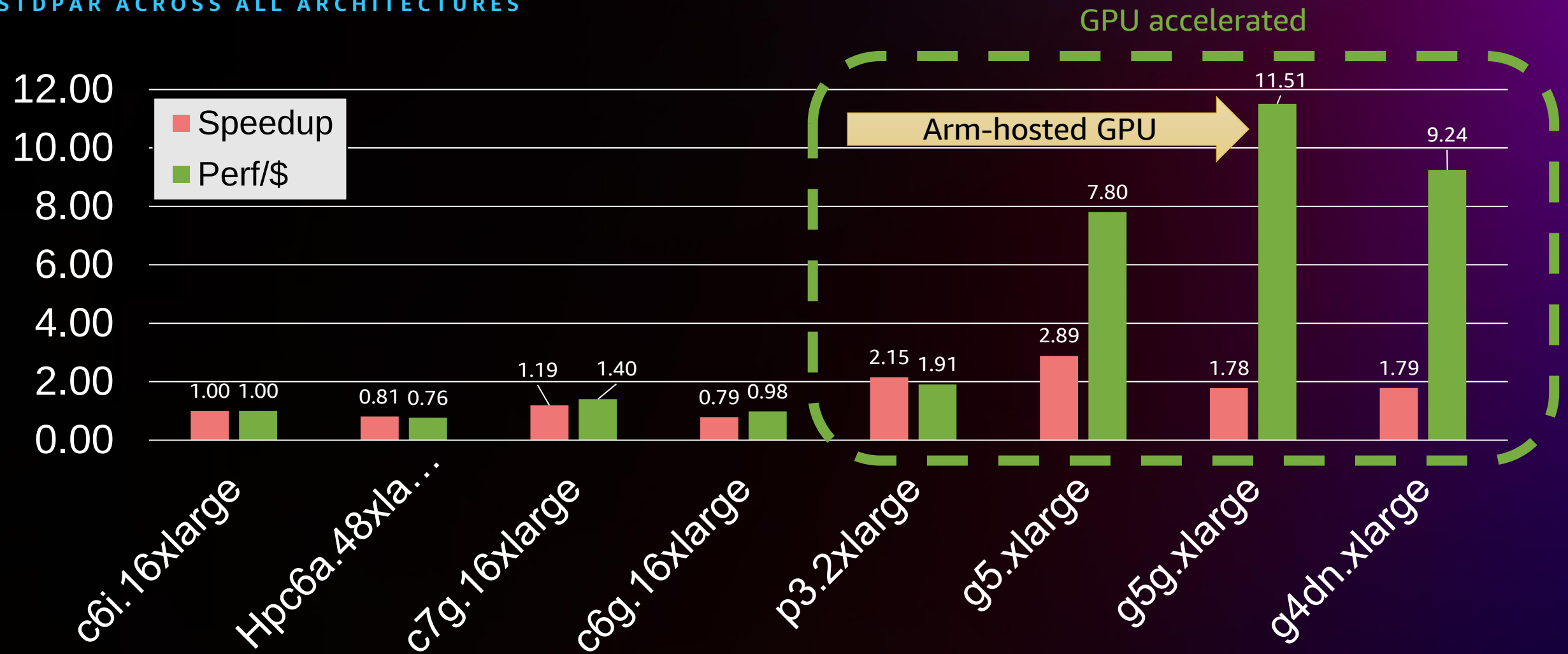
Standard C++

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

codesign.llnl.gov/lulesh

# ISO C++ LULESH with NVIDIA HPC SDK

STDPAR ACROSS ALL ARCHITECTURES




aws

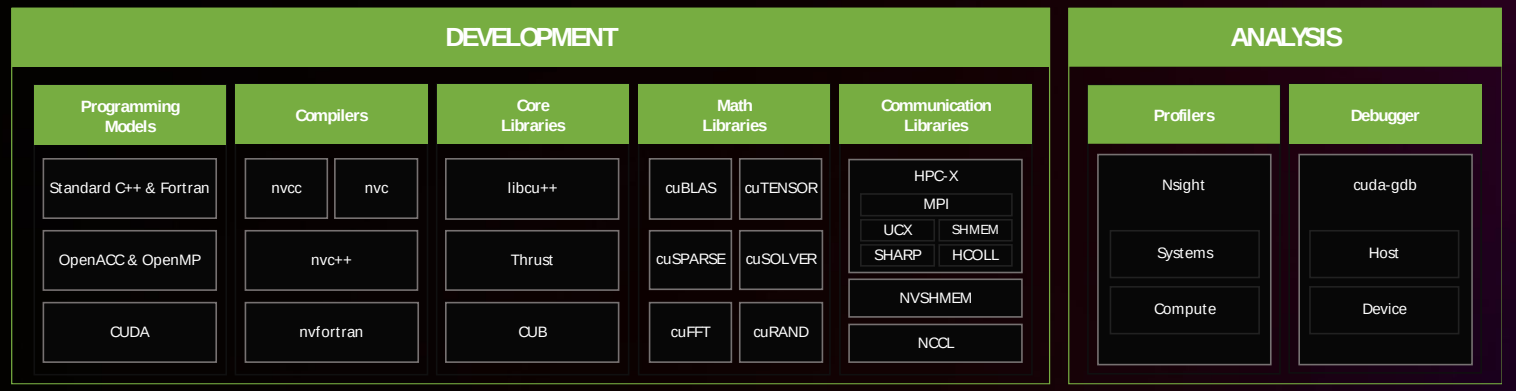
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Wrap-up

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# developer.nvidia.com/hpc-sdk

- **NVIDIA HPC SDK** is a complete solution for HPC on Amazon EC2
  - **Productivity:** full support for all HPC programming models on CPU or GPU
  - **Performance:** compilers, math libraries, and tools for performance optimization
  - **Portability:** target Amazon EC2 instances built on Arm or x86 with or without GPUs
- NVIDIA HPC SDK is available in AWS Marketplace
- Freely available and trivial to install on any Amazon EC2 instance




© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Join the NVIDIA Inception Program for Startups

Accelerate your startup's growth and build your solutions faster with engineering guidance, free technical training, and preferred pricing on NVIDIA products, opportunities for customer introductions and co-marketing, and exposure to the VC community.



**APPLY TO INCEPTION TODAY**
**https://www.nvidia.com/en-us/startups**



**GET THE LATEST NEWS, UPDATES, AND MORE**
**https://developer.nvidia.com/developer-program**

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Thank you!



Please complete the session survey in the **mobile app**

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.