
NVIDIA
TECHNOLOGICAL UPDATE
SEPT 2022

NVIDIA H100

NVIDIA H100

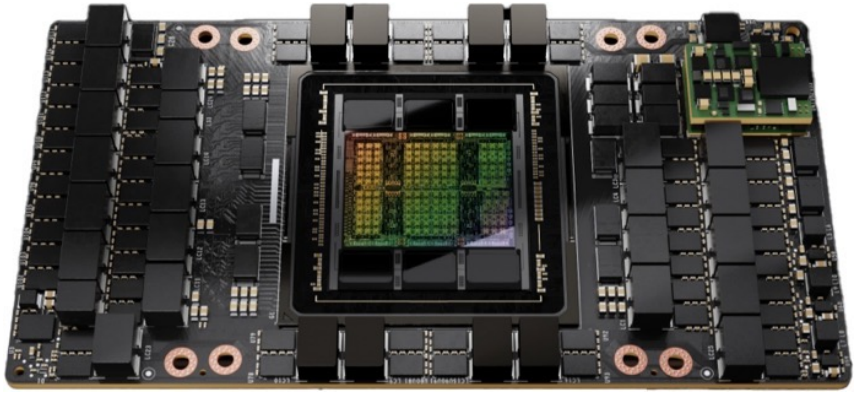
Unprecedented Performance, Scalability, and Security for Every Data Center

HIGHEST AI AND HPC PERFORMANCE
4PF FP8 (6X)| 2PF FP16 (3X)| 1PF TF32 (3X)| 60TF FP64 (3X)
3TB/s (1.5X), 80GB HBM3 memory

TRANFORMER MODEL OPTIMIZATIONS
6X faster on largest transformer models

HIGHEST UTILIZATION EFFICIENCY AND SECURITY
7 Fully isolated & secured instances, guaranteed QoS
2nd Gen MIG | Confidential Computing

FASTEST, SCALABLE INTERCONNECT
900 GB/s GPU-2-GPU connectivity (1.5X)
up to 256 GPUs with NVLink Switch | 128GB/s PCI Gen5

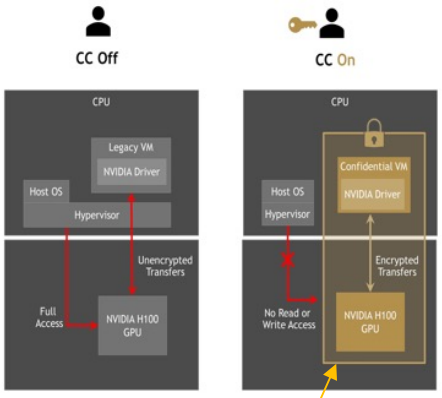


FP8, FP16, TF32 performance include sparsity. X-factor compared to A100

HOPPER TECHNOLOGICAL BREAKTHROUGHS

CONFIDENTIAL COMPUTING

Secure Data and AI Models In-Use



Confidential VM
(virtualization-based TEE)

Secure during compute, not just in storage or in motion

MULTI-GPU INSTANCE

7 Secure Tenants on 1 GPU

New vs A100:

- *6x compute, more mem bandwidth*
- *Dedicated NVDEC & NVJPG per MIG*
- *Each instance has its own telemetry*
- *Secure confidential compute MIG*

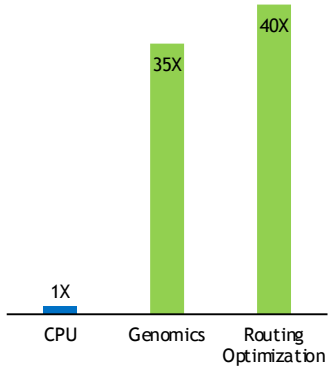
3X Compute
1.5X Bandwidth

NEW DYNAMIC PROGRAMING INSTRUCTIONS

Accelerate Dynamic Programming Algorithms

A BROAD RANGE OF USE CASES

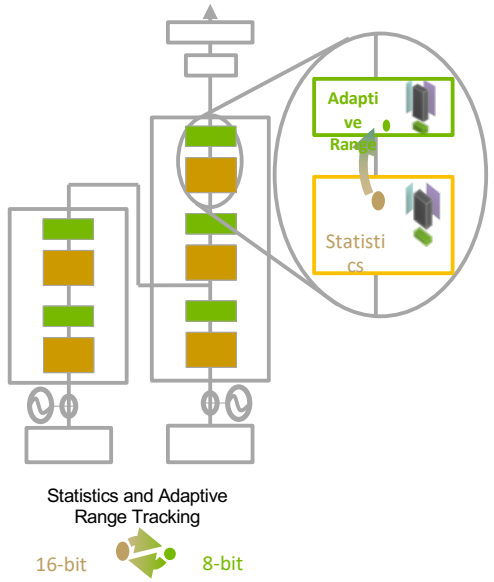
REAL-TIME PERFORMANCE



40X vs CPU
7X vs A100

TRANSFORMER ENGINE

Tensor Core Optimized for Transformer Models



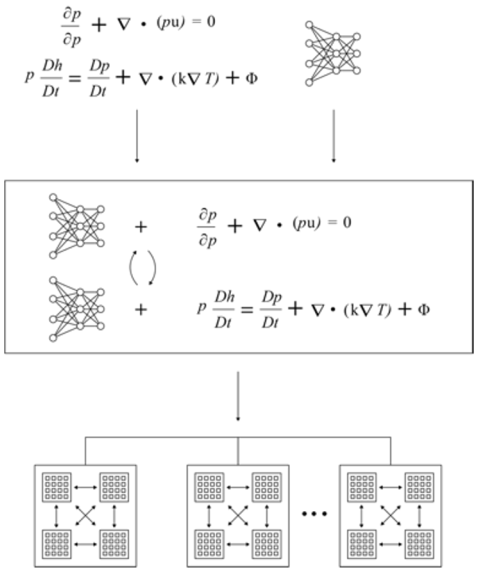
6X on LLM

Dynamic Programming Instructions HGX H100 4-GPU vs dual socket 32 core IceLake. HGX H100 performance projections.

NVIDIA MODULUS

Physics Machine Learning Platform

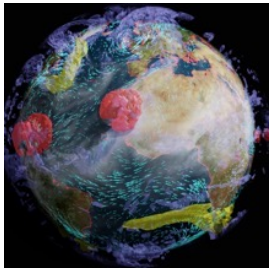
TRAINING NEURAL NETWORKS USING BOTH DATA AND THE GOVERNING EQUATIONS



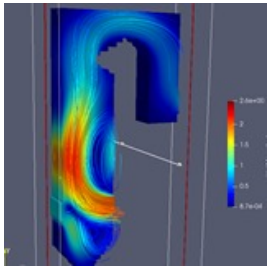
ADVANCING SCIENTIFIC DISCOVERY WITH MODULUS

RENEWABLE ENERGY
Siemens Gamesa: Up to 4000X Speedup of Wind Turbine Wake Optimization

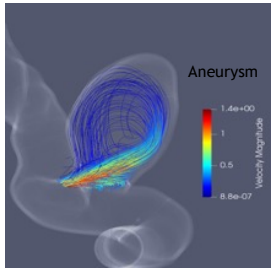
CLIMATE CHANGE
45,000X Speedup of Extreme weather Prediction with FourCastNet



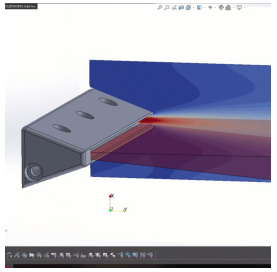
INDUSTRIAL HPC
NETL: 10,000X Faster Build Of high-fidelity surrogate models



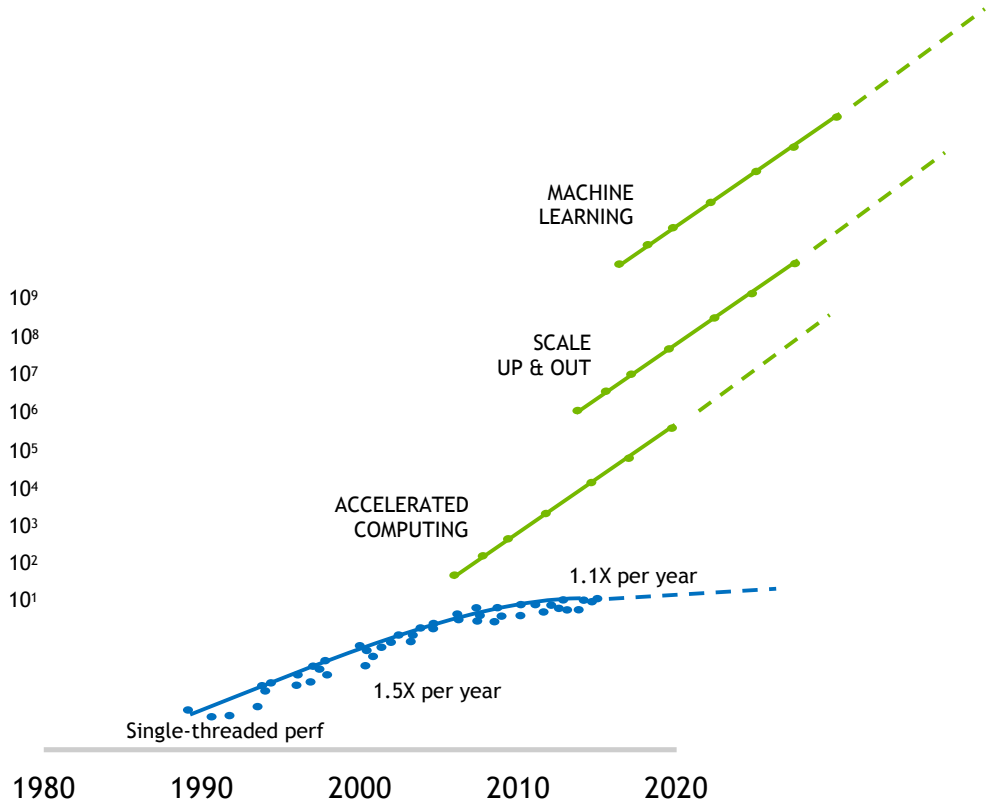
HEALTHCARE
Achieve high-fidelity results faster for blood flow in inter-cranial aneurysm

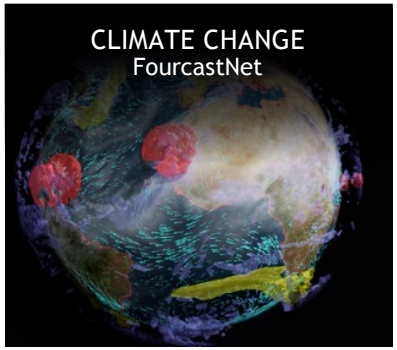


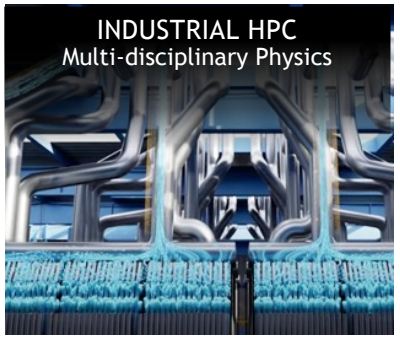
DIGITAL TWINS
Kinetic Vision: Design Optimization Using parameterized models



MILLION-X SPEEDUP FOR INNOVATION AND DISCOVERY





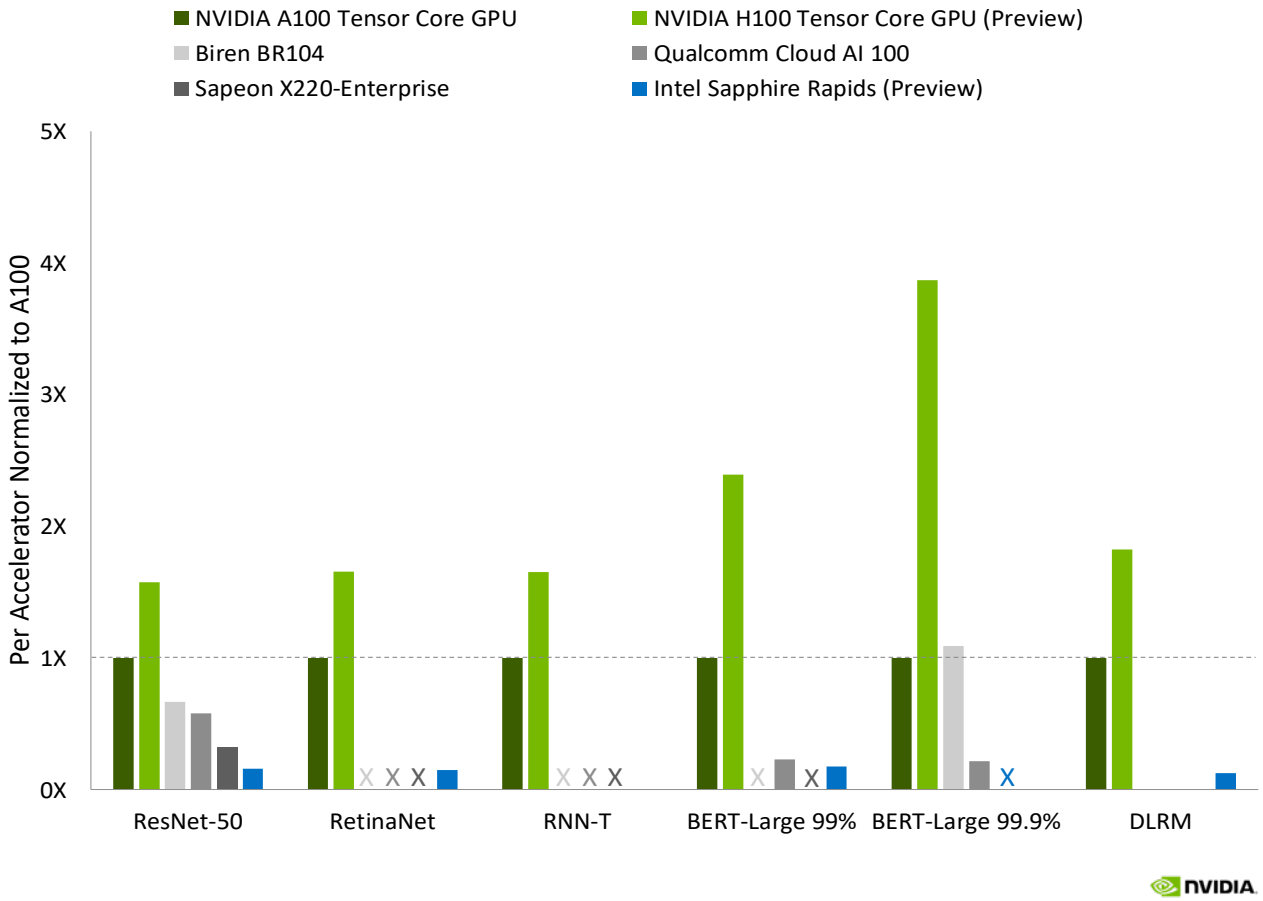


NVIDIA

H100 SUPERCHARGES NVIDIA AI

NVIDIA Furthers AI Inference Performance Leadership

- H100 Tops All Data Center Tests Up to 4.5X Higher Performance than A100
- 12 NVIDIA Partners Submitted
- A100 still delivery good performance and has improved by 6x since June 2020 thanks to software enhancements



GTC SESSION TO FOLLOW

GTC22 Fall

- A Deep Dive into the Latest HPC Software [A41133]
- CUDA: New Features and Beyond [A41100]
- Developing HPC Applications with Standard C++, Fortran, and Python [A41087]

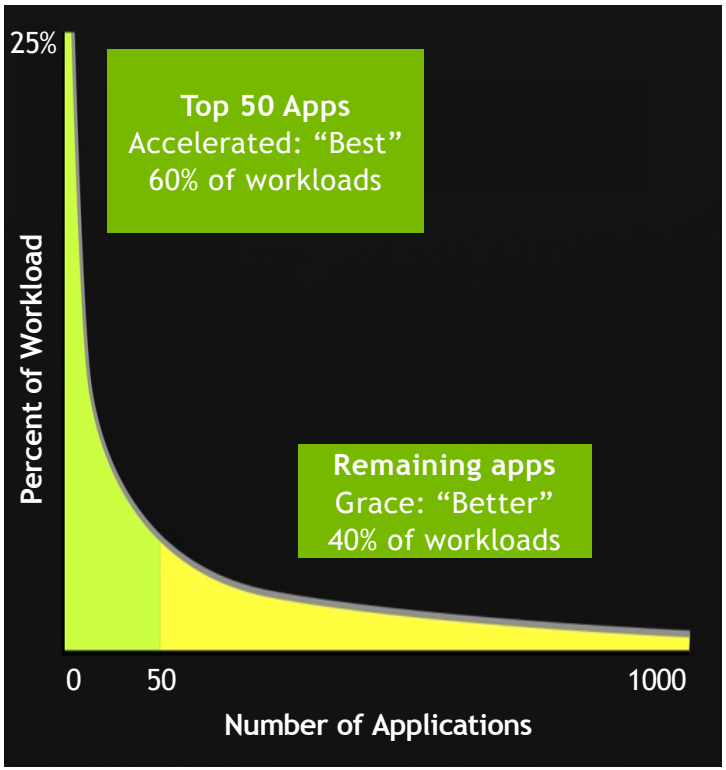
GTC22 Spring

- C++ Standard Parallelism [S41960]
- Future of Standard and CUDA C++ [S41961]
- Shifting through the Gears of GPU Programming: Understanding Performance and Portability Trade-offs [S41620]
- From Directives to DO CONCURRENT: A Case Study in Standard Parallelism [S41318]
- Evaluating Your Options for Accelerated Numerical Computing in Pure Python [S41645]
- How to Develop Performance Portable Codes using the Latest Parallel Programming Standards [S41618]

GENERAL PURPOSE MEETS HIGH PERFORMANCE

GRACE HELPS TO FILL THE GAP OF THE NOT-YET-ACCELERATED FRONTIER

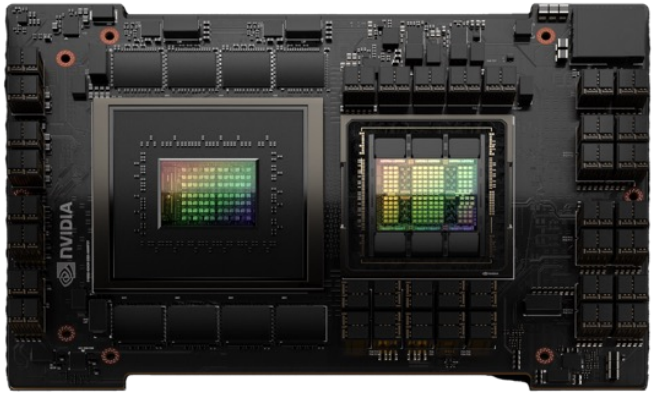
- There are over **1 billion lines of FORTRAN** in **HPC workloads today.**
  - *Fortran was first released in 1957.*
- The vast majority of applications **can** be accelerated.
  - …and a lot of what your customers run already is.
- Most **will** be accelerated ***someday.***
- Grace provides an NVIDIA solution for ***every*** HPC workload **today.**



NVIDIA GRACE PLATFORM

Grace Hopper Superchip

Giant Scale AI & HPC



Accelerated applications where CPU performance and system memory BW are critical since AI models continue to get bigger and our GPUs get even faster

Grace CPU Superchip

CPU Computing

Applications that are not accelerated yet but where absolute performance, energy efficiency, and datacenter density matter, such as in scientific computing, data analytics, and hyperscale computing applications

GRACE CPU SUPERCHIP

HIGHER HPC PERFORMANCE AT FRACTION OF POWER

Applications that are not accelerated yet but where absolute performance, energy efficiency, and datacenter density matter, such as in scientific computing, data analytics, and hyperscale computing applications

Specifications	Grace SuperChip
Architecture	Armv9, SVE2 with 4x 128b pipeline/core
Cores / Speed	144 cores
Memory	LPDDR5x soldered down, 1TB/s BW Up to 1TB per superchip
Cache	L1: 64KB i-cache + 64KB d-cache per core L2: 1MB per core L3: 240MB per superchip
Power	500W including LPDDR5x memory
Interfaces	Up to 8x PCIe Gen5 x16 HS interface
Process Node	TSMC 4N

GRACE SIMPLIFIES BUILDING AND RUNNING COMPUTE INFRASTRUCTURE

GRACE SUPERCHIP

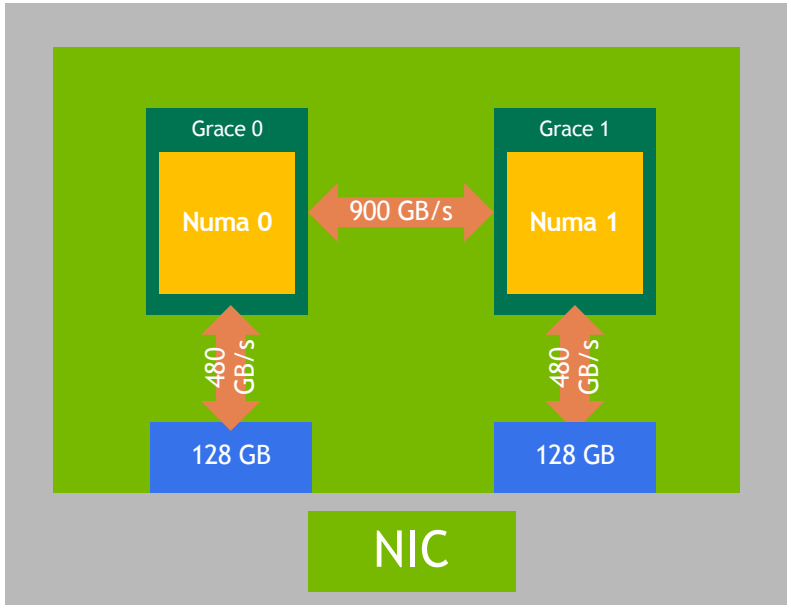
144 Cores

3Ghz Freq

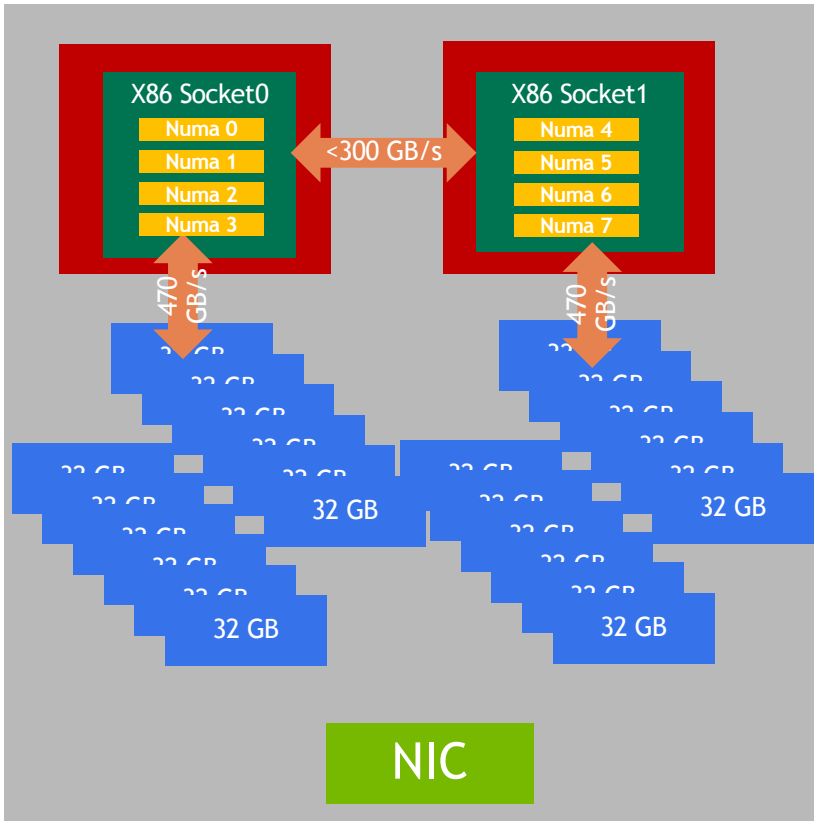
256GB RAM

500W CPU+MEM

2 NUMA



X86 2-SOCKET SERVER



192 Cores

2Ghz Freq

768GB RAM

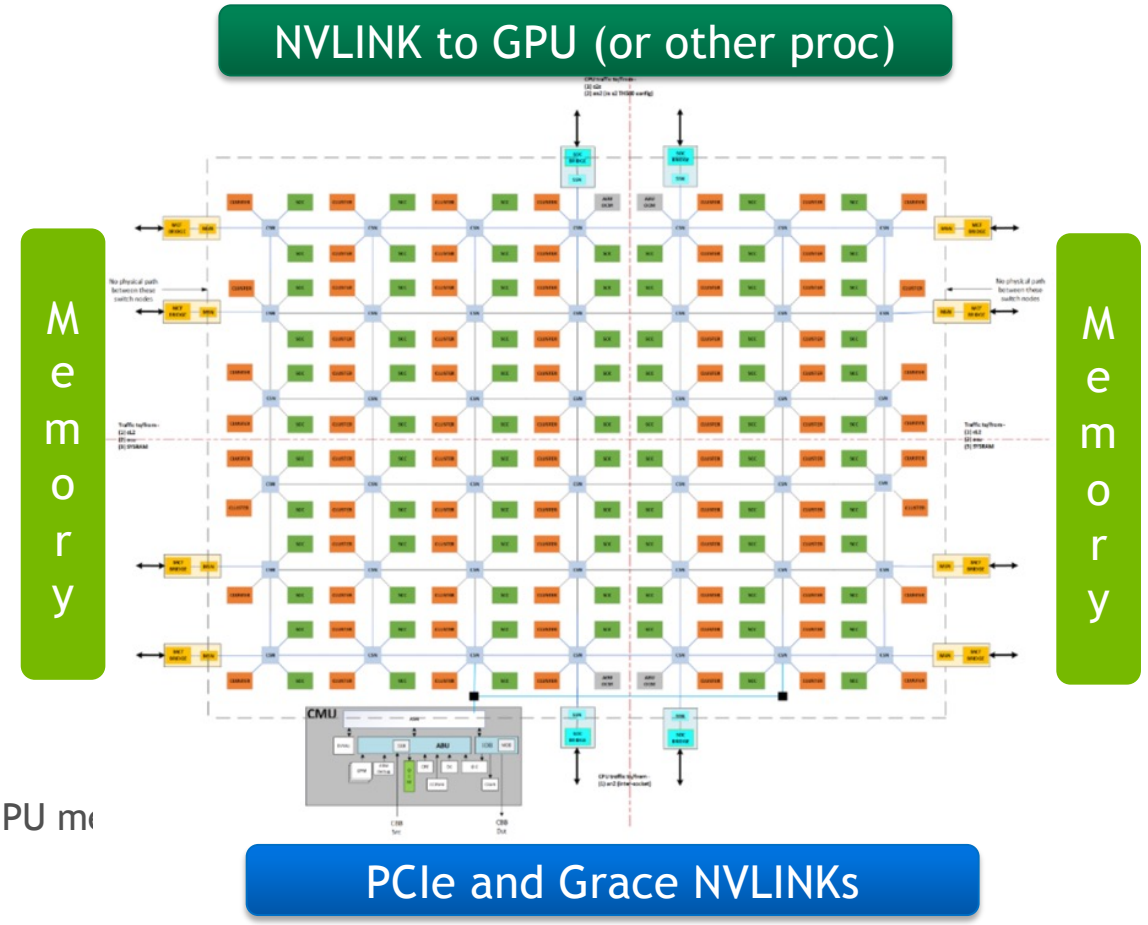
900W CPU+MEM

8+ NUMA

PROCESSOR SOC/NOC

Augmented custom ISA logic to support memory movement

- Monolithic SoC
  - Up to 120MB shared L3 cache
  - 3TB/s on-die mesh bisection BW
- Extensive set of Core and un-Core perf counters
- Thermal monitoring and power management
- DVFS support with multiple voltage domain
- Individual core power and clock gating support
- Tx and Rx paths optimized for 400Gbps fabric
- ARM V9 ISA virtualization and security support
- Custom SoC level logic support for GPUDirect, CPU-GPU m movement and synchronization



Preliminary disclosure, subject to change. Max chip capability shown, product SKUs specs can be different

NVIDIA GRACE VS. FUJITSU A64FX

A64FX is an outlier in every way – your Grace experience will be different

MAINSTREAM LEADERSHIP HPC

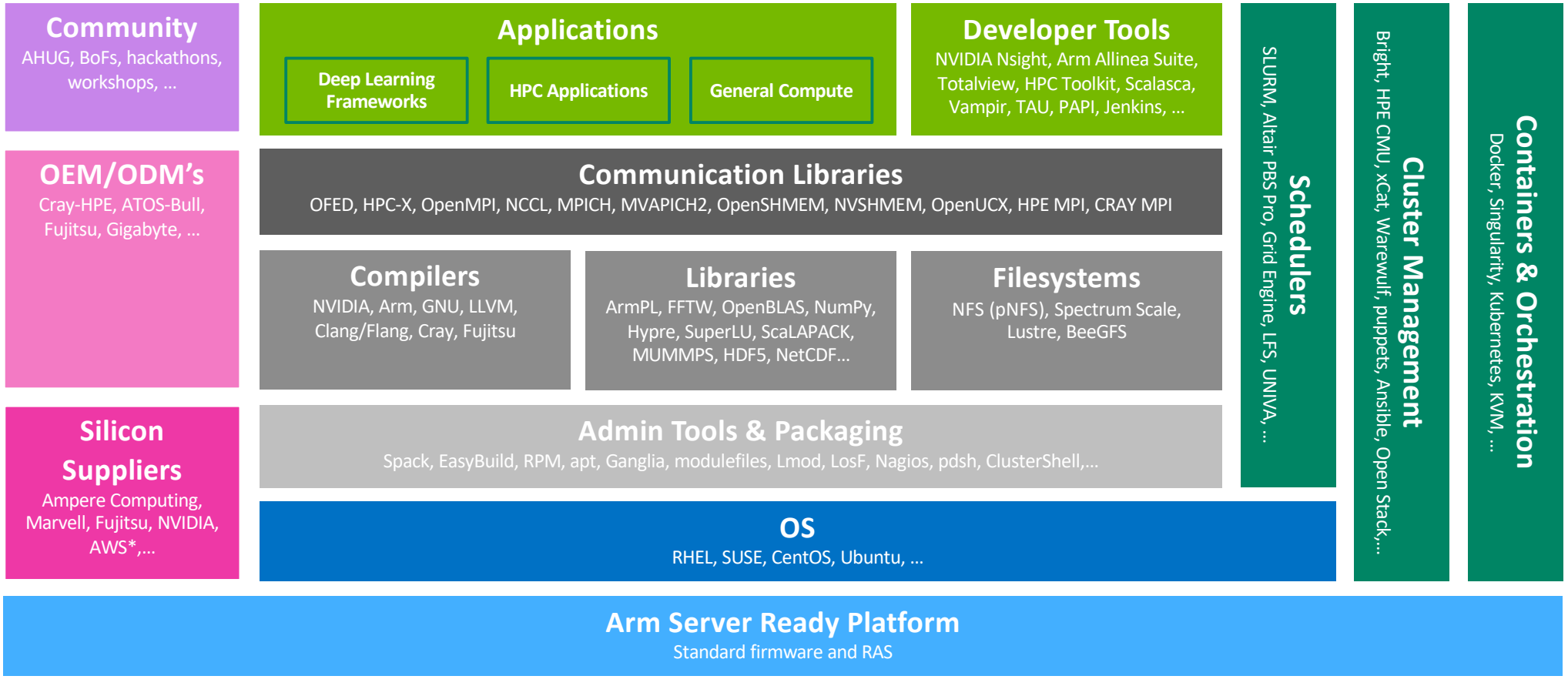
Familiar design	• High single-thread performance • Simple memory hierarchy
Large user community	• Runs key HPC applications out-of-the-box • Standard best practices hold true
Significant fraction of peak w/o tuning	• OSS toolchains (i.e. GNU) are tuned for u-arch • Performance curves generally follow expectation

EXTREME HPC CODESIGN

Codesigned for specific application	• Custom hardware or software • Trades generality for performance
Small userbase of extreme experts	• Nonstandard software environments • Common assumptions may hurt performance
Significant tuning effort required	• OSS toolchains unlikely to be performant • Plan for man-months of optimization effort

ARM IN HPC

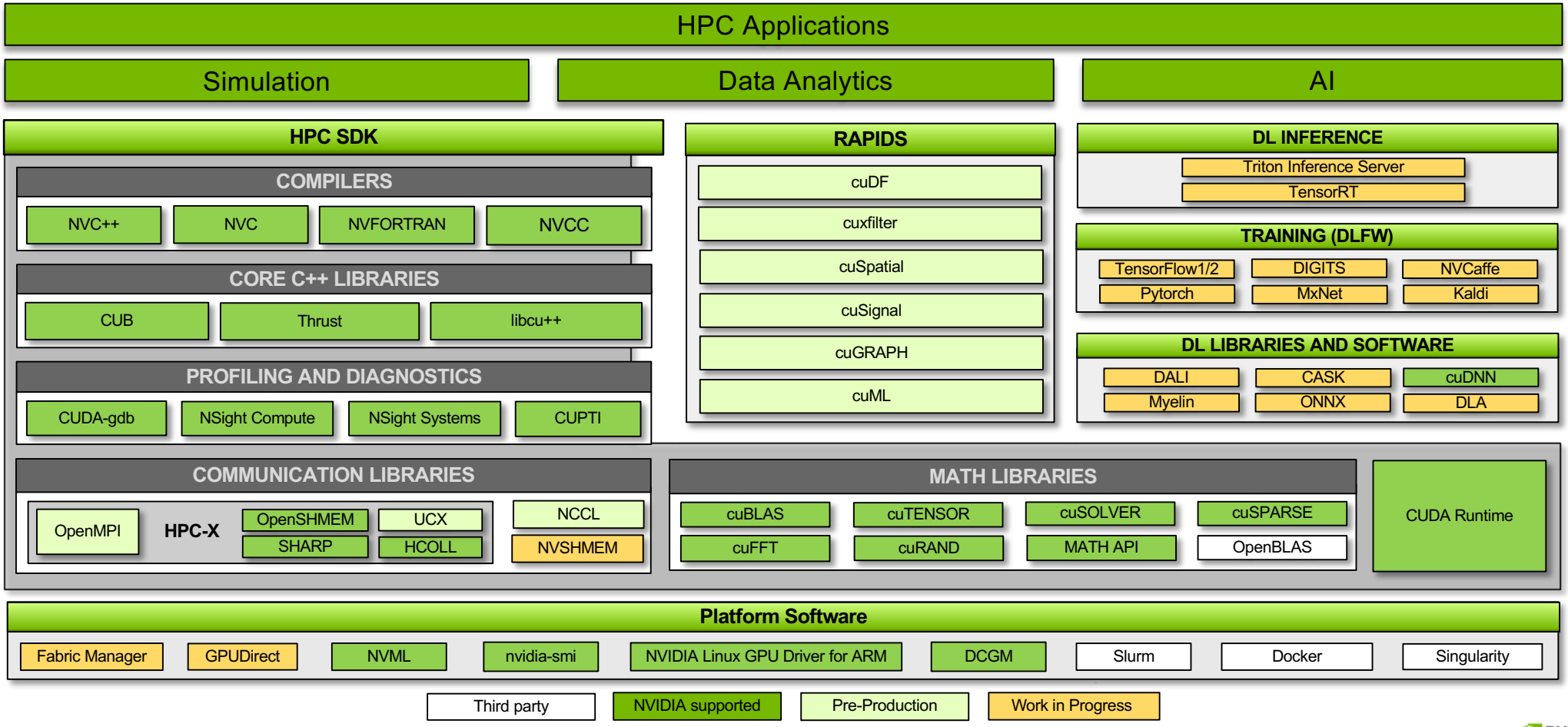
A Growing Ecosystem



NVIDIA

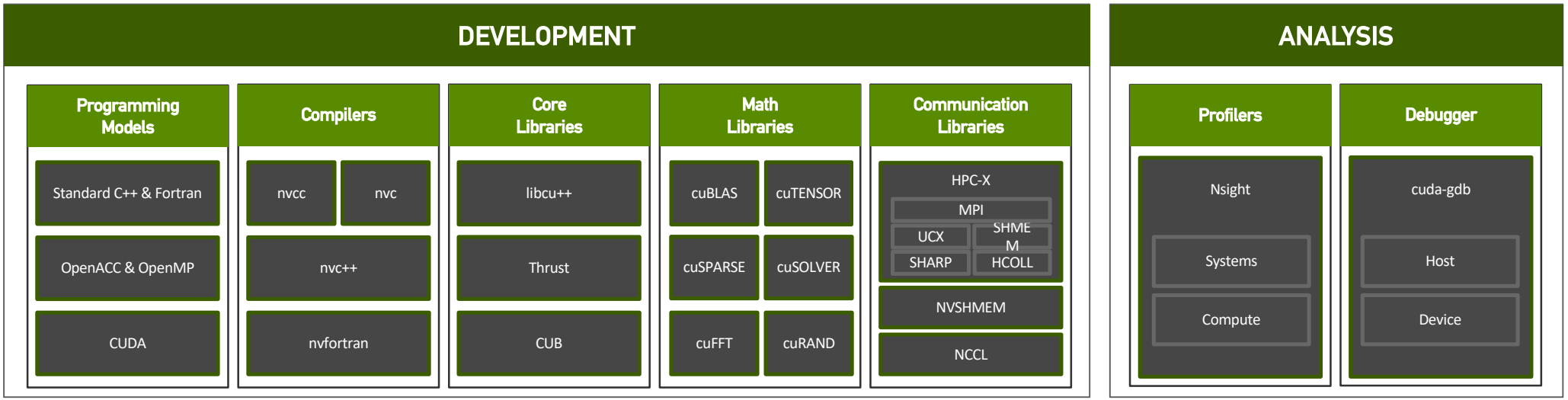
THE NVIDIA HPC SOFTWARE PLATFORM

Simplified Status Assessment



NVIDIA HPC SDK

Available at developer.nvidia.com/hpc-sdk, on NGC, via Spack, and in the Cloud



Develop for the NVIDIA Platform: GPU, CPU and Interconnect

Libraries | Accelerated C++ and Fortran | Directives | CUDA

x86_64 | **AArch64** | OpenPOWER

7-8 Releases Per Year | Freely Available

NVIDIA

ACCELERATED STANDARD LANGUAGES

Parallel performance for wherever your code runs

ISO C++

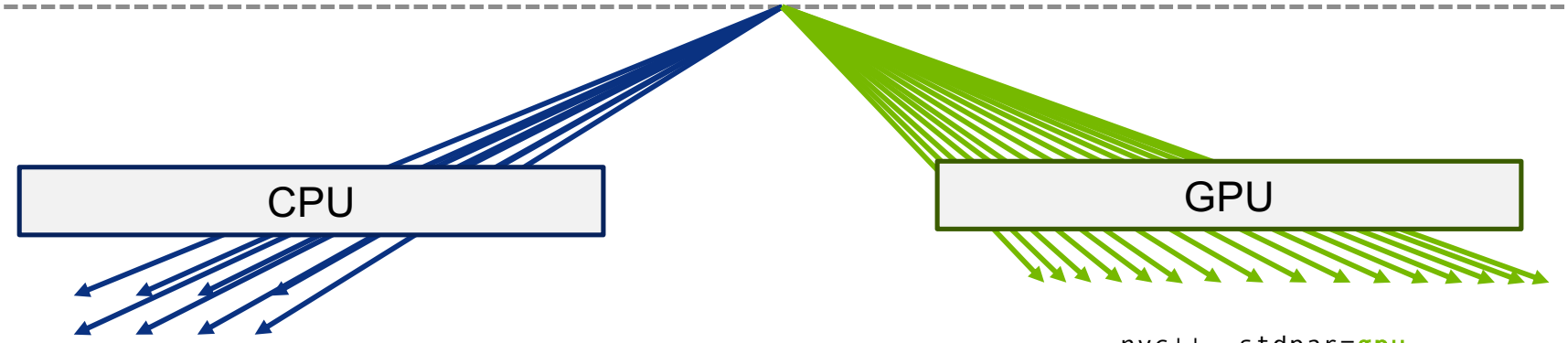
```
std::transform(par, x, x+n, y,
   y,[=](float x, float y){
      return y + a*x;
    }
);
```

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
```

35% BETTER PERFORMANCE ON AWS GRAVITON3

Amazon EC2 C7g instances

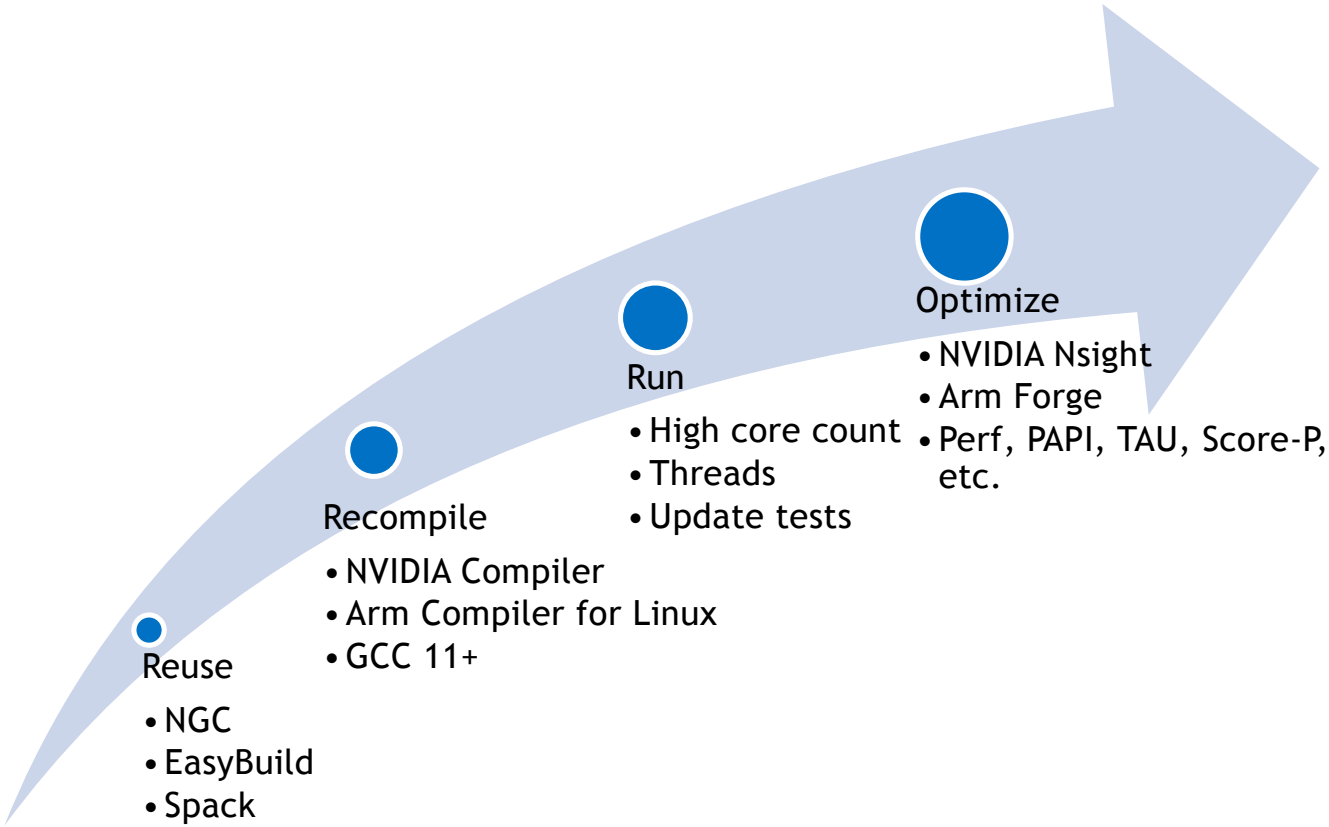
- *"We benchmarked 35% better performance on AWS Graviton3-based C7g instances compared to the previous generation instances.* ***With the support of LS-DYNA on the AWS Graviton3 processor****, Ansys customers will get the best of both worlds – access to a world-class multiphysics solver without comprising on speed, and lower energy and costs."*

-- Prith Banerjee, Chief Technology Officer - Ansys


“ Porting to Arm is boring.”
— Simon McIntosh-Smith … and many more

PORTING TO ARM

Most HPC applications recompile easily and work “out of the box”



Quick Launch

- NGC provides containerized apps
- EasyBuild and Spack can quickly build applications and common dependencies

Compilation Tips

- Most compiler flags are the same:
  - Use -mcpu=native
  - **Don’t use -march or -mtune**
  - You may need -fsigned-char
- Update your unit tests:
  - Aarch64 floating point is as accurate as all other platforms

USE STANDARDS-COMPLIANT MULTI-PLATFORM COMPILERS

You're not porting to Arm. You're porting *away* from ifort, xlf, etc!

- Use any portable multi-platform compiler: NVIDIA, GCC, LLVM, etc.
- Use the most recent compiler possible. **If using GCC, version 11+ is strongly recommended.**
- Beware of non-standard build systems
  - icc, ifort, xlf, etc. may be hard-coded into the build system
  - Be explicit about which compiler to use. Don't let the build system make assumptions
- Beware of non-standard default compilers
  - Make sure default compiler commands (cc, fc, gcc, etc.) invoke a recent cross-platform compiler
  - Use `` `mpicc -show` `` or similar to verify that MPI compiler wrappers are invoking the right compiler
- Log the build, then check the log afterward

PORTING ASSEMBLY AND X86 VECTOR INTRINSICS

Translate intrinsics to port functionality, then focus on performance tuning

- For a quick fix, use a drop-in header-based intrinsics translator
  - SIMD Everywhere (SIMDe): https://github.com/simd-everywhere/simde
  - SSE2NEON: https://github.com/DLTcollab/sse2neon
  - Quick tutorial using BWA-MEM2: https://www.nvidia.com/en-us/on-demand/session/gtcspring22-s41702/
- Follow Arm's documentation on rewriting x86 vector intrinsics
  - Porting and Optimizing HPC Applications for Arm SVE [https://developer.arm.com/documentation/101726/latest]
  - Coding for NEON [https://developer.arm.com/documentation/101725/0300/Coding-for-Neon]
- Arm assembly is simpler than x86
  - Arm processors have a much simpler and general set of registers than x86. Just assign a one-to-one mapping from an x86 register to an Arm register when porting code.
  - Complex x86 instructions will become multiple Arm instructions

ARM SIMD PROGRAMMING APPROACHES

Follow these recommendations in order, e.g. prefer auto-vectorization over intrinsics

- Compilers
  - Auto-vectorization: NVIDIA, GCC, LLVM, ACfL
  - Compiler directives, e.g. OpenMP
    - `#pragma omp parallel for simd`
    - `#pragma vector always`
- Libraries:
  - NVIDIA Math Libraries
  - Arm Performance Library (ArmPL)
  - Open Source Scientific Libraries (BLIS, FFTW, PETSc, etc.)
- Intrinsics (ACLE):
  - Arm C Language Extensions for SVE
  - Arm Scalable Vector Extensions and Application to Machine Learning
- Assembly:
  - SVE ISA Specification: The Scalable Vector Extension for Armv8-A

ARM SCALABLE VECTOR EXTENSION (SVE)

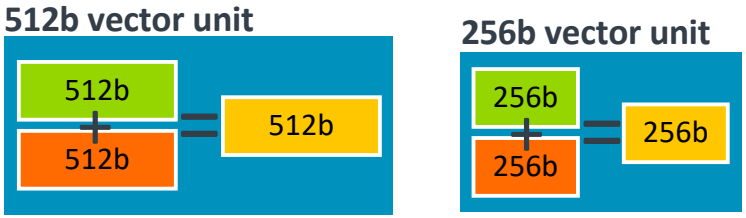
An ISA feature which Arm partners can implement at length - 128 to 2048 bits

How SVE works

The hardware sets the vector length

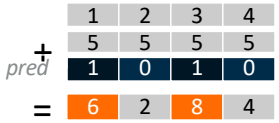
In software, vectors have no length

The *exact same* binary code runs on hardware with different vector lengths

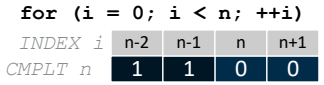


SVE improves auto-vectorization

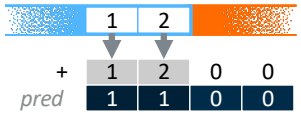
Gather-load and scatter-store



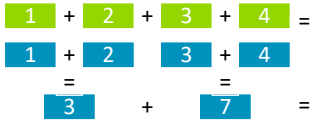
Per-lane predication



Predicate-driven loop control and management



Vector partitioning and software-managed speculation



Extended floating-point horizontal reductions

SVE SUPPORT IS MATURE

Arm actively posting SVE open source patches upstream since 2016

Beginning with first public announcement of SVE at HotChips 2016

Available upstream

GNU Binutils-2.28:	Released Feb 2017, includes SVE assembler & disassembler
GCC 8:	Full assembly, disassembly and basic auto-vectorization
LLVM 7:	Full assembly, disassembly
QEMU 3:	User space SVE emulation
GDB 8.2	HPC use cases fully included
LLVM:	Since Nov 2016, as presented at LLVM conference
Linux kernel:	Since Mar 2017, LWN article on SVE support

INSIDE GRACE WATCH PARTY

Ask us the invite to the

“Inside Grace”

watch party organized by

Filippo Spiga

AI generated picture

INSIDE GRACE WATCH PARTY

Ask us the invite to the

“Inside Grace”

watch party organized by

Filippo Spiga

AI generated picture

lovely Spanish village with castle in Extremadura