High Performance Scientific computing

S. Gopalakrishnan

EMERGING GRAND CHALLENGE APPLICATIONS

- Full systems modeling
  - Earth systems: Complex Climate Models
  - Biological systems: Human Microbiome Modeling
  - Complex Engineering Systems: Aircraft, Power Plant etc.
  - Environmental modeling like distribution of Micro-Plastics
  - Novel Drug designs, Material designs (Inverse problems)

EMERGING GRAND CHALLENGE APPLICATIONS

- They span across several orders of magnitude space and time scales
- Span across large number of physical, chemical, quantum phenomena

- Example could be influence of Aerosols to Ocean Circulation

EMERGING GRAND CHALLENGE APPLICATIONS

- At the same time silicon process technologies running out of steam

EMERGING GRAND CHALLENGE APPLICATIONS

- Classic brute force ModSim algorithms may not be sufficient or possible
- Classic Fortran + MPI is not sufficient
- Homogeneous CPU Nodes or GPU Nodes are not sufficient
- Homogeneous ModSim clusters may not be sufficient

EMERGING GRAND CHALLENGE APPLICATIONS

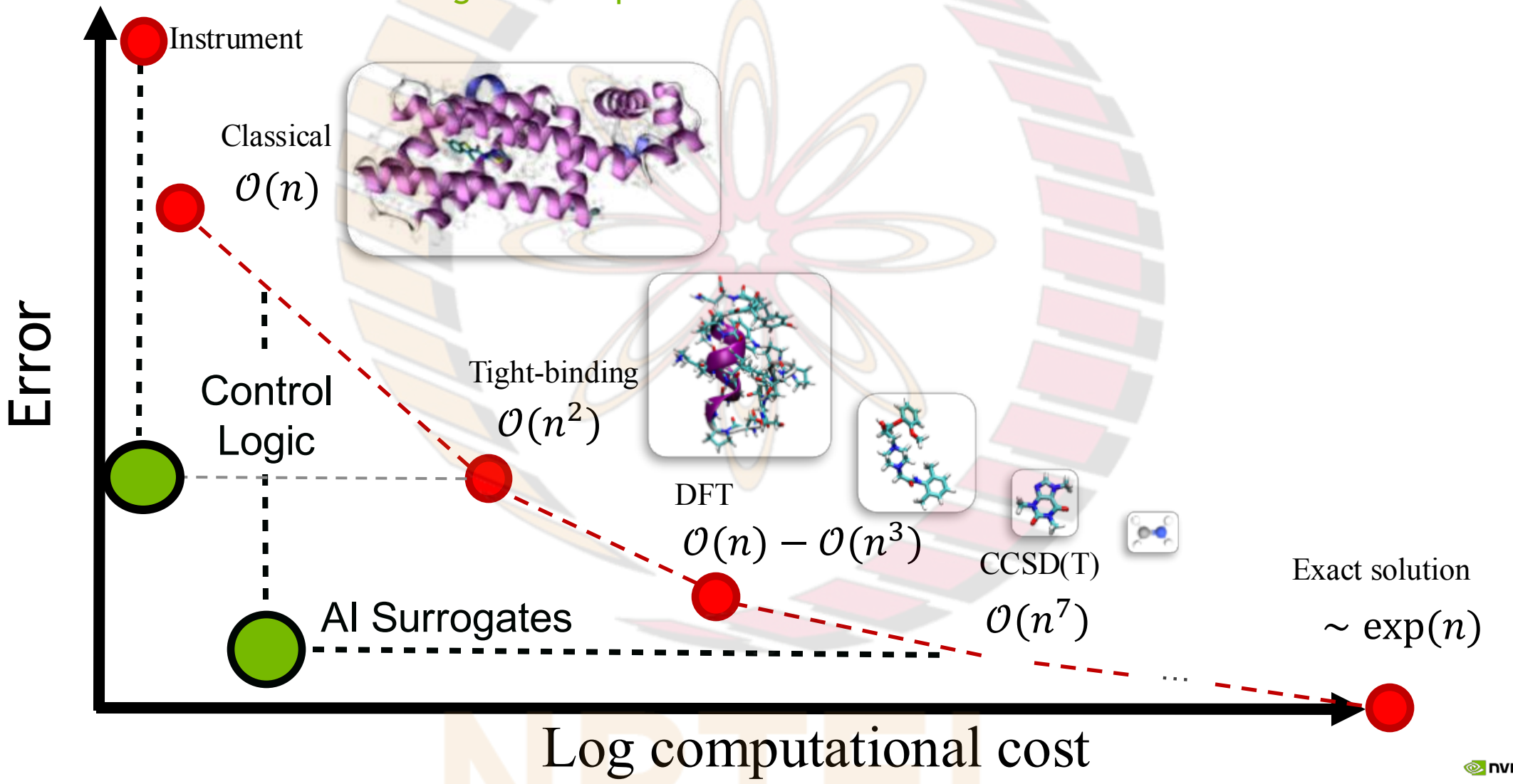
- Data and Algorithms (ML/DL) to the rescue
- ModSim + ML/DL + Data

ML/DL WILL CONTINUE MOORE'S OBSERVATION IN HPC

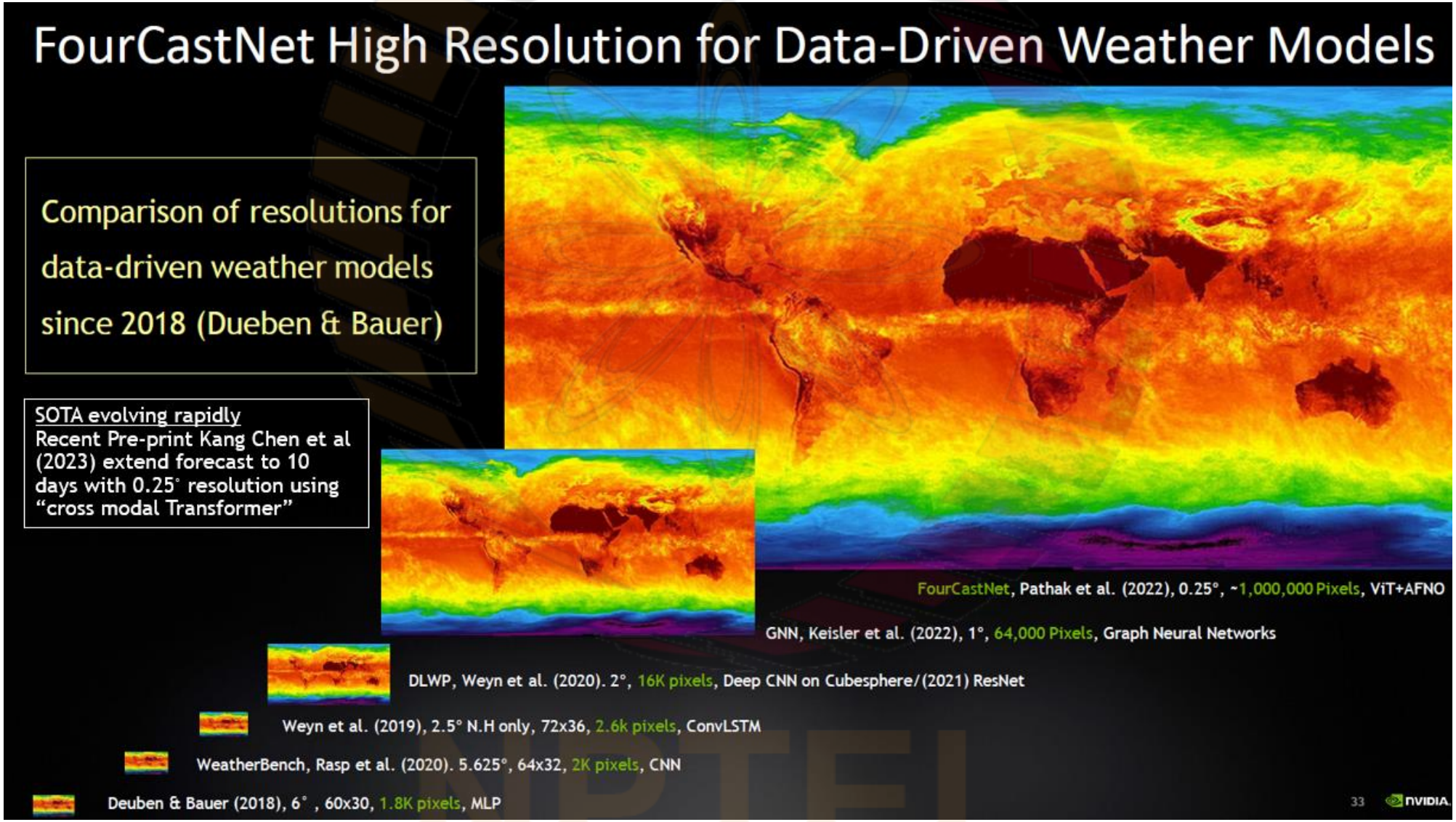
- ML/DL is rapidly becoming the 4th pillar of Scientific Discovery
- Approximations in Mod-Sim, Inverse problems, Insights from experimental/empirical data etc.
- Protein Structure Prediction became a solved problem
- Sub-Seasonal Forecast Competition is won by MSFT ML/DL Team
- Physics informed NNs in CFD

AI INTRODUCES NEW USE CASES FOR SCIENCE AND ENGINEERING

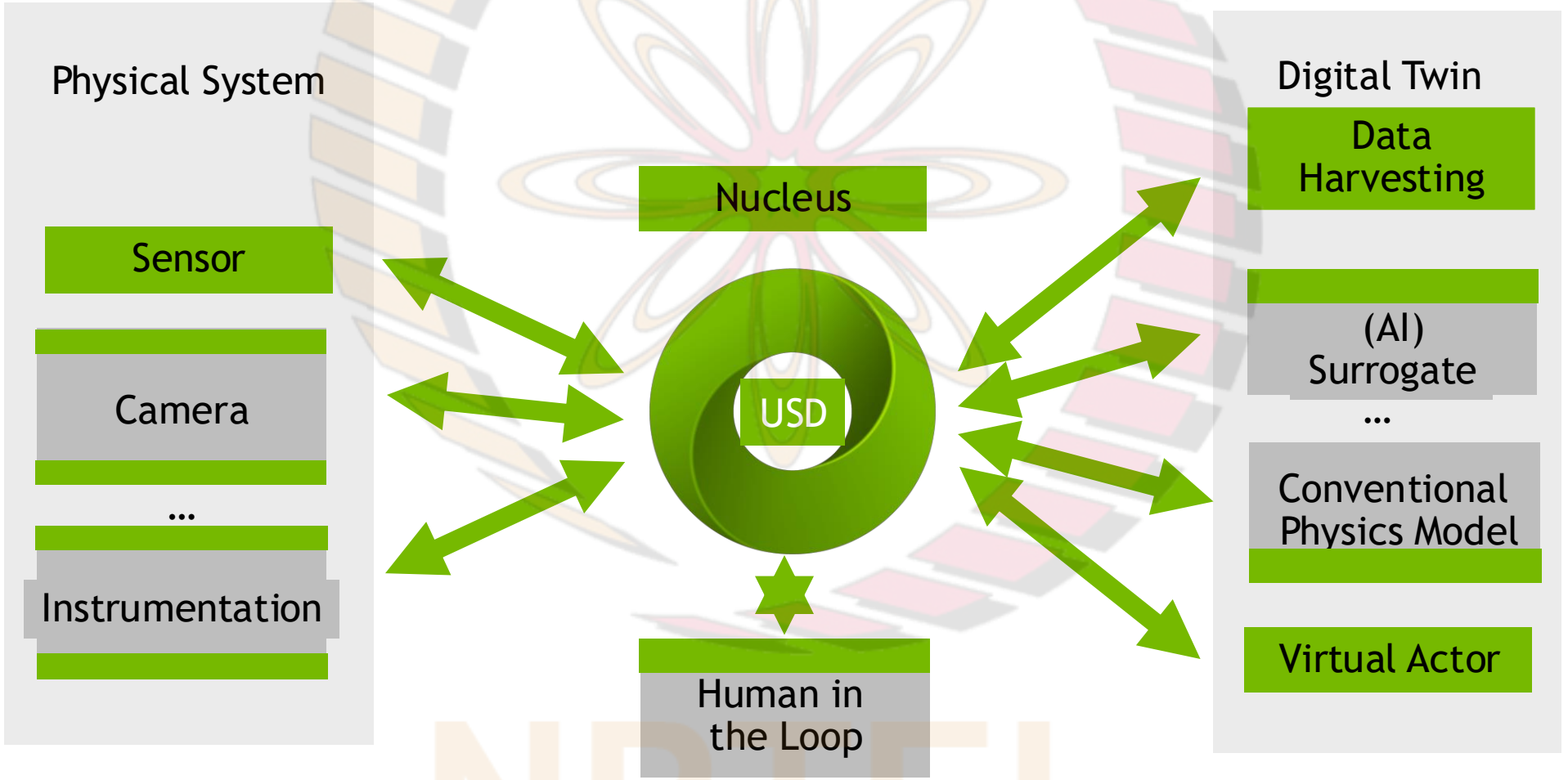
AI Bridges the Gap Between Simulation and Real-Time



AI ALGORITHMS EVOLVING AT UNPRECEDENTED PACE



Digital Twin For Science

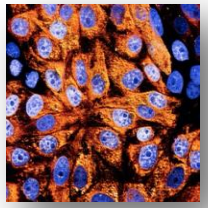


EMERGING GRAND CHALLENGE APPLICATIONS

- Supercomputers may need to integrate with sensor networks
- High throughput scientific instruments
- Most importantly, capable of solving HPC + ML in a tightly coupled loop

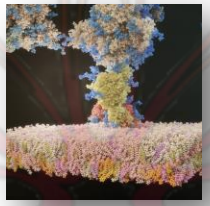
Workloads of the Modern Supercomputer

EDGE

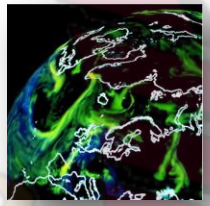


SIM + AI

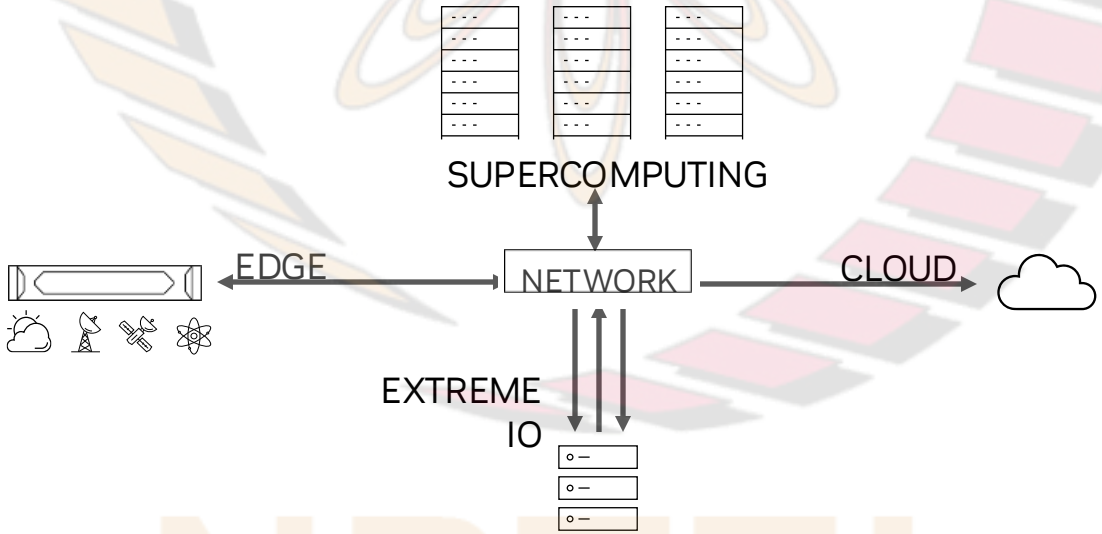
SIMULATION



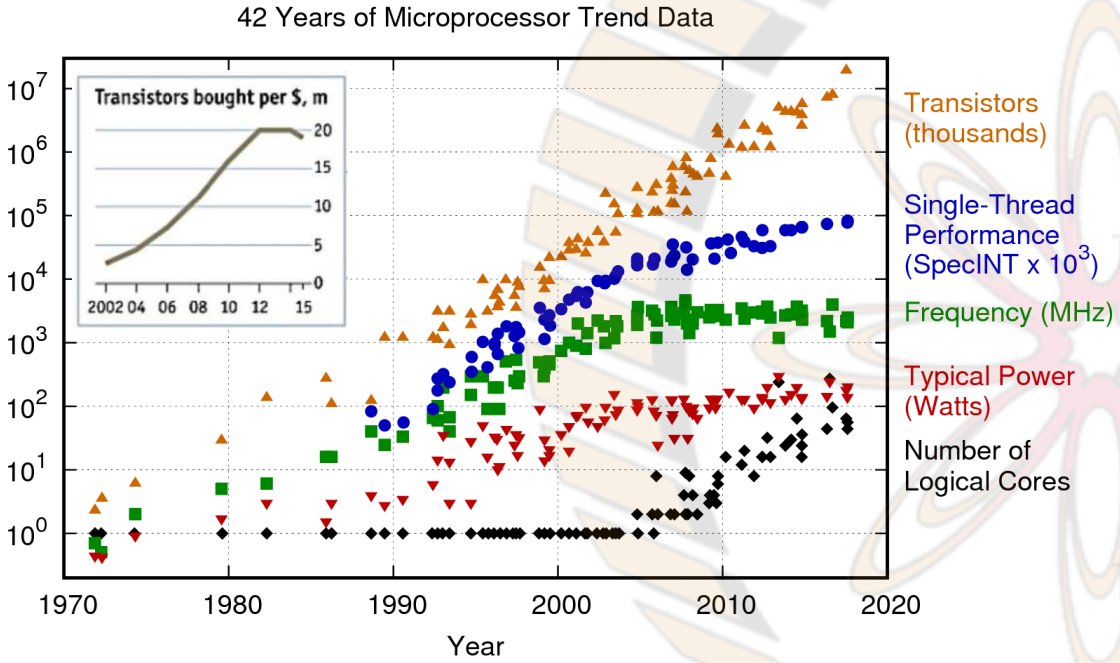
DIGITAL TWIN



QUANTUM COMPUTING

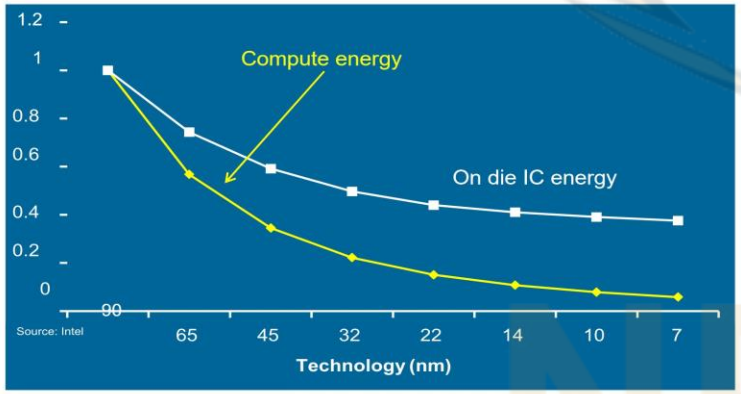


MOORE'S OBSERVATION



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2017 by K. Rupp

- Original:
  - Transistor density doubles
  - Power stays same and Cost stays same
- Last two are gone
  - Due to loss of Dennard scaling and Litho and other process complexities
- Different features scale differently with process
  - Logic, SRAM, Wires, and IOs
    - Distorting over time



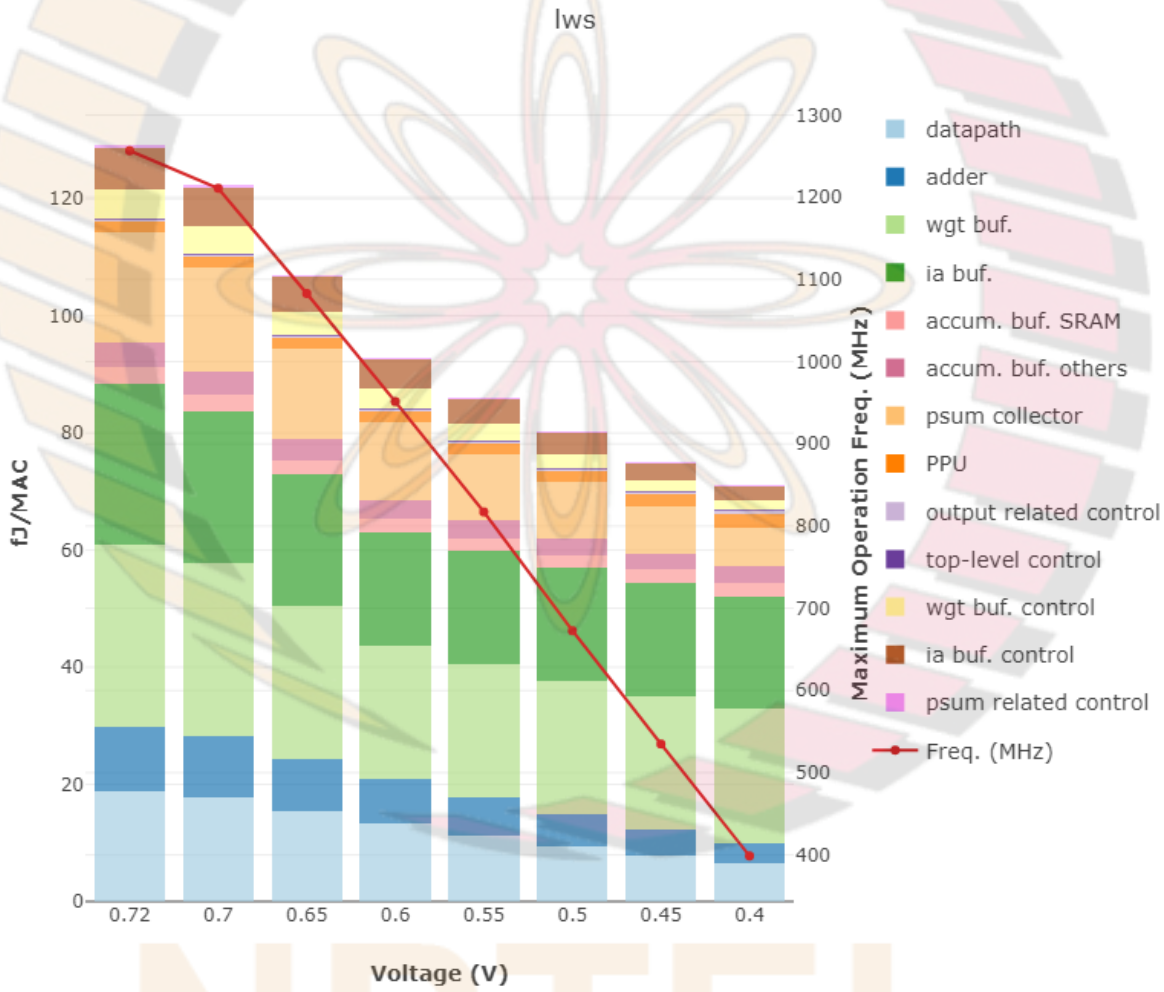
NVIDIA.

ENERGY DOMINATED BY MEMORY AND DATA

70 fJ/MAC

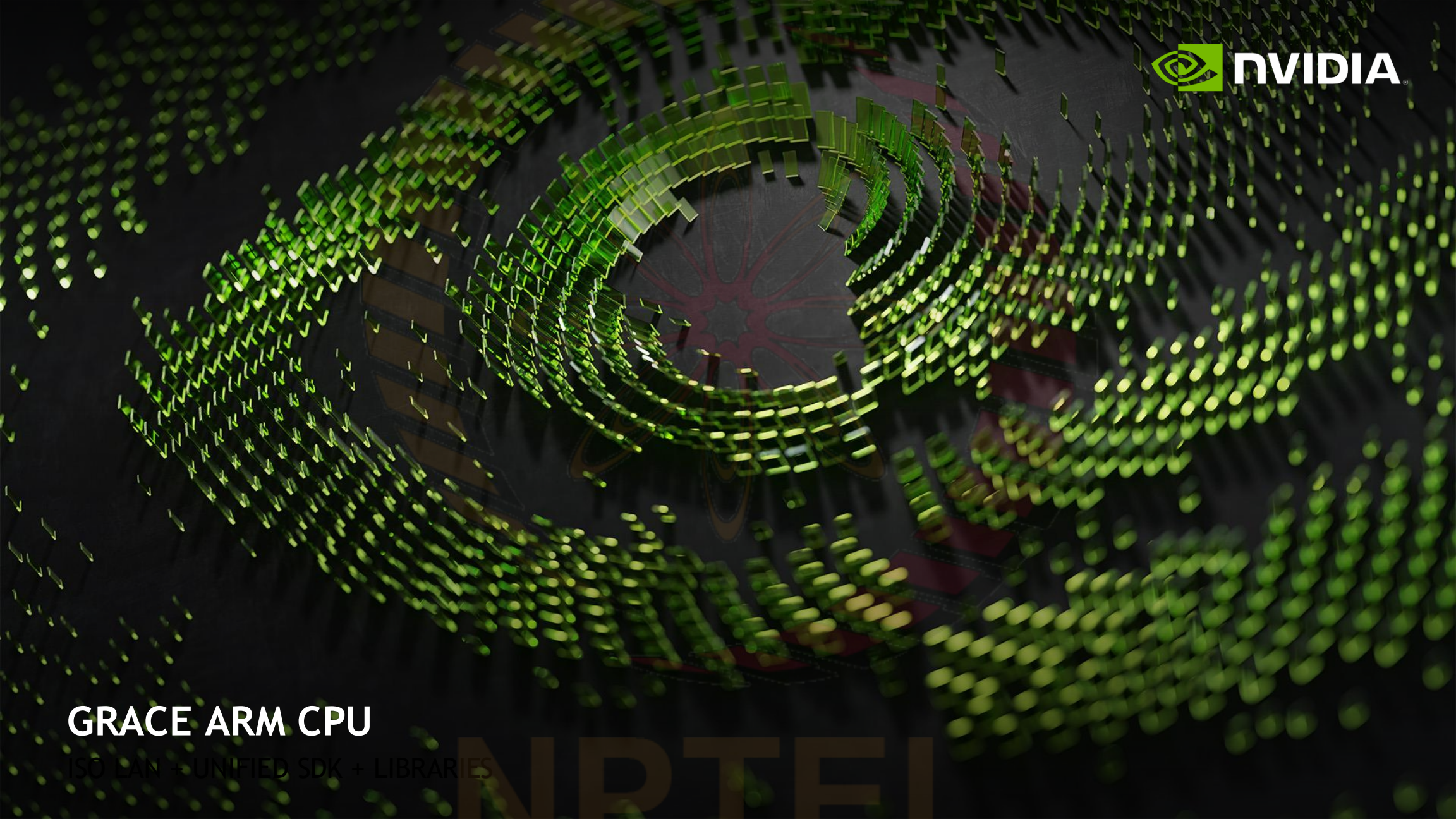
35 fJ/OP

29 TOPS/W



MORE THAN MOORE

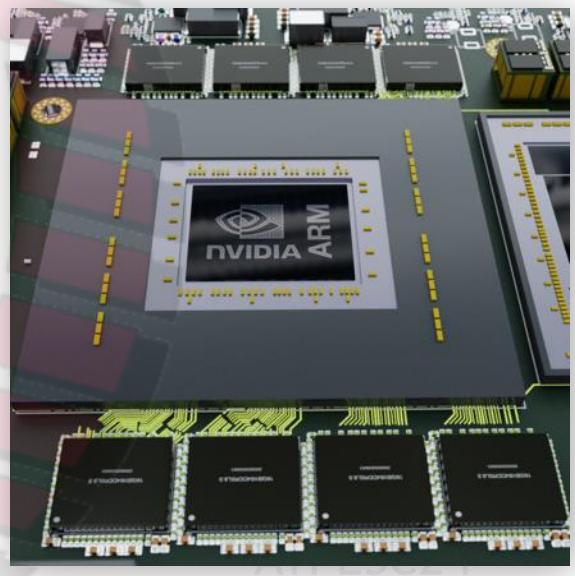
- Nvidia maintains “More than Moore” by optimizing
  - End-to-End mapping of Applications to Supercomputing System
  - Algorithmic, SW, architectural, Packaging, Process Technology
  - Try to maintain 2X to 6X Gen-to-Gen perf improvement
  - Requires close collaboration with HPC/ML/DL Community
  - Transformer Acceleration in Hopper is the best recent example
  - FLOPs not executed, Bytes that are not moved are the best FLOPs and Bytes


NVIDIA
GRACE ARM CPU
ISO LAN + UNIFIED SDK + LIBRARIES

NV ARM OVERVIEW

Server Class ARM CPU

- 64bit Server Class Core and SoC
  - Arm V9.0 ISA Compliant aarch64 core
  - Full SVE-2 Vector Extensions support, inclusive of NEON instructions
  - Supports 48-bit Virtual and 48-bit Physical address space
- Balanced architecture between Single Core Perf, Core count, Memory and IO subsystems
- Supports Arm Server ready (SBSA), Boot compliant (SBBR) and manageability (SBMG) open standards
  - Linux, system management, HPC stacks will run out of the box



NVIDIA ARM CORE

Optimized for Single Thread Performance

Scalar Side

- Wide Super Scalar OoO single-threaded high performance Core Pipeline
  - Supporting Multi-stage branch prediction and advanced prefetching algorithms
  - 1LD + 4LD/ST pipes, 6 ALU pipes
- 64B Cacheline
- 64KB L1ICache, 64KB L1DCache
- 1MB Private L2Cache

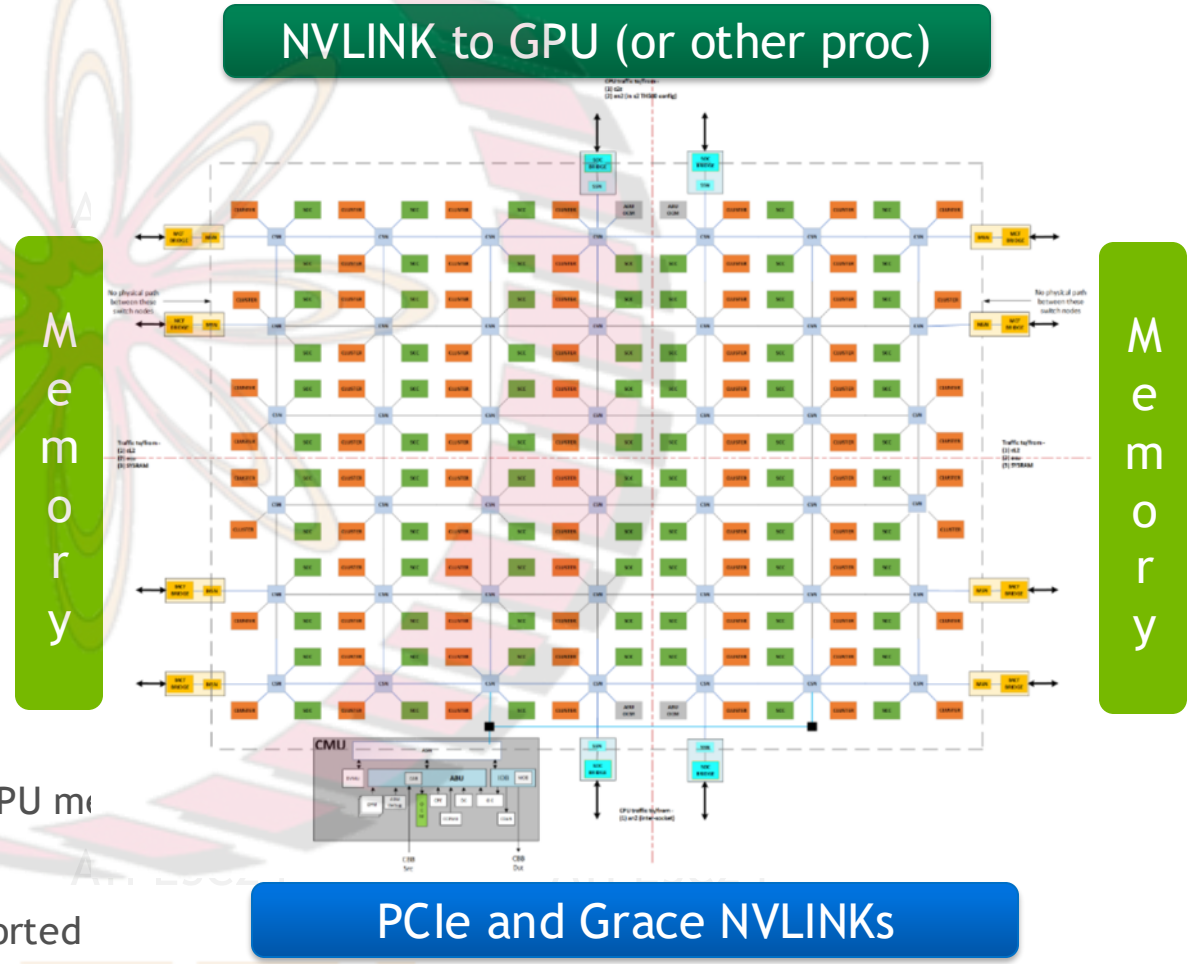
Vector Side

- Four 128-bit SVE2 Execution pipes capable of 16 DP FLOPS per Cycle
- Support 64bit, 32bit, 16bit and bfloat16 FP and int8
- Complex datatypes and math
- Enables easier vectorization through Predicate and mask instructions
- Lane widening and narrowing instructions
- Classic and non-temporal Gather Load and Scatter Store instructions
- Crypto extensions

PROCESSOR SOC

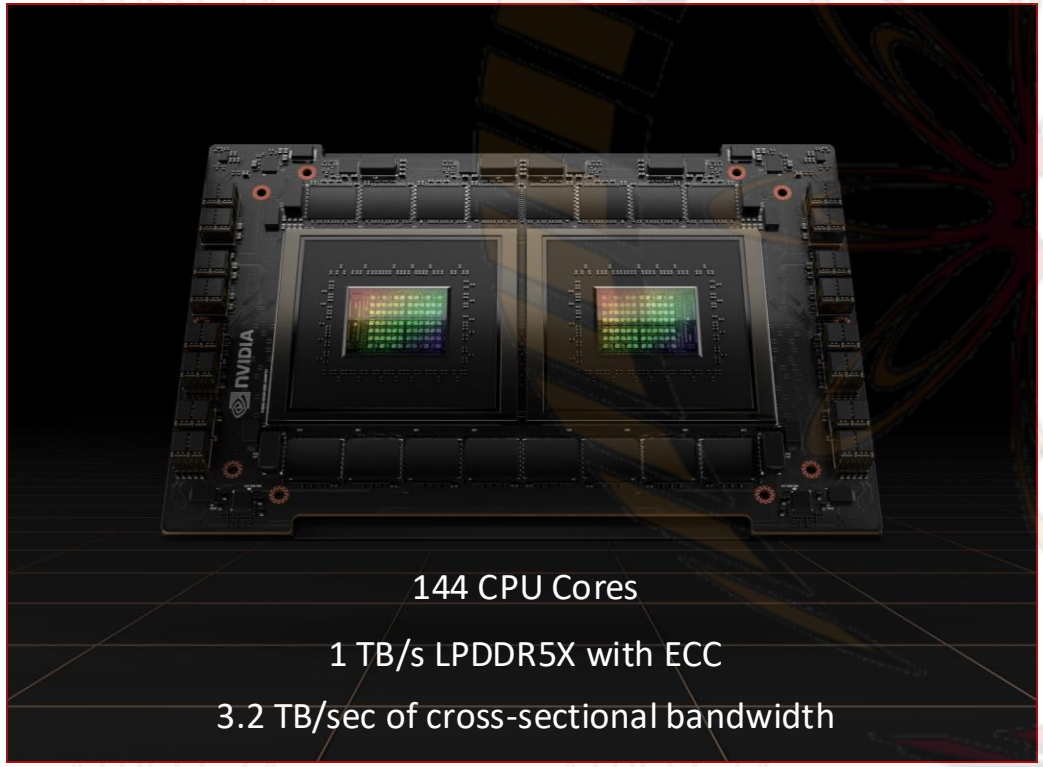
Augmented custom logic to support memory movement

- Monolithic SoC
  - Up to 117MB 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
- 120GB, 240GB, 480GB LPDDR5 CapaATPESC24es supported



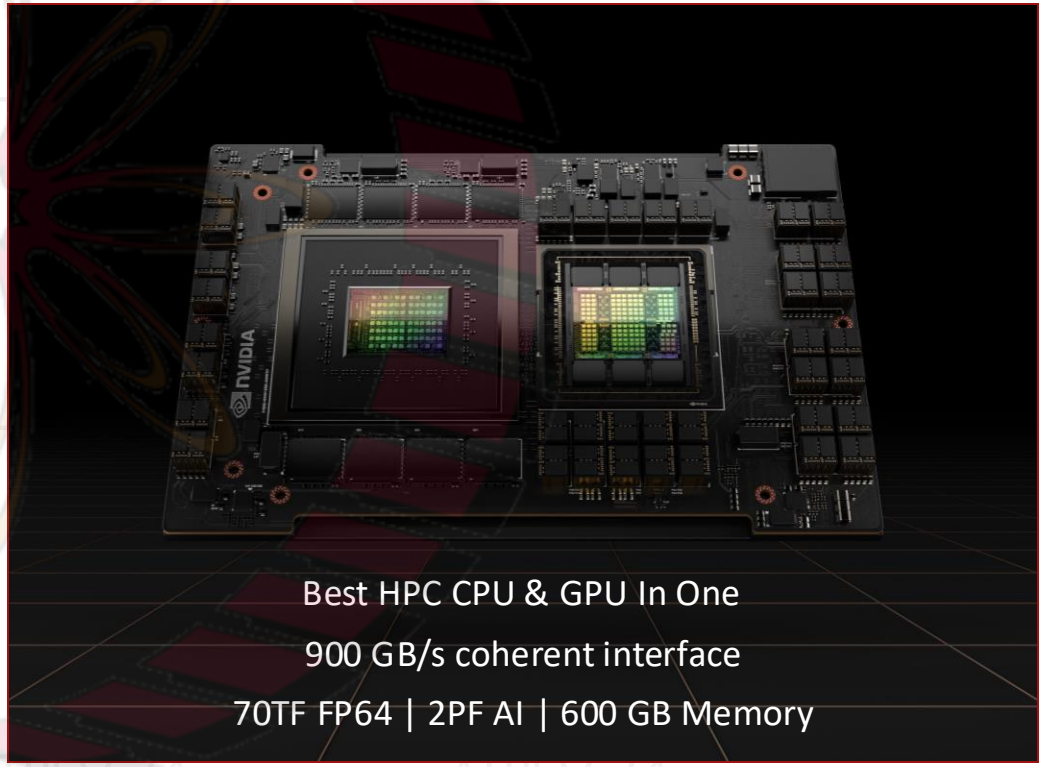
NVIDIA Grace and Grace Hopper

High Performance for an Energy Constrained World



Grace CPU Superchip

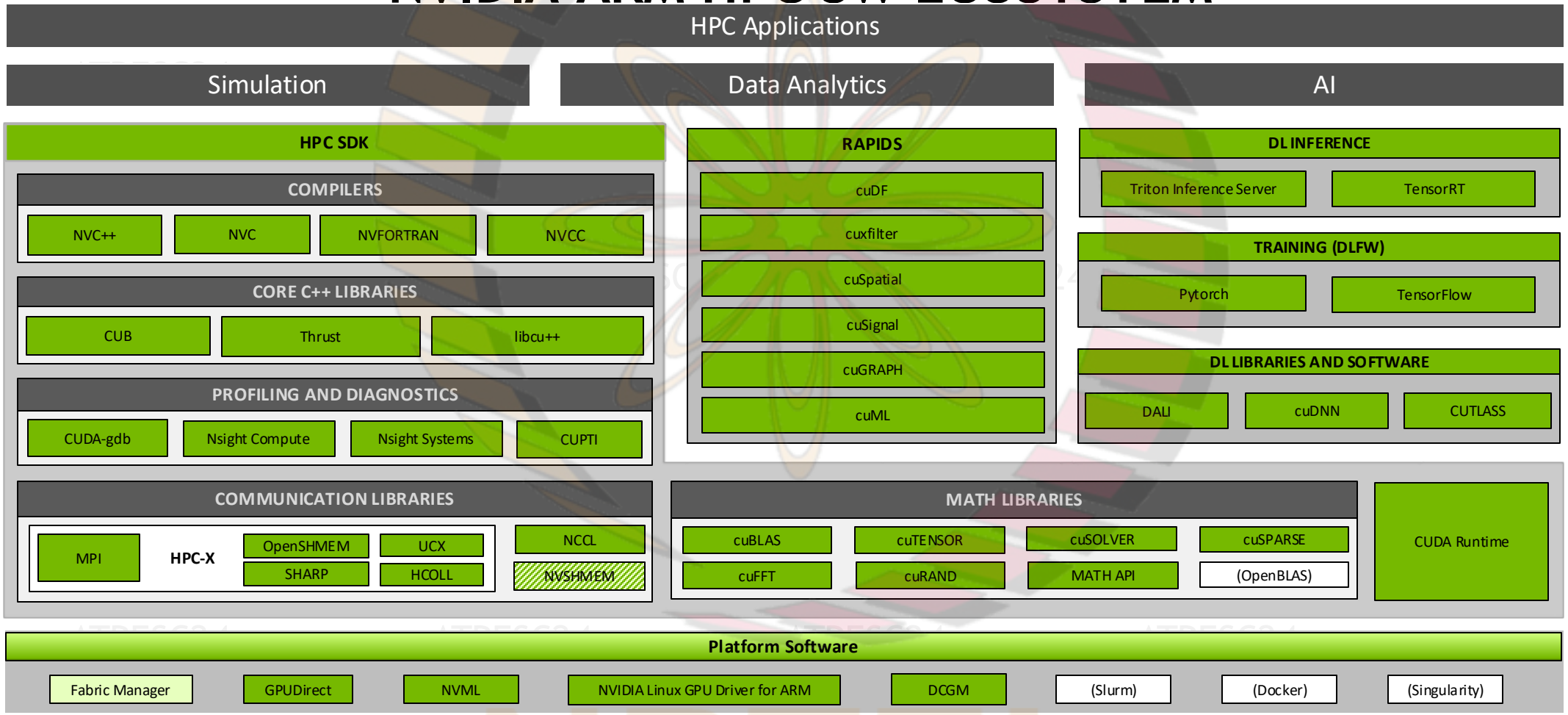
High-performance CPU for HPC and cloud computing

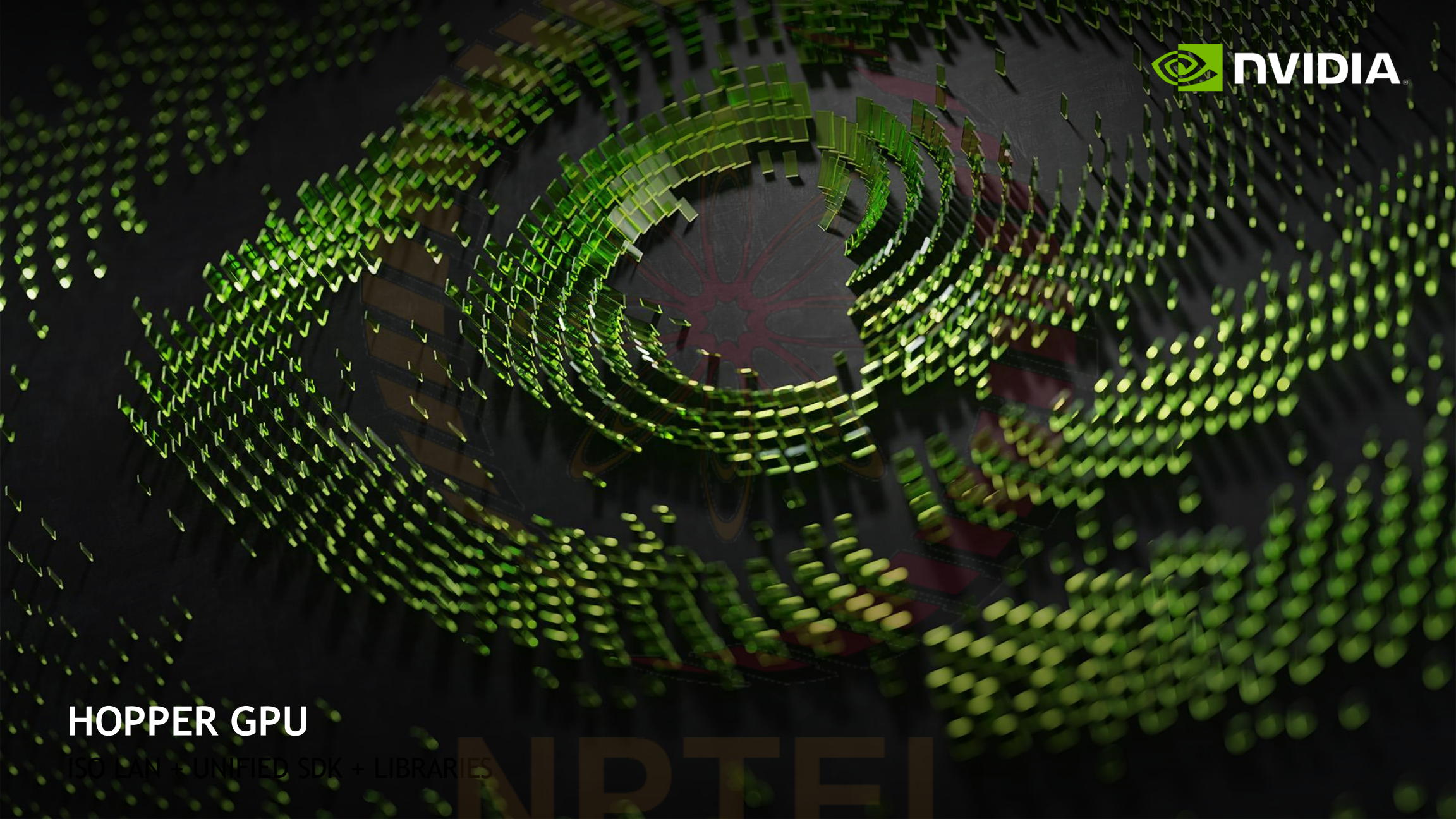


Grace Hopper Superchip

CPU+GPU designed for giant-scale AI and HPC

NVIDIA ARM HPC SW ECOSYSTEM

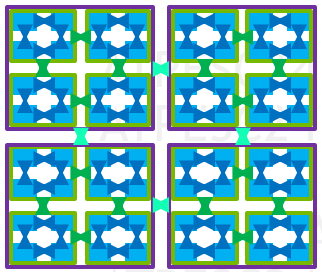



NVIDIA
HOPPER GPU
ISO LAN + UNIFIED SDK + LIBRARIES

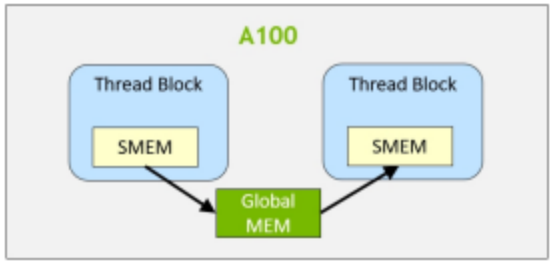
NEW HOPPER SM DOES MORE THAN IMPROVE RAW SPEEDS AND FEEDS

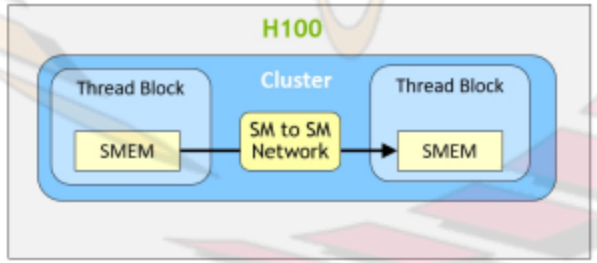
- New **Thread Block Clusters**
  - Turn locality into efficiency
  - Support **Distributed Shared Memory** between SMs
- New **Asynchronous Transaction Barriers**
  - Increased support for asynchronous programming
- New **Tensor Memory Accelerator**
  - Fully asynchronous data movement
- New **DPX** instruction set
  - Special Purpose Acceleration
- New **Transformer Engine** for AI Model Acceleration

THREAD BLOCK CLUSTERS



- New feature introduces programming locality within clusters of SMs
- About 7X higher throughput vs. using global memory
- Shared memory blocks of SMs within a GPU Processing Cluster (GPC) can communicate directly (w/o going to HBM)
- Leveraged with CUDA cooperative groups API



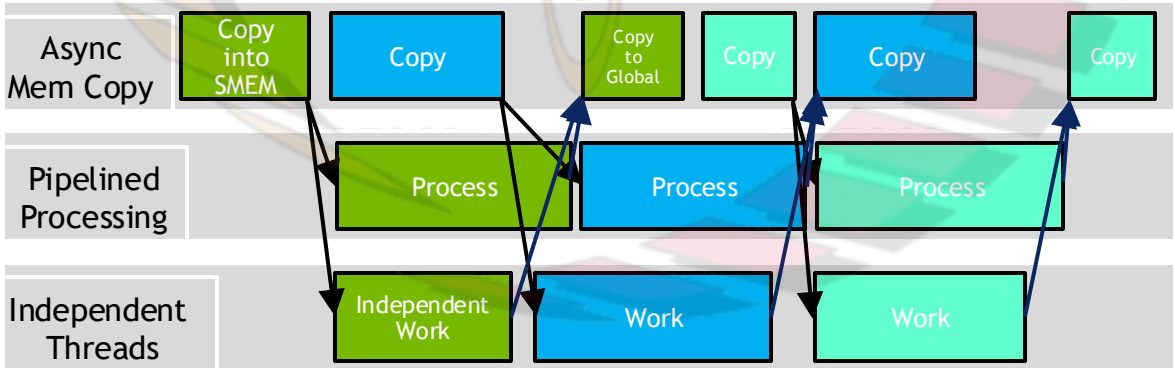


For details, see "NVIDIA H100 Tensor Core GPU Architecture" white paper available for download

ASYNCHRONOUS ENHANCEMENTS

Hopper enables end-to-end fully asynchronous pipelines

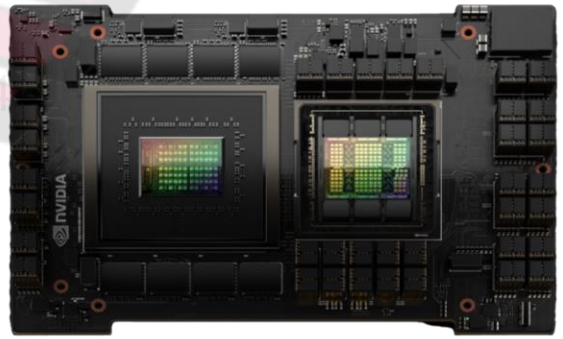
- Async Transaction Barriers – Atomic data movement with synchronization
- More efficient Waiting on Barriers
- Async Mem_copy via Tensor Memory Accelerator (TMA)

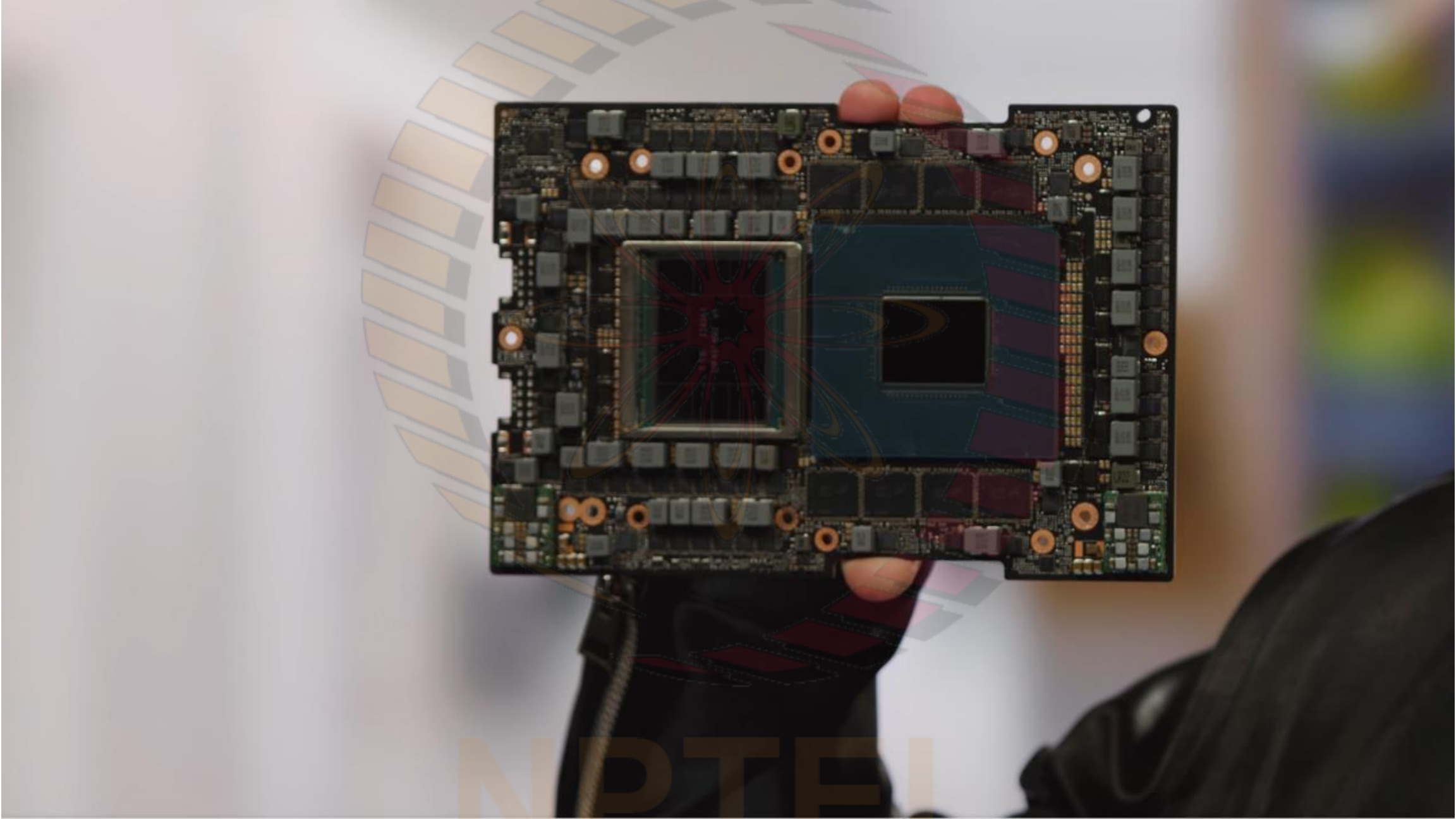


GRACE-HOPPER

A revolutionary Architecture

- Nvidia GPUs
  - Latency hiding Throughput Machines => Async Computational Graph solvers
    - Can effectively map *Dataflow-Complex* portions of the Algorithm
  - Custom (compute and memory) IP Blocks for energy efficiency
- Nvidia CPUs
  - SuperScalar, OOO Core based tightly coupled SoC with balanced Bytes/s/FLOPS
  - Can effectively map *ControlFlow-Complex* portions of the Algorithm
  - Strong Vector and Tensor performance in future
- Unified Compute Substrate





GB200 With NVL72 Enabling Trillion Parameter AI

NVL72: One Big GPU

GB200 NVL72

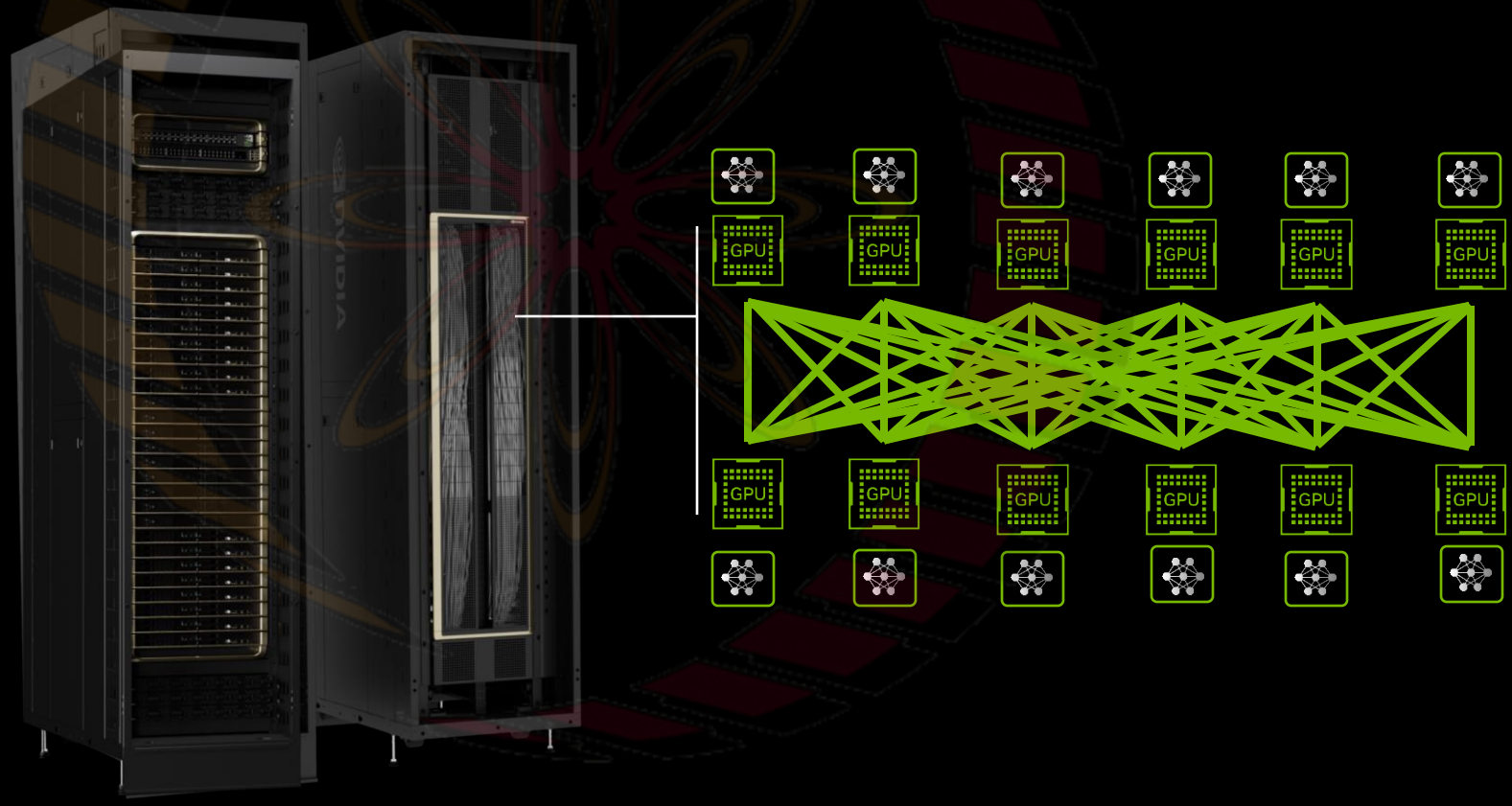
36 GRACE CPUs

72 BLACKWELL GPUs

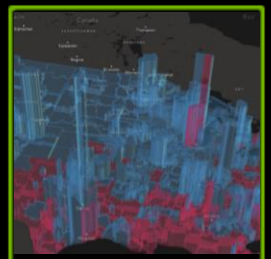
One NVLink Domain

130 TB/sec All-to-All Bandwidth

18x vs HDR



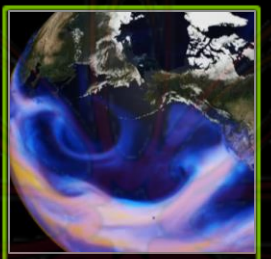
NVIDIA AI Accelerated Computing Platform



DATA PROCESSING

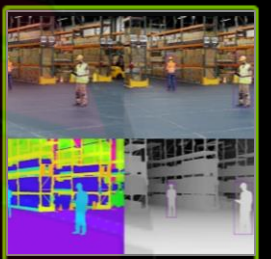
CAD, CAE, SDA

COMPUTER-AIDED DRUG DESIGN



CLIMATE SIMULATION

QUANTUM SIMULATION

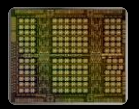


ROBOTICS & INDUSTRIAL DIGITAL TWINS

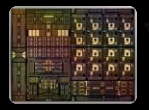
ENTERPRISE AI

CUDA-X LIBRARIES

CPU



GPU



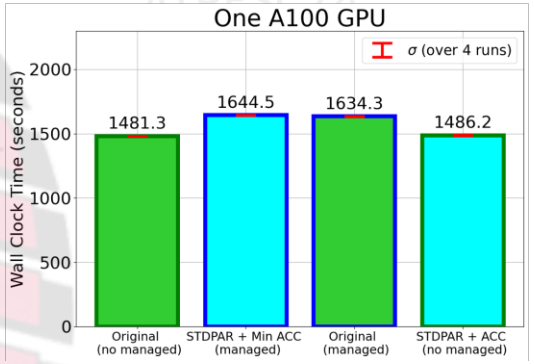
DPU

ACCELERATED COMPUTING

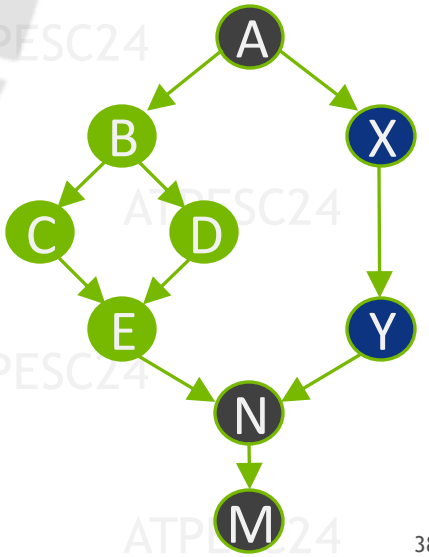
COMPUTATIONAL GRAPH + GRACE HOPPER

Same programming model for CPU and GPU, plus existing applications ready-to-run Day 1.

- True Unified Memory + High Bandwidth Link make data movement less of a bottleneck
- Existing CPU programming models continue to work on Grace CPU with high performance



Senders/Receivers enable defining hybrid execution graphs to take advantage of the strengths of each processor.



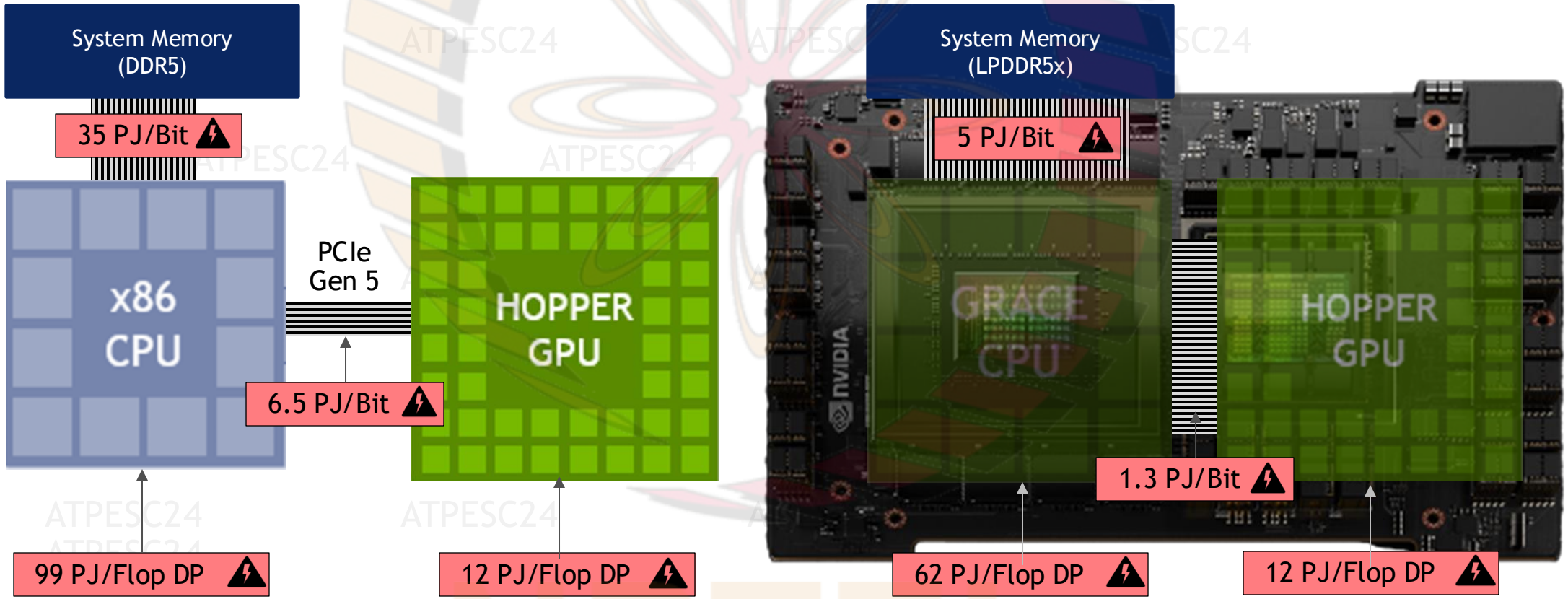
UNIFIED COMPUTE SUBSTRATE

- Grace + Hopper Enables broader set (all!) of codes to be accelerated
- Grace + Hopper enables architectural mapping of both control and data flow complex portions of the algorithm efficiently
  - Enables mapping of Multi-Scale, Multi-Physics Apps, post-Foundation AI and Complex workflows
- Standards-based parallelism enables productive portability
- Non-accelerated, fully-accelerated and mixed controlflow-dataflow complex applications can be run on Grace-Hopper
- Mapping these complex workflows to Distributed Heterogeneous Compute platforms require Omniverse like Digital Twin Frameworks

Energy Efficient Design

NVIDIA SUPERCHIPS SAVE ENERGY

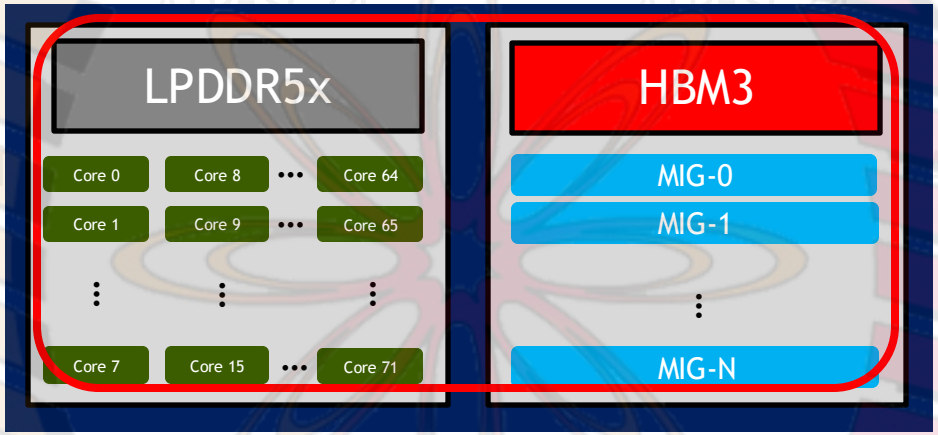
Low Power Data Motion and Computation

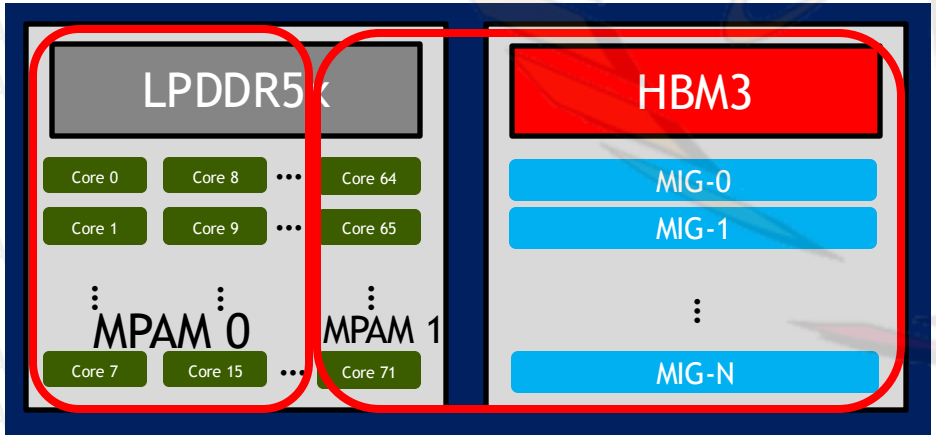


CO-SCHEDULING

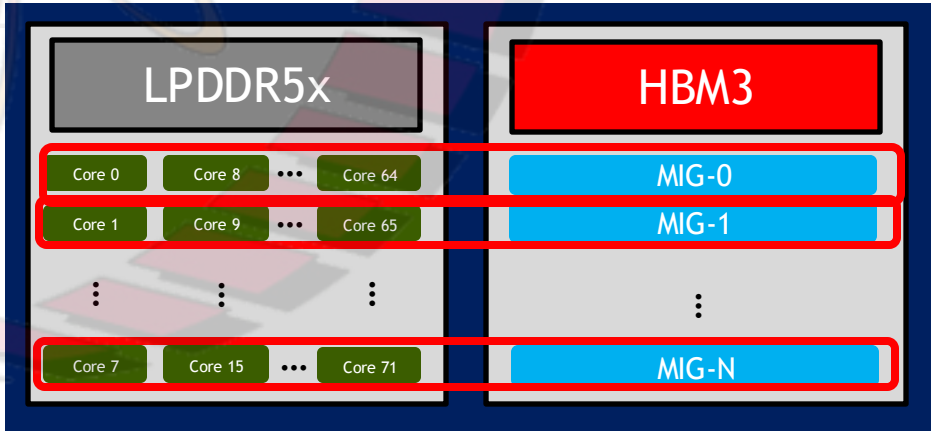
Alternative to CPU-only partition

One Job exclusive to the node





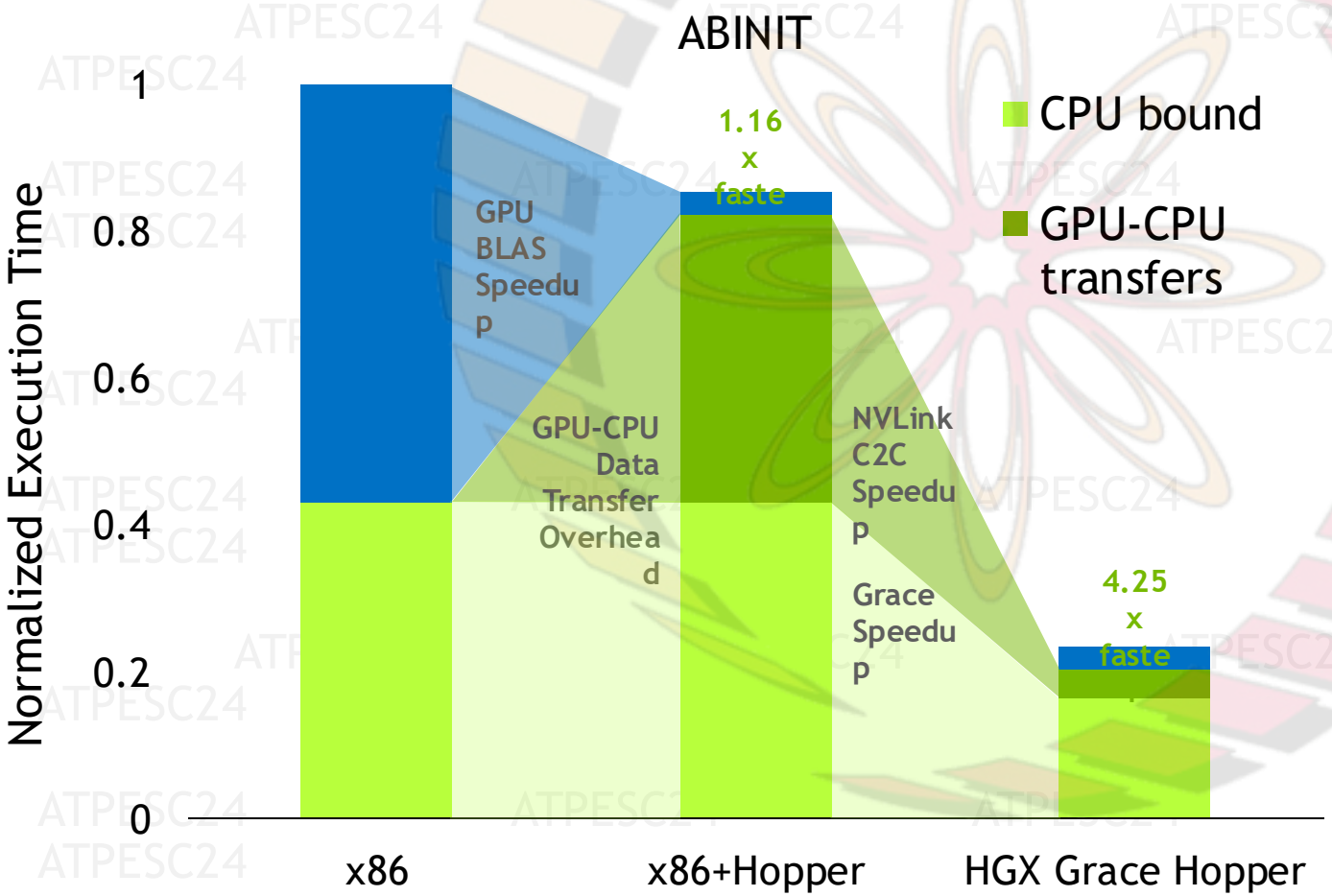
Job A – 64 Grace CPU MPAM
Job B – 8 Grace Cores MPAM + Hopper GPU



Job A – 8 Grace Cores MPAM + Hopper MIG
Job B – 8 Grace Cores MPAM + Hopper MIG
Etc.

GRACE+HOPPER MAKES ACCELERATION MORE ACCESSIBLE

Delivers Superior Performance and Efficiency for HPC



White Paper - Grace Hopper Superchip Architecture

A performance simulation for ABINIT with NVBLAS featuring Titanium 255 Atoms using the LOBPCG algorithm

PARALLEL EXPRESSION

- Moving from Prescriptive to descriptive parallel expression
  - ISO-Language Parallelism
- Parallel Expression as a Computational Graph
- Decouple Scheduling from the expression of algorithm (aka Halide/XLA)
- OS/Runtime provide HW capabilities as hints to the scheduler

- Lowering through MLIR
- Scheduler/Runtime tools like XLA

CONCLUSIONS

Grand Challenge applications are becoming complex, Heterogeneous, data driven and ML/DL aided

Silicon Process Technology alone loosing steam

Architectural innovations, co-Design, ML/DL aiding the continuation of “Moore’s Observation”
Grace-Hopper architecture is a step in that direction

Accurate definition of performance is critical

Efficiently mapping Heterogeneous applications to heterogeneous HW is becoming complex
STD PAR, MLIR, XLA

Power and Energy are becoming the hard barriers to performance