DE LA RECHERCHE À L'INDUSTRIE

Architecture evolutions for HPC

30 Novembre 2020

Guillaume Colin de Verdière

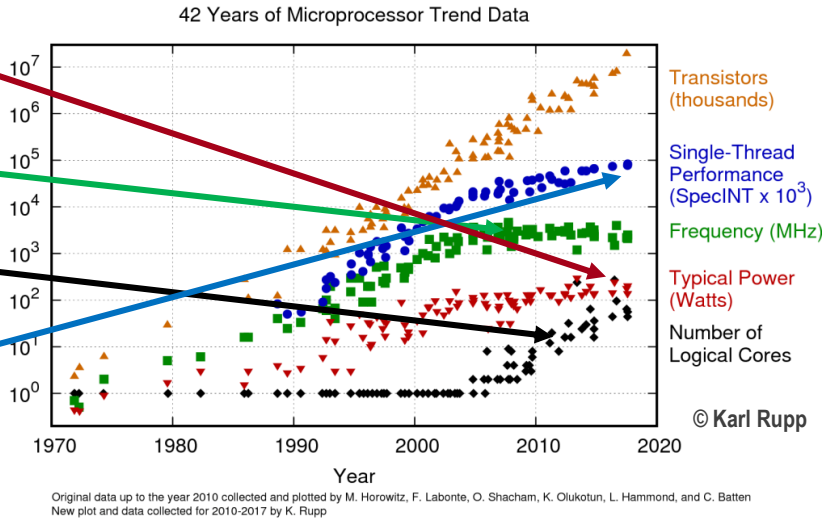
EVOLUTION DRIVER: TECHNOLOGICAL CONSTRAINTS

- Power wall
- Scaling wall
- Memory wall
- Towards accelerated architectures
- SC20 main points

$$P = cV^2F$$

P: power
V: voltage
F: frequency

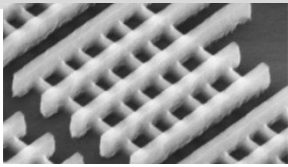
- Reduce V
  - Reuse of embedded technologies
- Limit frequencies
- More cores
- More compute, less logic
  - SIMD larger
  - GPU like structure



https://software.intel.com/en-us/blogs/2009/08/25/why-p-scales-as-cv2f-is-so-obvious-pt-2-2

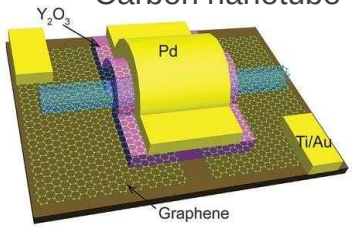
SCALING WALL

FinFET

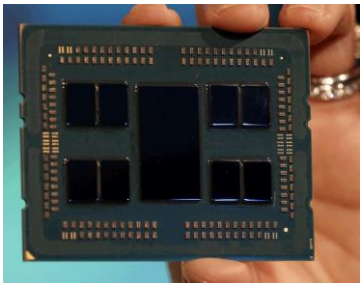


- Moore's law comes to an end
  - Probable limit around 3 - 5 nm
  - Need to design new structure for transistors

Carbon nanotube



- Limit of circuit size
  - Yield decrease with the increase of surface
    - Chiplets will dominate



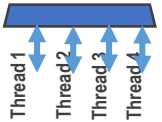
© AMD

  - Data movement will be the most expensive operation
    - 1 DFMA = 20 pJ, SRAM access= 50 pJ, DRAM access= 1 nJ (source NVIDIA)

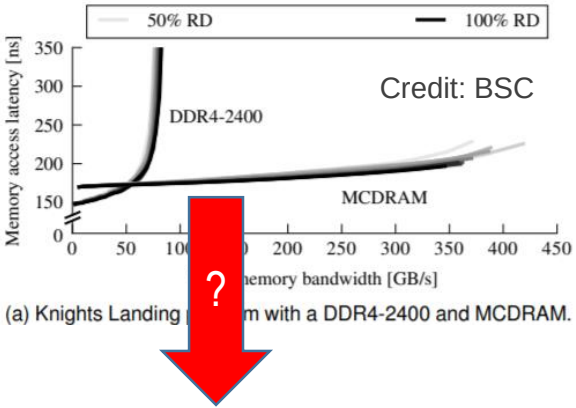
1 nJ = 1000 pJ, Φ Si =110 pm

MEMORY WALL

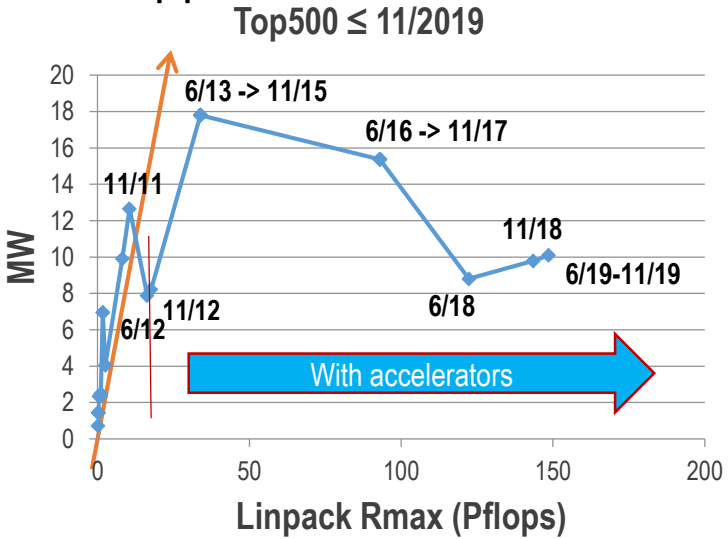
- Better bandwidth with HBM
  - DDR5 @ 5200 MT/s 8ch = **0.33 TB/s**
  - HBM2 @ 4 stacks = **1.64 TB/s**
- Latencies don't improve
  - More hyperthreads
- Deeper memory hierarchy
  - caches + HBM + DDR (+ NVM)
- Impact of non volatile memories?



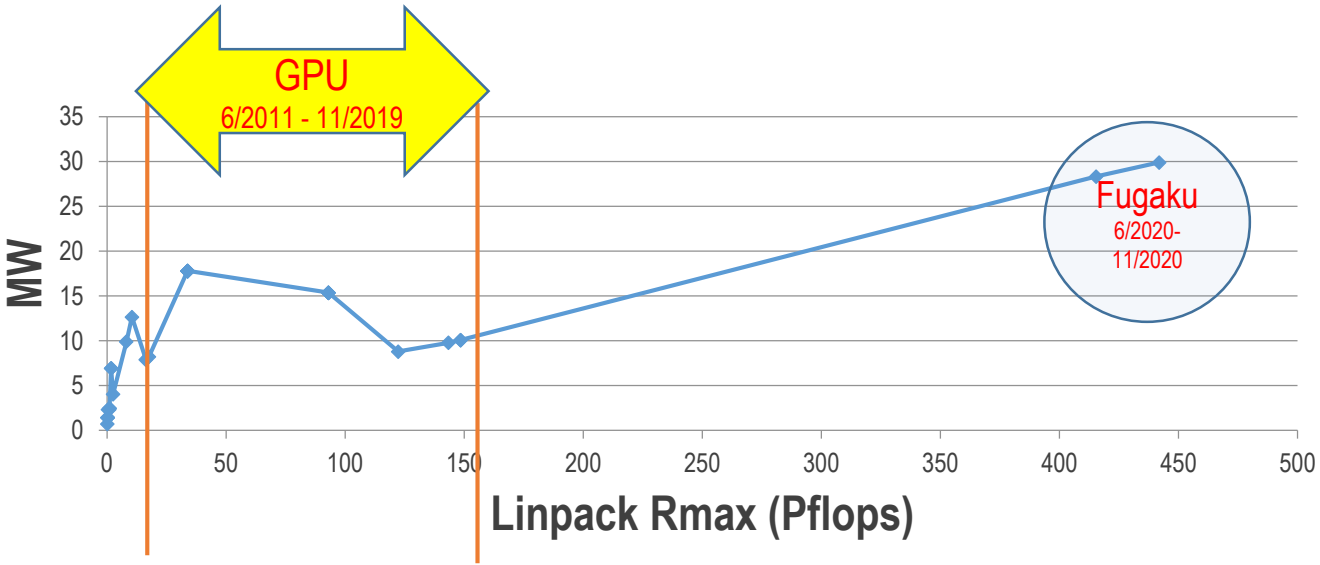
Skylake:	*SMT2*
ThunderX2:	*SMT4*
KNL:	*SMT4*
Power8:	*SMT8*



- Exaflop with regular CPUs will be too power hungry
- One or more accelerator per node
- Accelerator type will depend on applications
  - GPU
    - Compute and IA
  - FPGA
  - TPU
  - DSP
  - Neuromorphic
  - Quantum accelerator

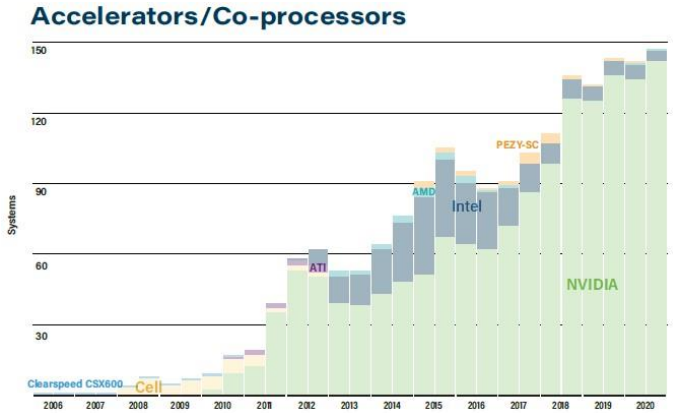


ACCELERATION IS STILL REQUIRED ☺



- All known (pre)Exascale machines are accelerated

Year	Site	Name	Vendors	Technology	Exaflops
2021	Argonne	Aurora	Intel + HPE(Cray)	CPU+GPU Intel	1
2021	Cineca	Leonardo	Atos	CPU Intel + GPU Nvidia	0.2+
2022	Oak Ridge	Frontier	HPE(Cray) + AMD	CPU+GPU AMD	1.5
2023	Livermore	El Capitan	HPE(Cray)	CPU+GPU AMD	> 2



© TOP500.org

SC20 announcements

GPU HARDWARE ANNOUNCEMENTS @ SC20

- NVIDIA
  - A100 80 GB — 9.7 TF FP64 / 19.5 TF FP64 TC
- INTEL
  - Ponte Vecchio — announcements in 2021 for Aurora
  - Xe-HP — 41908 GFLOPS FP32
- AMD
  - MI 100 — 11.5 TF FP64

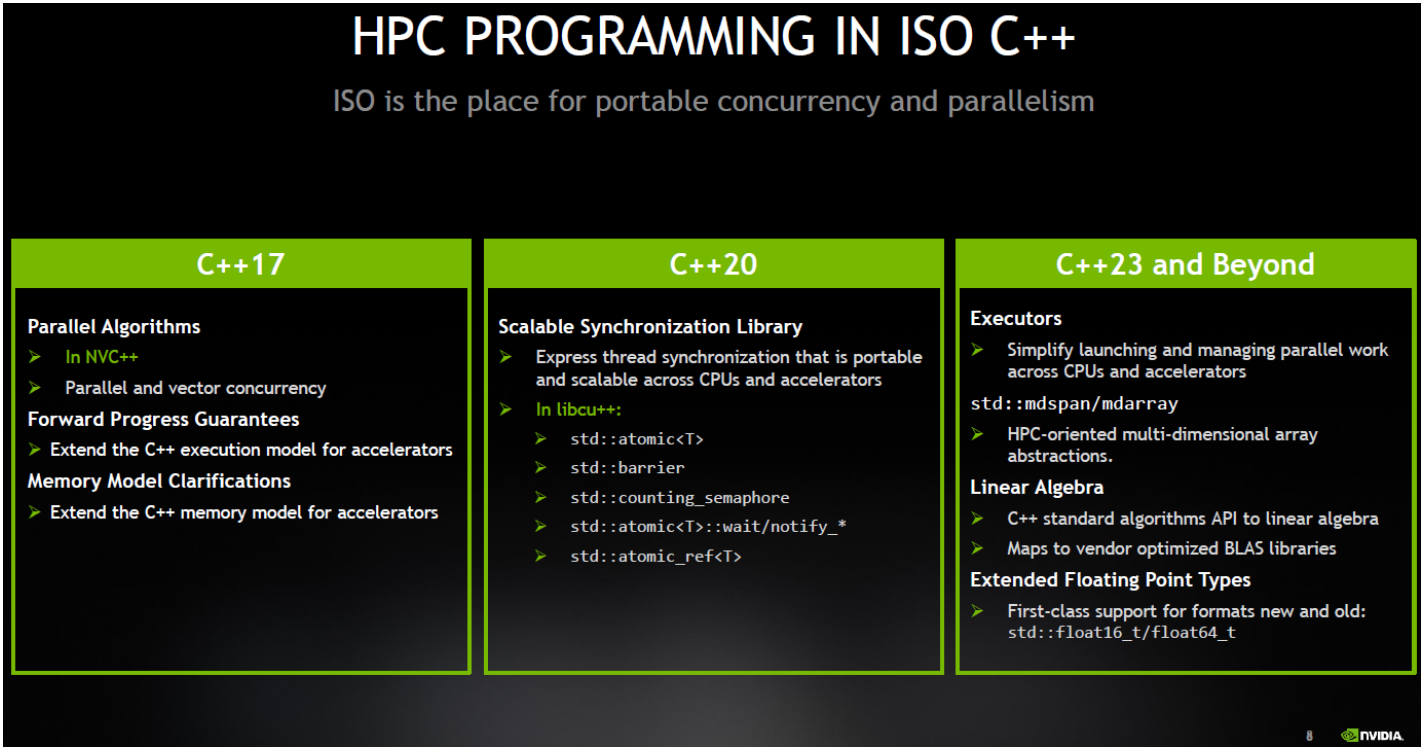
Note : All connected via PCI-Express
⇒ Data management/movement of utter importance
⇒ No hardware help before some years

- Big trends
  - C++ is getting momentum => new codes
  - Fortran is loosing attraction => what about legacy codes ?
- In details

	Intel	AMD	NVIDIA
Directives	OpenMP 5.0	OpenMP 5.0	OpenMP 4.5+ OpenACC
Low level		HIP	CUDA
C/C++	OpenCL/SYCL DPC++	OpenCL/SYCL	OpenCL/SYCL DPC++ (codeplay)

- OpenMP 5.1 announced

- Intel & NVIDIA push to extend C++ standard towards heterogeneous programming



TOOLS AND LIBRARIES

- Intel and NVIDIA have a pretty comprehensive suite
  - Intel Advisor Offload
  - NVIDIA software acceleration of libraries
- AMD is improving on this front
- All vendors are investing on / contributing to LLVM
  - Including Arm ☺

CONCLUSION

- It is high time to program for GPUs
- We have various hardware for performance portability testing
- OpenMP looks like the best bet for the future
  - Especially for Legacy codes and Fortran codes
- C++ might be the best option in the long run

DE LA RECHERCHE À L'INDUSTRIE

Questions?

EUROfusion 30/11/2020

EUROfusion webinar on GPUs #6
MONDAY 30 November : 11h00 am- 11h30 am

Chair France Boillod-Cerneux (CEA)

EUROfusion

SC20 MATERIAL

A compilation by France Boillod-Cerneux

At CEA

cea

EUROfusion https://indico.euro-fusion.org/category/48/

This work has been carried out within the framework of the EUROfusion Consortium and has received funding from the Euratom research and training programme 2014-2018 under grant agreement No 633053. The views and opinions expressed herein do not necessarily reflect those of the European Commission.

YouTube https://www.youtube.com/channel/UCQNxXoQUPMOo_ETr09N3OtA

AMD

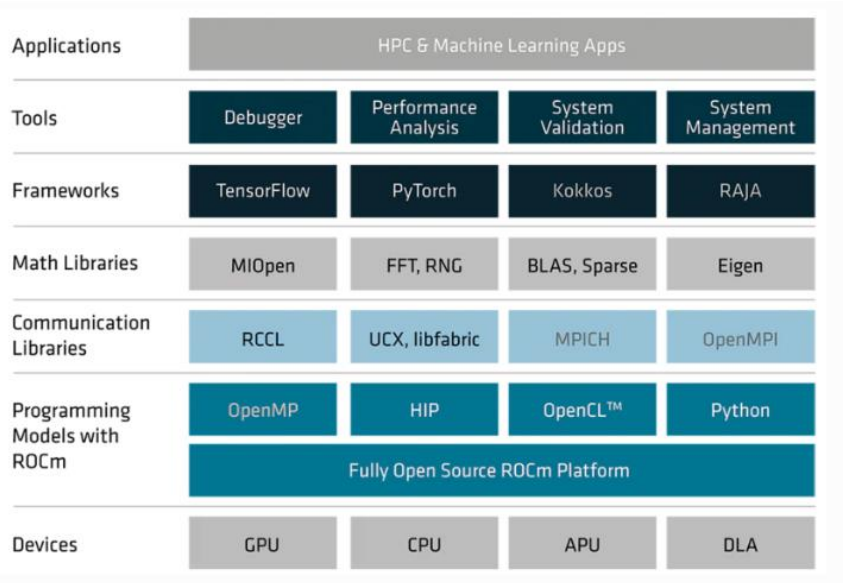
AMD

- AMD Instinct M100
- Sources
  - https://www.amd.com/en/products/server-accelerators/instinct-mi100?utm_source=pardot&utm_content=&utm_campaign=2020-11-17-commercial-server-mi100-launch-en-email&utm_medium=email&utm_term=btn
  - https://rocmdocs.amd.com/en/latest/
  - pdf

AMD

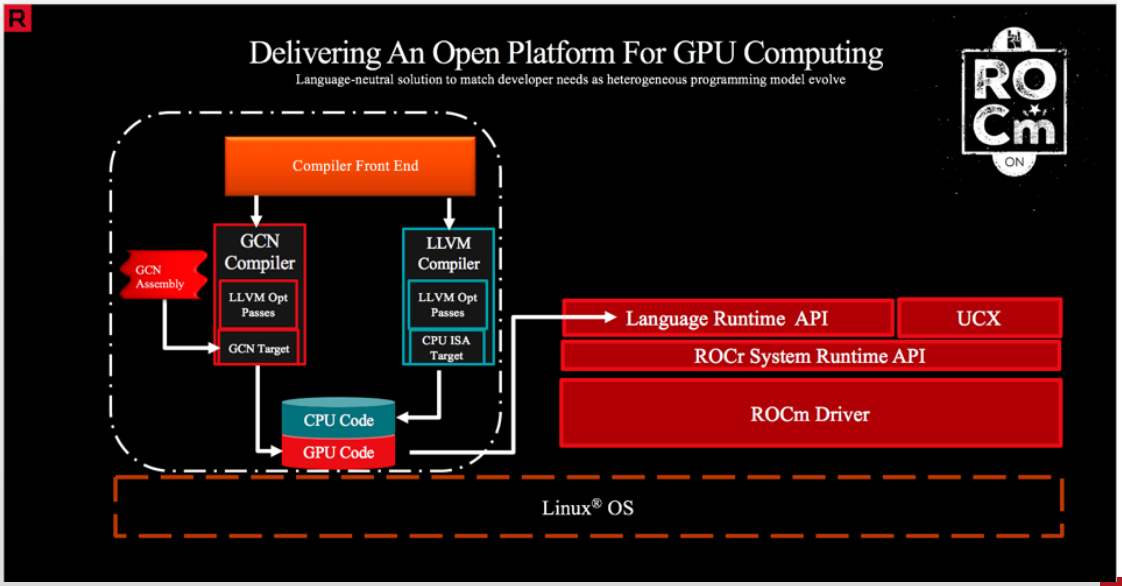
- AMD Instinct M100
  - SC20 releasing the M100 information
    - Delivering up to 11.5 TFLOPs of double precision (FP64) theoretical peak performance
  - Target: HPC + AI
- Compared to previous AMD generation:
  - HPC applications and a substantial up-lift in performance over previous gen AMD accelerators
  - ~~The MI100 delivers up to a 74% generational double precision performance boost for HPC applications.[13]~~

- AMD Software
  - AMD ROCm is the first open-source software development platform for HPC/Hyperscale-class GPU computing
  - UNIX philosophy of choice, minimalism and modular software development to GPU computing
  - Currently: ROCm 4.0



- AMD Software
  - AMD ROCm ecosystem is comprised of open technologies:
    - frameworks (Tensorflow / PyTorch),
    - libraries (MIOpen / Blas / RCCL),
    - programming model (HIP),
    - inter-connect (OCD) and up streamed Linux® Kernel support,
    - Tools, guidance and insights are shared freely across the ROCm GitHub community and forums
      - AMD ROCm profiler, debugger…
  - AMD ROCm is built for scale
    - support multi-GPU computing in and out of server-node communication through RDMA
    - AMD ROCm also simplifies the stack when the driver directly incorporates RDMA peer-sync support.

- AMD Software
  - **AMD ROCm Programming-Language Run-Time**
    - The AMD ROCr System Runtime is language independent and makes heavy use of the Heterogeneous System Architecture (HSA) Runtime API.
    - provides a rich foundation to execute programming languages such as HCC C++ and HIP



- AMD Software
  - **AMD ROCm Programming-Language Run-Time**
    - Multi-GPU coarse-grain shared virtual memory
    - Process concurrency and preemption
    - Large memory allocations
    - HSA signals and atomics
    - User-mode queues and DMA
    - Standardized loader and code-object format
    - Dynamic and offline-compilation support
    - Peer-to-peer multi-GPU operation with RDMA support
    - Profiler trace and event-collection API
    - Systems-management API and tools
  - **Solid Compilation Foundation and Language Support**
    - LLVM compiler foundation
    - HCC C++ and HIP for application portability
    - GCN assembler and disassembler

- AMD Software
  - ROCm 4.0 provides the foundation for exascale computing
    - open source toolset consisting of compilers, programming APIs and libraries
    - ROCm 4.0 has been optimized to deliver performance at scale for MI100-based systems.
    - ROCm 4.0 has upgraded the compiler to be open source and unified to support both OpenMP® 5.0 and HIP.
    - PyTorch and Tensorflow frameworks, which have been optimized with ROCm 4.0, can now achieve higher performance with MI100.
    - ROCm 4.0 is the latest offering for HPC, ML and AI application developers which allows them to create performance portable software

AMD

- Deployed at Oak Ridge
  - performance boosts, up to 2-3x compared to other GPUs
  - recognize is the impact software has on performance
  - ROCm open software platform and HIP developer tool are open source and work on a variety of platforms

- Features of the AMD Instinct MI100 accelerator include:
  - Delivers 11.5 TFLOPS peak FP64 performance and 23.1 TFLOPS peak FP32 performance.
  - Matrix Core technology for HPC and AI delivering single and mixed precision matrix operations, such as FP32, FP16, bFloat16, Int8 and Int4, for converged HPC and AI.
  - 2nd Gen AMD Infinity Fabric Technology – Instinct MI100 provides ~2x the peer-to-peer (P2P) peak I/O bandwidth over PCIe 4.0 with up to 340 GB/s of aggregate bandwidth per card with three AMD Infinity Fabric Links. MI100 GPUs can be configured in a server with up to two fully-connected quad GPU hives, each providing up to 552 GB/s of P2P I/O bandwidth for fast data sharing.
  - 32GB high-bandwidth HBM2 memory at a clock rate of 1.2 GHz delivering 1.23 TB/s of memory bandwidth to support large data sets and help eliminate bottlenecks in moving data in and out of memory.
  - Support for PCIe Gen 4.0 providing up to 64GB/s peak theoretical transport data bandwidth from CPU to GPU.

AMD

- Deployed at Oak Ridge
  - performance boosts, up to 2-3x compared to other GPUs
  - recognize is the impact software has on performance
  - ROCm open software platform and HIP developer tool are open source and work on a variety of platforms

Key Features

PERFORMANCE

Compute Units	120
Stream Processors	7,680
Peak BFLOAT16	Up to 92.3 TFLOPS
Peak INT4 | INT8	Up to 184.6 TOPS
Peak FP16	Up to 184.6 TFLOPS
Peak FP32 Matrix	Up to 46.1 TFLOPS
Peak FP32	Up to 23.1 TFLOPS
Peak FP64	Up to 11.5 TFLOPS
Bus Interface	PCIe® Gen 3 and Gen 4 Support[3]

MEMORY

Memory Size	32GB HBM2
Memory Interface	4,096Bits
Memory Clock	1.2 GHz
Memory Bandwidth	Up to 1.2 TB/s

RELIABILITY

ECC (Full-chip)	Yes[4]
RAS Support	Yes[5]

SCALABILITY

Infinity Fabric™ Links	3
OS Support	Linux® 64-bit
AMD ROCm™ Compatible	Yes

BOARD DESIGN

Board Form Factor	Full-Height, Dual Slot
Length	10.5" Long
Thermal	Passively Cooled
Max Power	300W TDP
Warranty	Three Year Limited[6]

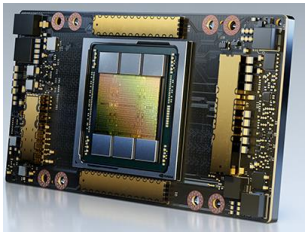
NVIDIA

NVIDIA

- Source
- https://insidehpc.com/2020/11/nvidia-announces-a100-80gb-gpu/
- https://nvidianews.nvidia.com/news/nvidia-doubles-down-announces-a100-80gb-gpu-supercharging-worlds-most-powerful-gpu-for-ai-supercomputing
- https://www.nvidia.com/en-us/data-center/a100/
- https://www.hpcwire.com/2020/11/16/nvidia-unveils-a100-80gb-gpu-powerhouse-supercomputing-chip/
- https://www.tomshardware.fr/nvidia-lance-un-gpu-ampere-a100-avec-80-go-de-memoire-hbm2e/

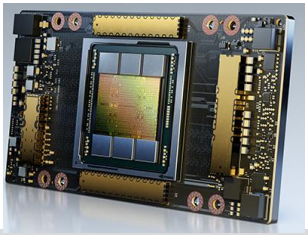
NVIDIA

- The NVIDIA® A100 Tensor Core GPU acceleration at every scale to power AI, data analytics, and HPC applications
- A100 provides up to 20X higher performance over the prior NVIDIA Volta™ generation
- A100 can efficiently scale up or be partitioned into seven isolated GPU instances,
  - with Multi-Instance GPU (MIG) providing a unified platform
  - enables elastic data centers to dynamically adjust to shifting workload demands



NVIDIA

- NVIDIA A100 Tensor Core technology supports
  - a broad range of math precisions, providing a single accelerator for every workload.
- A100 80GB doubles GPU memory and as a memory bandwidth at 2 terabytes per second (TB/s)



NVIDIA

- Ampere architecture
  - Whether using MIG to partition an A100 GPU into smaller instances, or NVIDIA NVLink® to connect multiple GPUs to speed large-scale workloads,
  - A100 can readily handle different-sized acceleration needs, from the smallest job to the biggest multi-node workload.
  - A100 versatility means IT managers can maximize the utility of every GPU in their data center, around the clock.

- MULTI-INSTANCE GPU (MIG)
  - An A100 GPU can be partitioned into as many as seven GPU instances, fully isolated at the hardware level with their own high-bandwidth memory, cache, and compute cores.
  - MIG gives developers access to breakthrough acceleration for all their applications, and IT administrators can offer right-sized GPU acceleration for every job, optimizing utilization and expanding access to every user and application.

NVIDIA

- THIRD-GENERATION TENSOR CORES
  - NVIDIA A100 delivers 312 teraFLOPS (TFLOPS) of deep learning performance.
  - That's 20X the Tensor FLOPS for deep learning training and 20X the Tensor TOPS for deep learning inference, compared to NVIDIA Volta GPUs.

- HBM2E
  - With up to 80 gigabytes (GB) of high-bandwidth memory (HBM2e), A100 delivers a world's first GPU memory bandwidth of over 2TB/sec, as well as higher dynamic random-access memory (DRAM) utilization efficiency at 95%.
  - A100 delivers 1.7X higher memory bandwidth over the previous generation

NVIDIA

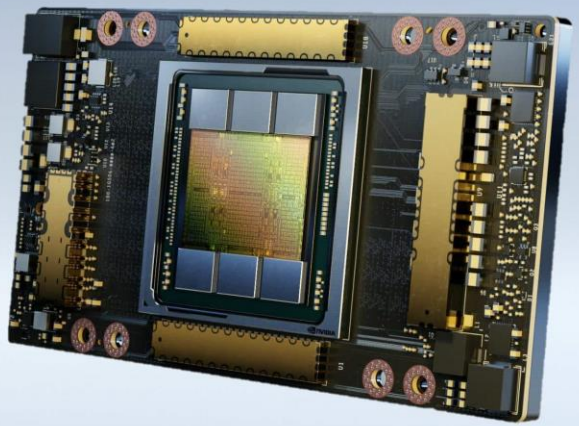
- NEXT-GENERATION NVLINK
  - NVIDIA NVLink in A100 delivers 2X higher throughput compared to the previous generation.
  - When combined with NVIDIA NVSwitch™, up to 16 A100 GPUs can be interconnected at up to 600 gigabytes per second (GB/ sec), unleashing the highest application performance possible on a single server.
  - NVLink is available in A100 SXM GPUs via HGX A100 server boards and in PCIe GPUs via an NVLink Bridge for up to 2 GPUs.

- STRUCTURAL SPARSITY
  - AI networks have millions to billions of parameters.
  - Not all of these parameters are needed for accurate predictions, and some can be converted to zeros, making the models “sparse” without compromising accuracy.
  - Tensor Cores in A100 can provide up to 2X higher performance for sparse models.
  - While the sparsity feature more readily benefits AI inference, it can also improve the performance of model training.

- NVIDIA A100 SC20 release

SUPERCHARGED AI SUPERCOMPUTING WITH DOUBLE THE MEMORY

1.8x Faster Simulations Quantum Espresso	**2x** Faster Big Data Analytics 10 TB Retail Benchmark	**3x** AI Training DLRM Recommender
1.25x MIG Inference RNN-T Speech Recognition	**1.25x** Energy Efficiency Shatters 25 GF/W	



Speedups Normalized to Number of GPUs | Comparisons A100 80GB to A100 40GB | Measurements performed DGX A100 servers
AI Training running DLRM using Huge CTR framework on a 450 GB Criteo dataset. Normalized speedup ~2.6X | Data Analytics: big data benchmark with 10TB dataset, 30 analytical retail queries, ETL, ML, NLP. Normalized speedup ~1.9X
HPC: Quantum Espresso with CNT10POR8 dataset on a 1.6TB dataset. Normalized speedup ~1.8X | AI Inference: RNN-T Single stream latency on A100 80GB on 1MIG@10GB when configured for 7MIGs. Normalized speedup ~1.25X

- NVIDIA A100 SC20 release

NEW DGX A100 640GB SYSTEM

For the Largest AI Workloads

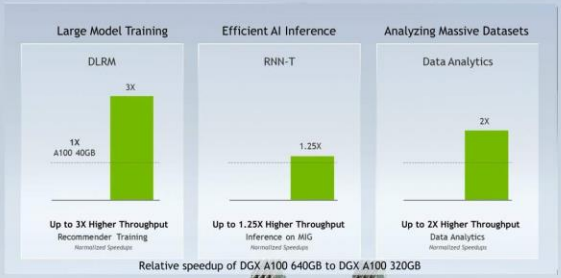
640 GB of GPU memory per system to increase model accuracy and reduce-time-to-solution

Up to 3X higher throughput for large-scale workloads

Double the GPU memory for MIG for more flexible AI development, analytics, and inference

Available individually, or part of DGX SuperPOD Solution for Enterprise

Upgrade option for current DGX A100 customers



Speedups Normalized to Number of GPUs | Comparisons to A100 40GB | Measurements performed DGX A100 servers . AI Training: DLRM (Huge CTR) | DGX A100: 16x A100 40GB vs 8x A100 80GB | speedup = 1.4X. | Speedup normalized to number of GPUs = 2.8X. Data Analytics: big data benchmark with RAPIDS(0.16), BlazingSQL(0.16), DASK(2.2.0) | 30 analytical retail queries, ETL, ML, NLP | 96x A100 40GB vs 48x A100 80GB | AI Inference: RNN-T (MLPerf 0.7 Single stream latency) | DGXA100: A100 40GB vs A100 80GB on 1MIG@10GB when configured for 7MIGs

NVIDIA

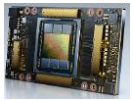
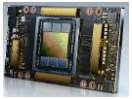
- NVIDIA A100 vs V100



GPU	A100 (80 Go)	A100 (40 Go)	V100
Cœurs CUDA FP32	6912	6912	5120
Fréquence Boost	1,41 GHz	1,41 GHz	1,53 GHz
Vitesse mémoire	3,2 Gbit/s HBM2e	2,4 Gbit/s HBM2	1,75 Gbit/s HBM2
Taille bus mémoire	5120-bit	5120-bit	4096-bit
Bande passsante mémoire	2,0 To/sec	1,6 To/sec	900 Go/sec
VRAM	80 Go	40 Go	16 Go/32 Go
Simple précision	19,5 TFLOPs	19,5 TFLOPs	15,7 TFLOPs
Double précision	9,7 TFLOPs	9,7 TFLOPs	7,8 TFLOPs
INT8 Tensor	624 TOPs	624 TOPs	N/A
FP16 Tensor	312 TFLOPs	312 TFLOPs	125 TFLOPs
TF32 Tensor	156 TFLOPs	156 TFLOPs	N/A

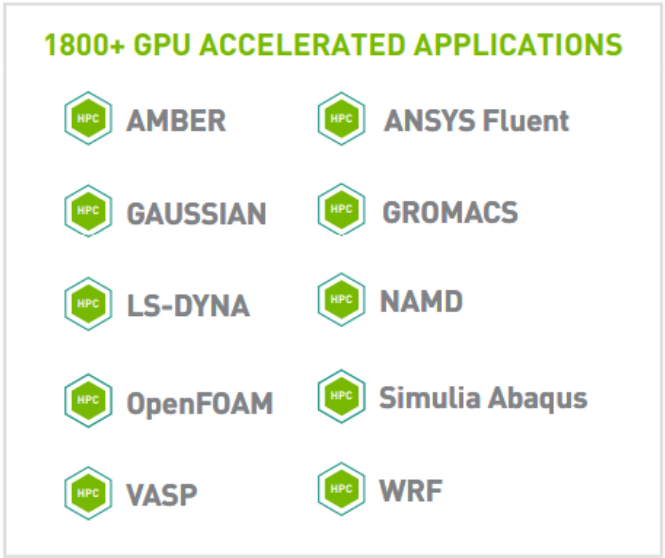
NVIDIA

- NVIDIA A100 vs V100


GPU	A100 (80 Go)	A100 (40 Go)	V100
Interconnexion	NVLink 3 – 12 Liens (600 Go/sec)	NVLink 3 – 12 Liens (600 Go/sec)	NVLink 2 – 6 Liens (300 Go/sec)
GPU	GA100 (826 mm2)	GA100 (826 mm2)	GV100 (815 mm2)
Nombre de transistors	54,2 milliards	54,2 milliards	21,1 milliards
TDP	400 W	400 W	300 W/350 W
Process	TSMC 7N	TSMC 7N	TSMC 12 nm FFN
Interface	SXM4	SXM4	SXM2/SXM3
Architecture	Ampere	Ampere	Volta

- NVIDIA ecosystem



NVIDIA

- A100 80GB GPU for the Nvidia HGXTM AI supercomputing platform
  - twice the memory of its predecessor.
  - with HBM2e doubles the A100 40GB GPU's high-bandwidth memory to 80GB and delivers more than 2TB/sec of memory bandwidth, according to Nvidia
  - The A100 80GB GPU is available in Nvidia DGX A100 and DGX Station A100 systems

- A100 80GB version is intended for a range of applications with large data memory requirements.

- For AI training, recommender system models like DLRM have large tables representing billions of users and billions of products.
  - A100 80GB delivers up to a 3x speed-up, designed for rapid retraining of models.

- A100 can be partitioned into up to seven GPU instances, each with 10GB of memory, according to the company

NVIDIA

- On a big data analytics benchmark for retail in the terabyte-size range, the A100 80GB boosts performance up to 2x, “making it an ideal platform for delivering rapid insights on the largest of datasets. Businesses can make key decisions in real time as data is updated dynamically,” Nvidia said.

- For scientific applications, such as weather forecasting and quantum chemistry, the A100 80GB can deliver acceleration – Quantum Espresso, a materials simulation, achieved throughput gains of nearly 2x with a single node of A100 80GB.

- Nvidia said the A100 80GB includes the following features of the Nvidia Ampere architecture:
  - Third-Generation Tensor Cores: Provides up to 20x AI throughput of the previous Volta generation with a new format TF32, as well as 2.5x FP64 for HPC, 20x INT8 for AI inference and support for the BF16 data format.
  - HBM2e GPU Memory: Doubles the memory capacity and is the first in the industry to offer more than 2TB per second of memory bandwidth.
  - MIG technology: Doubles the memory per isolated instance, providing up to seven MIGs with 10GB each.
  - Structural Sparsity: Delivers up to 2x speedup inferencing sparse models.
  - Third-Generation NVLink and NVSwitch: Provides twice the GPU-to-GPU bandwidth of the previous generation interconnect technology, accelerating data transfers to the GPU for data-intensive workloads to 600 gigabytes per second.
  - The A100 80GB GPU is part of the Nvidia HGX AI supercomputing platform, which brings together Nvidia GPUs, NVLink, InfiniBand networking and an AI and HPC software stack.

- NVIDIA A100 SC20 release

SERVER-CLASS SOLUTION IN OFFICE-FRIENDLY FORM

Data Center Technology Outside the Data Center

First and only workstation with 4-way NVIDIA A100, NVLink and MIG

- Four A100 Tensor Core GPUs, 320GB total HBM2E
- Multi-Instance GPU (MIG)
- 3rd generation NVLink
- 200GB/s bi-directional bandwidth between any GPU pair, almost 3x compared to PCIe Gen4
- New Maintenance-free Refrigerant Cooling System

CPU and Memory

- 64-core AMD® EPYC® CPU, PCIe Gen4
- Up to 512GB system memory

Internal Storage

- 1.92 TB NVME M.2 SSD for OS, up to 7.68TB NVME U.2 SSD for data cache

Connectivity

- 2x 10GbE (RJ45)
- 4x Mini DisplayPort for display out
- Remote management 1GbE LAN port (RJ45)


INTEL

INTEL

- Sources
  - https://insidehpc.com/2020/11/a-bridge-to-ponte-vecchio-argonne-aurora-developers-using-substitute-intel-xe-hp-gpus-oneapi-for-scientific-applications/
  - https://www.hpcwire.com/2020/11/17/intel-xe-hp-gpu-aurora-exascale-development/
  - https://newsroom.intel.com/news/intel-xpu-vision-oneapi-server-gpu/#gs.kni86m

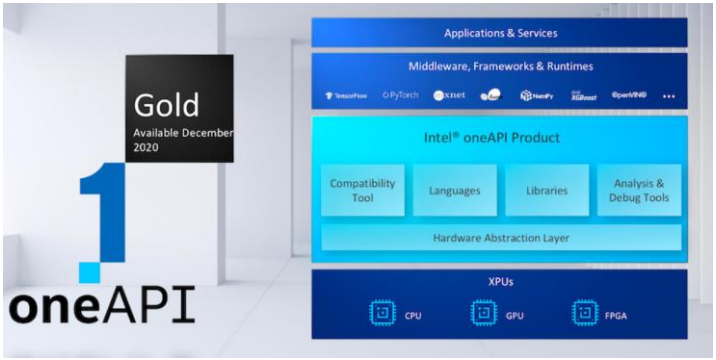
INTEL PONTE VECCHIO

Intel and Argonne National Laboratory partnership with Aurora supercomputer

- GPUs based on Intel's Xe-HP microarchitecture
- Intel oneAPI toolkits for development of scientific applications to be used on the Aurora exascale
  - Later delivery of 'Ponte Vecchio' GPUs, to be deployed in 2022

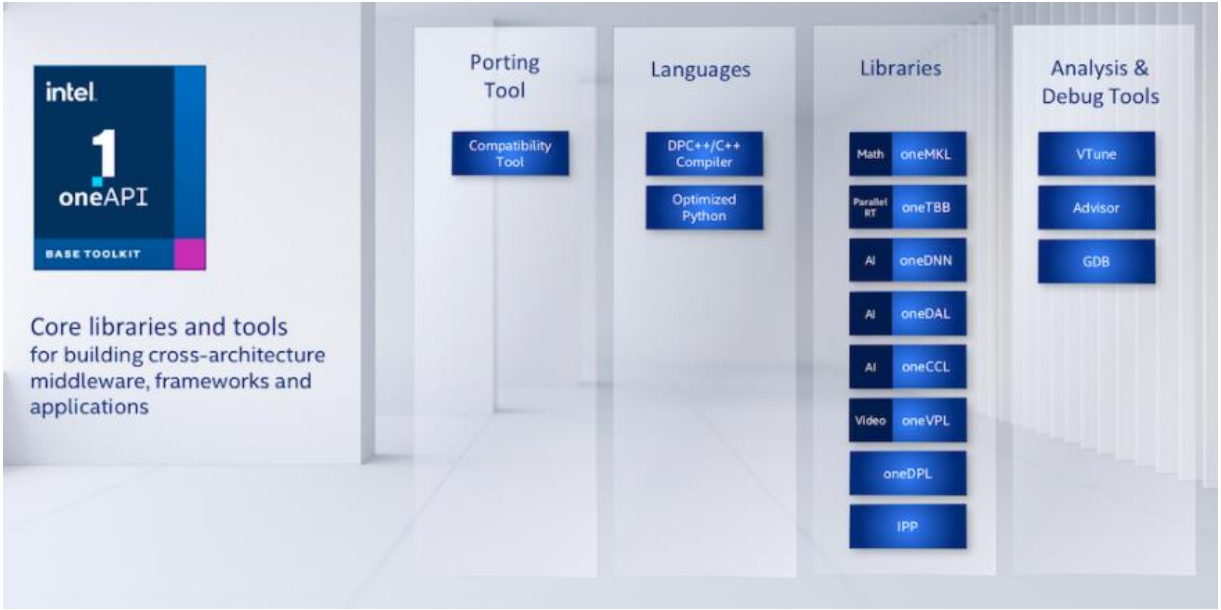
OneAPI:

- Using Intel heterogeneous computing programming environments
- scientific applications are ready for the scale and architecture of the Aurora supercomputer at deployment

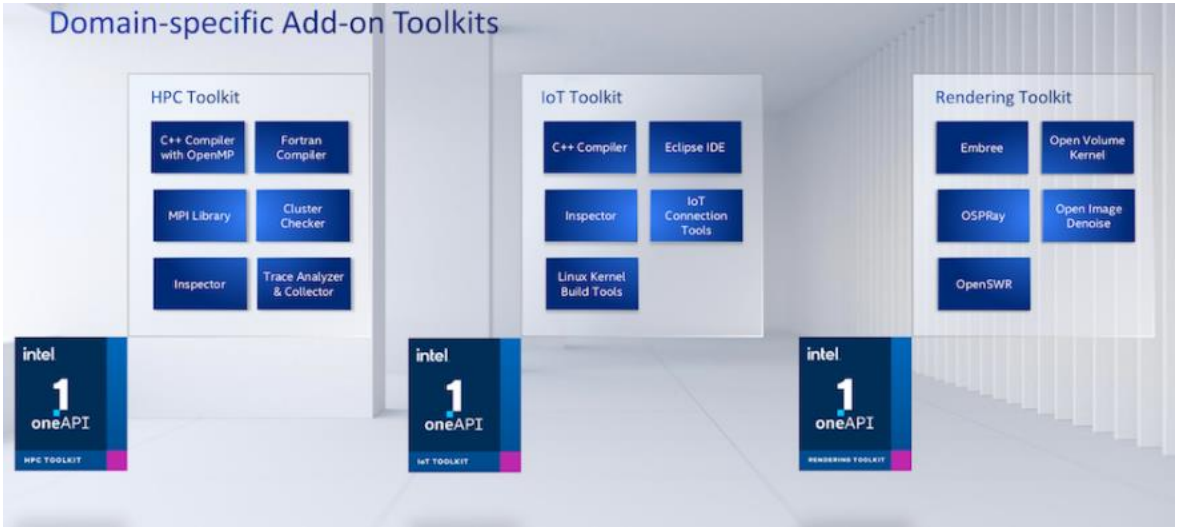


INTEL ONEAPI

OneAPI:



OneAPI:


Domain-specific Add-on Toolkits
HPC Toolkit
C++ Compiler with OpenMP
Fortran Compiler
MPI Library
Cluster Checker
Inspector
Trace Analyzer & Collector
IoT Toolkit
C++ Compiler
Eclipse IDE
Inspector
IoT Connection Tools
Linux Kernel Build Tools
Rendering Toolkit
Embree
Open Volume Kernel
OSPRay
Open Image Denoise
OpenSWR
intel
1
oneAPI
HPC TOOLKIT
intel
1
oneAPI
IoT TOOLKIT
intel
1
oneAPI
RENDERING TOOLKIT

INTEL

Argonne Leadership Computing Facility (ALCF) researchers are using software development platforms based on Intel Xe-HP GPUs

- designed to prepare key applications, libraries and infrastructure for Aurora
- Intel and Argonne teams are working together to co-design, test and validate several exascale applications
- perform software optimizations across Intel CPUs and GPUs
- investigate scenarios that would be difficult to replicate in software-only environments

The Intel Server GPU is based on Xe-LP microarchitecture, Intel's most energy-efficient graphics architecture, offering a low-power, discrete system-on-chip design, with a 128-bit pipeline and 8GB of dedicated onboard low-power DDR4 memory.