Nvidia Hopper GPU and Grace CPU Highlights

Anne C. Elster[1] and Tor A. Haugdahl[1]

[1]Affiliation not available

March 28, 2022

Abstract

At GTC 2022 Nvidia announced a new product family that aims to cover from small enterprise workloads through exascale HPC and trillion-parameter AI models. This column highlights the most interesting features of their new Hopper GPU and Grace CPU computer chips and the Hopper product family. We also discuss some of the history behind Nvidia technologies and their most useful features for computational scientists such as the Hopper DPX dynamic programming instruction set, increased number of SMs, and FP 8 tensor core availability. Also included are descriptions of the new Hopper Clustered SMs architecture and updated NVSwitch technologies that integrates their new ARM-based Grace CPU.

Systems with GPUs have impacted scientific computing since CUDA was first released in June 2007. The first CUDA tutorial was held at SC07 the same year (*). After just three years, Nvidia GPUs were featured as accelerators in several of the Top- 10 of the top500.org list. Nvidia was, at that time, still viewed primarily as a company focused on the gaming market. Today Nvidia is marketing itself as an AI technology company that continues to build products of interest to computational scientists.

(*) Author 1 remembers having to push for the CUDA Tutorial for SC07 as many on the Tutorials Committee were skeptical.

In this feature, we will highlight the most interesting features of the new Hopper and Grace computer chips and the Hopper product family (Fig. 1), just announced by Nvidia at their GTC 2022 event (*Nvida Hopper Architecture In-Depth — NVIDIA Technical Blog*, n.d.) and detailed in the Hopper Whitepaper (*Nvidia Hopper GPU Architecture*, n.d.). The product family aims to cover from small enterprise workloads to exascale HPC and trillion-parameter AI models. We will include some of the history behind the Nvidia technologies and their most useful features for computational scientists.

Figure 1: Nvidia Hopper product family with racked pods as background (extract from Nvidia source)

Connecting the CPU and GPU—Intel and PCIe

In early 2009 Intel sued Nvidia over their 2004 chipset licensing agreement that let Nvidia make core-logic (chipsets) for Intel in exchange for Intel licensing NVIDIA's 3D, GPU and other patents. Nvidia made the chips for their nForce series chipset and the two-chip ION Platform. ION's combination of Nvidia and Intel chipsets provided a 10x performance advantage over Intel-only chipsets and was popular in Apple laptop computers, etc. Intel claimed their agreement did not cover their newer Nehalem architecture, which featured an integrated memory controller. Nvidia chose to countersue. The USD 1.5 billion settlement in January 2011 barred Nvidia from making CPUs with Intel's x86 technology. Nvidia GPUs have since been relegated to connecting to Intel CPUs via the relatively slow PCIe buses. This is also why, at present, IBM's Power CPUs are connected to Nvidia's GPUs over NVLINK, while no such connections exist for Intel CPUs. AMD have similarly connected their CPUs and GPUs via their HyperTransport technology.

Nvidia interconnects: Mellanox, NVLink and NVSwitch

By announcing the new ARM-based Grace CPU at GTC2022 (available in 2023), Nvidia will again provide fast links between CPU and GPU as they will be connected via their upgraded NVLink technology. Nvidia's fourth-generation NVLink technology provides 1.5x higher bandwidth compared to the previous generation, and improved scalability for multi-GPU system configurations. A single Nvidia H100 Tensor Core GPU supports up to 18 NVLink connections for a total bandwidth of 900 gigabytes per second (GB/s)—over 7x the bandwidth of PCIe Gen5 .

Nvidia also announced in March 2019 that they had reached an agreement with Mellanox Technologies to buy the company for USD 6.9 Billion (*Nvidia press release on aquiring Mellanox for USD 6.9 Billion*, n.d.). The Israeli-American company is known for its InfiniBand technology used between by high-end servers. This has led Nvidia to develop their interconnected technology further.

Their new NVLink Switch System can support clusters of up to 256 connected H100s and promises to deliver 9x higher bandwidth than InfiniBand HDR on Ampere. In addition, Nvidia is taking advantage of its Mellanox purchase with NVLink now supporting in-network computing called SHARP, previously only available on InfiniBand. Nvidia states that it will deliver one exaFLOPS of FP8 compute performance while delivering 57.6 terabytes/s (TB/s) of All2All bandwidth.

Figure 2: Nvidia DGX H100 SuperPOD (Source: Nvidia)

Embedded, AI, and Crypto

The week before the announcement of the Intel-Nvidia settlement, Nvidia shares rose almost 30% on the announcement of their embedded ARM-based Tegra 2 chips. Nvidia has since made leaps in the embedded market and targeted their more powerful GPUs for AI and Machine Learning, to the point where they today market themselves heavily as an AI tech company that also produces consumer-grade GPUs for gaming and workstation workloads, as well as other embedded technologies.

Nvidia has profited greatly from their high-end GPUs being used for crypto mining, so much so that it has been hard the last couple of years for many gamers and scientists to get a hold of their newest most powerful consumer GPUs.

This article has been accepted for publication in Computing in Science & Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MCSE.2022.3163817

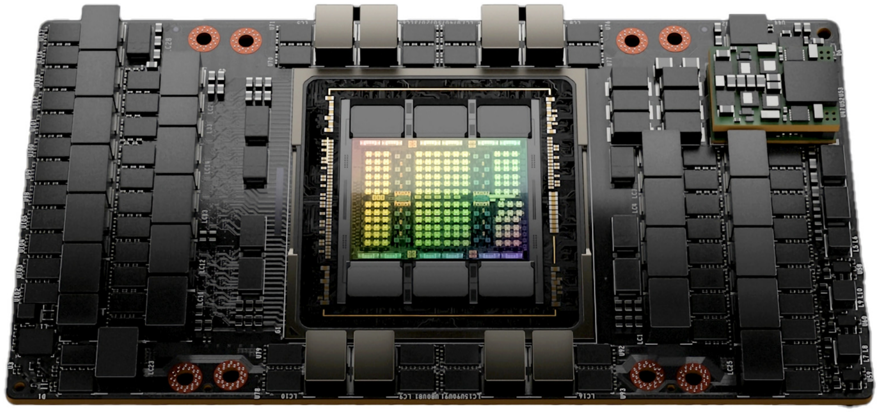

Figure 3: Nvida H100 GPU on new SXM5 Module (Source: Nvidia)

Hopper GPU H100 Overview

The Hopper GPU, introduced as Nvidia H100 Tensor Core GPU, is implemented using the Taiwan Semiconductor (TMSC) 4N process customized for Nvidia with 80 billion transistors. The H100 architecture includes several noteworthy architectural advances.

The custom Nvidia H100 SXM5 module houses the H100 GPU and HBM3 RAM chips, and also provides connection to other systems via their fourth-generation NVLink and PCIe Gen 5 ports. (Figure 3). Note that these modules do not include display connectors, Nvidia RT Cores for ray tracing acceleration, or an NVENC encoder since they, like the A100, are data center modules.

The H100 GPU consists of up to **144 Streaming Multiprocessors** (SM) per full GPU, which have many performance and efficiency improvements over earlier versions.

An overview comparing the A100 and H100 architectures is shown in the following table.

GPU Features	A100	H100 SXM5	H100 PCIe
GPU Architecture	Ampere	Hopper	Hopper
GPU Board Form Factor	SXM4	SXM5	PCIe Gen 5
SMs	108	132	114
TPCs 54 62 57	54	62	57
FP32 Cores / SM	64	128	128
FP32 Cores / GPU	6912	16896	14592
FP64 Cores / SM (excl. Tensor)	32	64	64
FP64 Cores / GPU (excl. Tensor)	3456	8448	7296
INT32 Cores / GPU (64 per SM)	6912	8448	7296
Tensor Cores / GPU (4 per SM)	432	528	456
Peak FP64 TFLOPS (non-Tensor)	9.7	30	24
Texture Units	432	528	456
Memory Interface, all 5120-bit	HBM2	HBM3	HBM2e
Memory Size	40 GB	80 GB	80 GB
Memory Bandwidth(GB/Sec)	1555	3000	2000
L2 Cache Size	40 MB	50 MB	50 MB
Shared Memory Size / SM (Config)	¡= 164 KB	¡= 228 KB	¡= 228KB
Register File Size / SM	256 KB	256 KB	256 KB
Total drawn power	400W	700W	350W
Transistors	54.2 billion	80 billion	80 billion
TSMC Manufacturing Process	7nm N7	4N custom	4N Custom

Table 1: A100 vs. H100 – Comparing Main Features (source: Nvidia)

Hopper Features useful for Scientific Computing

Key new features useful for Scientific computing include: (*Nvidia Hopper GPU Architecture*, n.d.)

- **2x faster clock-for-clock performance per SM** contributes significantly to 3x faster FP32 and FP64 instructions.
- **Fourth-generation Tensor Cores,** which are up to 6x faster chip-to-chip compared to A100, announced to deliver 2x the MMA (Matrix Multiply- Accumulate) computational rates of the A100 SM on equivalent data types, and 4x using the new **FP8 data type**, compared to old FP16.
- **New DPX Instructions** that should accelerate Dynamic Programming algorithms by up to 7x over the A100 GPU. *A more detailed description of DPX will follow in the next section.*
- New **Thread Block Cluster feature** allowing programmatic control of locality at a granularity larger than a single Thread Block on a single SM. *Note that this adds another synchronization layer. This will also be discussed in the next section.*
- New **Asynchronous Execution features** including a new **Tensor Memory Accelerator (TMA).** TMA is designed to transfer large data blocks efficiently between global and shared memory. TMA also supports asynchronous copies between *Thread Blocks in a Cluster*. There is also a new *Asynchronous Transaction Barrier* for doing atomic data movement and synchronization. .
- **HBM3** memory subsystem providing nearly a **2x bandwidth** increase over the previous generation. The H100 SXM5 GPU is the world's first GPU with HBM3 memory delivering a class-leading 3 TB/sec of memory bandwidth.
- **50 MB L2 cache** (versus A100s 40 MB L2) reducing trips to HBM3.
- **Second-generation Multi-Instance GPU (MIG)** technology provides approximately 3x more compute capacity and nearly 2x more memory bandwidth per GPU Instance compared to A100.

This article has been accepted for publication in Computing in Science & Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MCSE.2022.3163817

Tensor Memory Accelerator

The newly added Tensor Memory Accelerator (TMA) enables asynchronous transfers of multidimensional blocks of data. An elected thread within a thread group takes on responsibility for interacting with the TMA by passing along a *Copy Descriptor* detailing the information the TMA needs to correctly transfer a multidimensional block of data, or tensor. The remaining threads are free to perform other instructions while the TMA operation is underway.

Fourth-Generation Tensor Cores

Fourth-generation tensor cores further improve upon the efficiency of the previous generation. Nvidia has now added support for an 8-bit floating-point datatype: **FP8**. They support two flavors of FP8, namely E4M3 and E5M2, enabling the choice between dynamic range or precision. The number following the E and the number following the M represent the number of exponent- and mantissa bits respectively. Generic computations that natively match FP8 ranges are few and far between. In the cases where FP8 is sufficient, one can expect great performance improvements over, e.g., FP16. Nvidia expects their new DGX SuperPOD to be able to deliver 1 exaFLOPS of sparse FP8 compute.

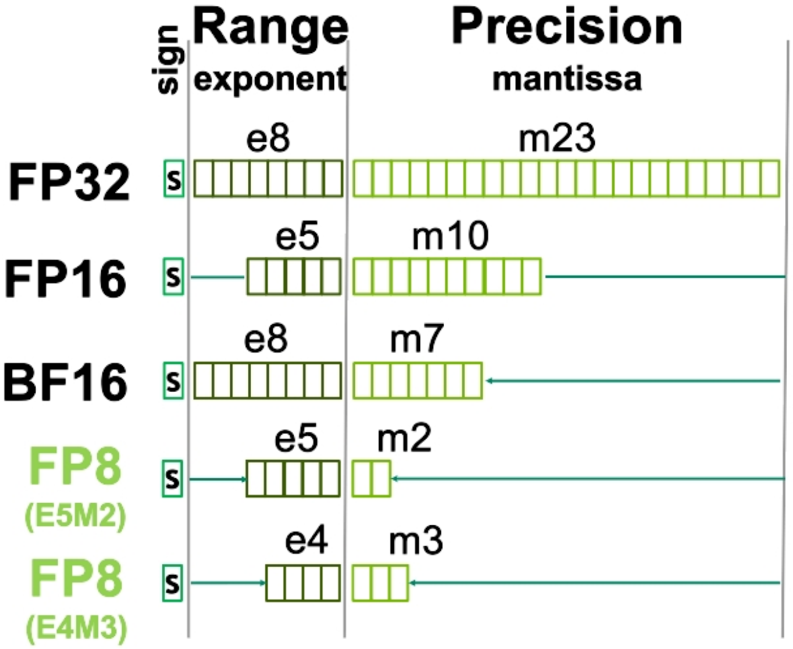

Figure 4: Hopper Floating Point layouts (source: Nvidia)

This article has been accepted for publication in Computing in Science & Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MCSE.2022.3163817

DPX Instructions (*Nvidia Hopper GPU – DPX Instructions — NVIDIA Blog*, n.d.)

Algorithms built upon problems where optimal solutions to subproblems constitute an optimal solution to the problem itself rely on *dynamic programming (DP)*. Application areas include healthcare (e.g., genomics), robotics (e.g., path finding), quantum computing, and data science.

A simple example comes from the Fibonacci numbers. The n-th Fibonacci number is known to be the sum of the two previous Fibonacci numbers. Finding the n-th Fibonacci number is thus solved by recursively solving sub-problems. Furthermore, subproblems of Fibonacci *overlap*. Other DP algorithms include Dijkstra's shortest path, Floyd-Warshall all-pairs shortest path and Smith-Waterman for sequence alignment.

DP problems benefit from the tabulation (building a solution bottom-up) and memoization (top-down) strategies. Both strategies store results of sub-problems such that recomputation is avoided. The new DPX instruction set aims to speed up dynamic programming with specialized instructions that presumably exploit the characteristics of the DP problems.

Thread Block Clusters

The CUDA programming model has long encompassed Threads, Thread Blocks, and Grids. The Hopper architecture adds another level to the hierarchy: *Thread Block Clusters*. The new level in the hierarchy exists between Grids and Thread Blocks. Its functionality enables increased programmatic control of data locality.

Thread Block Clusters further the capabilities of, among others, the CUDA Cooperative Groups API. Thread Blocks participating in Thread Block Clusters are guaranteed to be scheduled concurrently, allowing for finer-grained parallelism across Thread Blocks running on SMs. Going even further, Nvidia introduces a specialized SM-to-SM interconnect network, allowing SMs to exchange shared memory directly instead of through Global Memory.

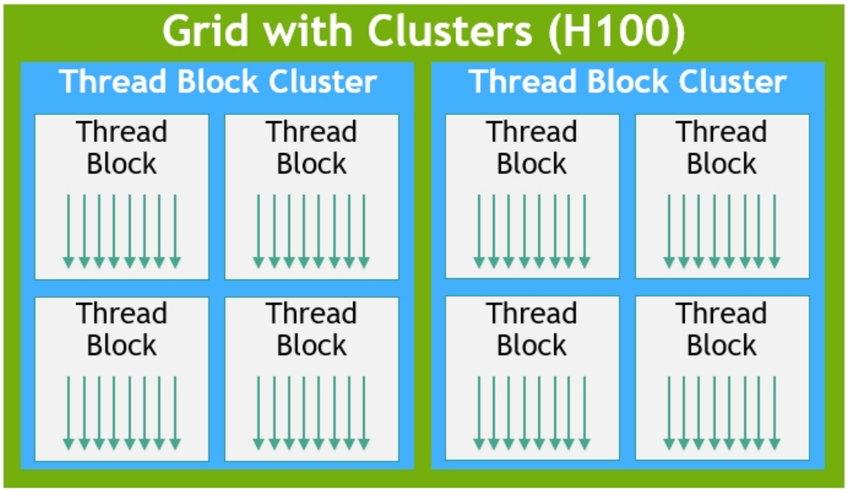

Figure 5: H100 features clusters of Thread Blocks (source: Nvidia)

Distributed Shared Memory (DSMEM)

DSMEM is Nvidia's name for the new capabilities for sharing data between SM shared memories. The CUDA environment ensures a contiguous, virtual memory address space for the participating shared memories.

Sharing data between SMs is thus done by simple pointer references. As mentioned earlier, the data exchanges between SM's are sped up by roughly 7x (*Nvida Hopper Architecture In-Depth — NVIDIA Technical Blog*, n.d.).

Configurable shared memory

Like the A100, which has Compute Capability (CC) 8.0, the H100s CC 9.0 has the same parameters (e.g., warp size of 32, max warps/SM of 64, etc.) as the V100 (CC 7.0), except for the shared memory sizes per SM, which are configurable up to 96KB, 164KB and 228KB, respectively, for the V100, A100 and H100. Keeping data local to the SMs is thus made even easier.

Application speedups over A100

Nvidia highlighted four HPC applications when presenting potential speedups for their new Hopper architecture at GTC 2022: Climate modeling, genomics, lattice quantum chromodynamics, and 8K-point three-dimensional FFT. The FFT in particular was shown to perform about 30x faster over A100 on a multi-GPU H100 system where the GPUs were connected via an NVLink network. It will be interesting to see how our own applications perform on such systems.

Hopper Power Efficiency

The SXM5 form factor H100 is stated to have a TDP (Thermal Design Power, the maximum heat dissipation a hardware component is designed to endure) of 700W. This generated its fair share of discourse on social media with proponents and opponents of the seemingly high TDP. Opponents are discussing, e.g., direct operating costs from the power draw, cooling costs due to increased cooling needs, among other concerns. The TDP is relatively high, yet Nvidia states that the Hopper generation of GPUs is their most energy-efficient yet. How does one defend a TDP of 700W for a GPU when previous generations have had TDPs around 3-400W?

The SXM5 variant uses HBM3 memory modules. Utilizing the full bandwidth of HBM3 likely requires an increase in memory clock over the PCIe Gen 5 variant; the PCIe Gen 5 variant uses the lower bandwidth HBM2e modules. The SXM5 variant is also, in likelihood, going to have a faster boost clock than the PCIe Gen 5 variant. Increased memory and boost clock frequencies are, however, unlikely to be the factors pushing the TDP towards 700W, due to power consumption scaling linearly with frequency under the same voltage conditions.

At the time of writing, it is challenging for us to estimate the mean power draw of the SXM5 variant in a general AI or HPC workload. Benchmarking the performance-to-power ratio of the H100 when power-capped is another benchmark that could be interesting to investigate.

Note also that the Hopper architecture introduces additional capabilities for asynchrony. One of the significant benefits of asynchrony in this case is the potential for attaining high degrees of utilization. The increased asynchrony and potential for concurrency bodes well for latency hiding (Volkov, 2016). The Hopper architecture might look bad on paper with a TDP of 700W, but might look good when taking the performance-per-watt into account.

New AI and Security features

AI techniques, especially those related to Machine Learning (ML) and Deep Learning are increasingly useful for scientific computing as well. The H100 new *Transformer Engine* can accelerate Transformer model training and inference by dynamically choosing between FP8 and 16-bit calculations, which may deliver up to 9x faster AI training and up to 30x faster AI inference speedups on large language models compared to A100.

H100 also provides some features that enable safer multi-user environments, especially important in virtualized environments. The new *Confidential Computing* capability with MIG-level Trusted Execution Environments (TEE) supports up to seven individual GPU Instances, each with dedicated NVDEC and NVJPG units. Each Instance now includes its own set of performance monitors that work with Nvidia developer tools. The H100 extends the *Trusted Execution Environment* with CPUs at full PCIe line rate.

ARM and Grace

Announced for 2023, Nvidia's ARM-based Grace CPU seems also noteworthy. USA's Los Alamos National Laboratory and the Swiss National Computing Centre have already announced plans for Grace-based supercomputers (*Nvidia's Supercomputing CPU Puts Intel Under Pressure*, n.d.). ARM, originally a UK IP-only company, is now owned by the Japanese SoftBank Group Corp (SBC). Nvidia also tried buying ARM from them in a Fall 2020 agreement, but the deal was terminated in Feb. 2022 due to "regulatory challenges" in both US and China.

Intel still dominates the top500.org list of the world's largest supercomputers, but in 2021 Japan's ARM-based Fugaku supercomputer (> seven million cores, running at 442 petaFLOPS) took the top spot.

The **Grace CPU Superchip** https://www.nvidia.com/en-us/data-center/grace-cpu/(*Nvidia Grace CPU intro*, n.d.) features two Grace cores connected via the NVLink-C2C technology thus providing **up to 144 Arm v9 CPU cores**. It claims to be the world's first CPU using LPDDR5x memory with ECC and 1TB/s total bandwidth. Its 900 GB/s coherent interface is 7x faster than PCIe Gen 5.

Nvidia's **Grace Hopper Superchip** combines the **Grace CPU** and **Hopper GPU** architectures using Nvidia's NVLink-C2C technology to deliver a coherent CPU+GPU memory model.

The system targets both HPC and AI applications and can provide a 30x higher aggregate system memory bandwidth to GPU compared to the DGX A100. Both the Grace and Grace Hopper superchips will run Nvidia's software stacks, including NVIDIA HPC, NVIDIA AI, and NVIDIA Omniverse .

Dr. Grace Murray Hopper

Finally, it is nice to see computer chips named after a woman. Computer pioneer and Rear Admiral of the US Navy Dr. Grace Brewster Murray Hopper (1906-1992) (*Captain Grace M. Hoppper Navy file, July 1981*, n.d.) developed, among several other things, the first compiler called A-0, and also coined the computer term "bug" and "de-bug" after discovering an actual bug in a computer. She got her Masters and PhD degrees from Yale University where she was the first woman to earn a PhD in Mathematics. The Cray XE6 "Hopper" supercomputer at NERSC (National Energy Research Scientific Computing Center) is named after her (*Five Fast Facts About Technologist Grace Hopper*, n.d.) as is Hopper Hall, the new Center for Cyber Security Studies at the US Naval Academy (*USNA Hopper Hall Announcement*, n.d.).

About the authors: We have not had the chance to test any H100 systems ourselves, but our recent BAT - Benchmark suite for Autotuners (Sund et al., 2021 url: = https://hgpu.org/?p=25866) was benchmarked on both our 2-node IBM AC922 Power systems with NVLinked V100 GPUs and CPUs and our NVIDIA DGX2 system, with 16 tightly coupled NVSwitched GPUs, so we are looking forward to running it on a H100 system.

Anne C. Elster (IEEE Senior Member) is a Professor in HPC at the Norwegian University of Science and Technology and also the Associate Editor of the IEEE CiSE New Architecture department. She holds

This article has been accepted for publication in Computing in Science & Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MCSE.2022.3163817

a PhD in EE from Cornell University and was one of the original members of the MPI committee. She has worked on GPUs with her students since 2007.

Tor A. Haugdahl (IEEE Student Member) is currently finishing his Masters degree in Computer Science specializing in autotuning for HPC at the Norwegian University of Science and Technology (NTNU). As a member at the HPC-lab at NTNU, he works alongside other Masters students specializing in a range of HPC-related topics.

References

https://developer.nvidia.com/blog/nvidia-hopper-architecture-in-depth/. https://developer.nvidia.com/blog/nvidia-hopper-architecture-in-depth/

https://www.nvidia.com/en-us/technologies/hopper-architecture/. https://www.nvidia.com/en-us/technologies/hopper-architecture/

https://nvidianews.nvidia.com/news/nvidia-to-acquire-mellanox-for-6-9-billion

https://blogs.nvidia.com/blog/2022/03/22/nvidia-hopper-accelerates-dynamic-programming-using-dpx-instructions/. https://blogs.nvidia.com/blog/2022/03/22/nvidia-hopper-accelerates-dynamic-programming-using-dpx-instructions/

Understanding Latency Hiding on GPUs (Number UCB/EECS-2016-143). (2016). (Number) [PhD thesis, EECS Department, University of California, Berkeley]. http://www2.eecs.berkeley.edu/Pubs/TechRpts/2016/EECS-2016-143.html

https://spectrum.ieee.org/nvidia-supercomputing-cpu-puts-intel-under-pressure. https://spectrum.ieee.org/nvidia-supercomputing-cpu-puts-intel-under-pressure

https://www.nvidia.com/en-us/data-center/grace-cpu/. https://www.nvidia.com/en-us/data-center/grace-cpu/

https://www.history.navy.mil/content/dam/nhhc/research/histories/bios/HopperGrace/Hopper.pdf. https://www.history.navy.mil/content/dam/nhhc/research/histories/bios/HopperGrace/Hopper.pdf

https://www.energy.gov/articles/five-fast-facts-about-technologist-grace-hopper. https://www.energy.gov/articles/five-fast-facts-about-technologist-grace-hopper

https://www.usna.edu/NewsCenter/2020/10/NAVAL_ACADEMY_OPENS_NEW_CENTER_FOR_CYBER_SECURITY_STUDIES,_HOPPER_HALL.php. https://www.usna.edu/NewsCenter/2020/10/NAVAL_ACADEMY_OPENS_NEW_CENTER_FOR_CYBER_SECURITY_STUDIES,_HOPPER_HALL.php

BAT: A Benchmark suite for AutoTuners. (2021 url: = https://hgpu.org/?p=258662021 url: = https://hgpu.org/?p=25866). *NIK 2021 - Norwegian National ICT Conference.*