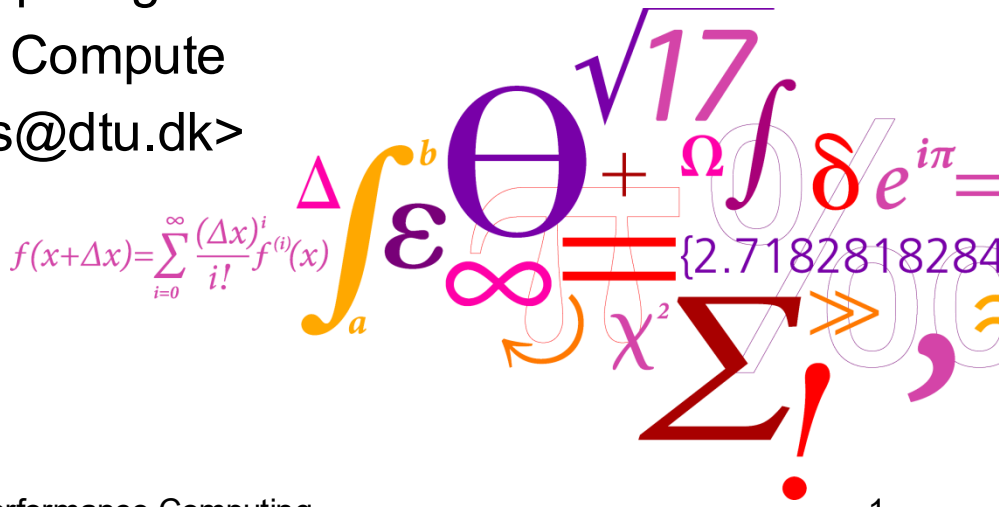
Introduction to GPU Computing


Hans Henrik Brandenborg Sørensen
DTU Computing Center
DTU Compute
<hhbs@dtu.dk>

DTU DCC Computing Center

GPUlab DTU Informatics



Practicalities

- Teacher(s) - Week 3
  - Hans Henrik B. Sørensen
  - Rasmus Kleist
- Lectures / GPU Exercises – Mon-Tue
  - Learning by doing!
  - Complementary to Assignment 3 (should not be handed in)
- Assignment #3: GPU Matrix Multiplication and GPU Poisson Problem – Wed-Fri
  - Collaborate on developing the code + report
  - Fill out “responsibilities” in the addendum like week 1+2

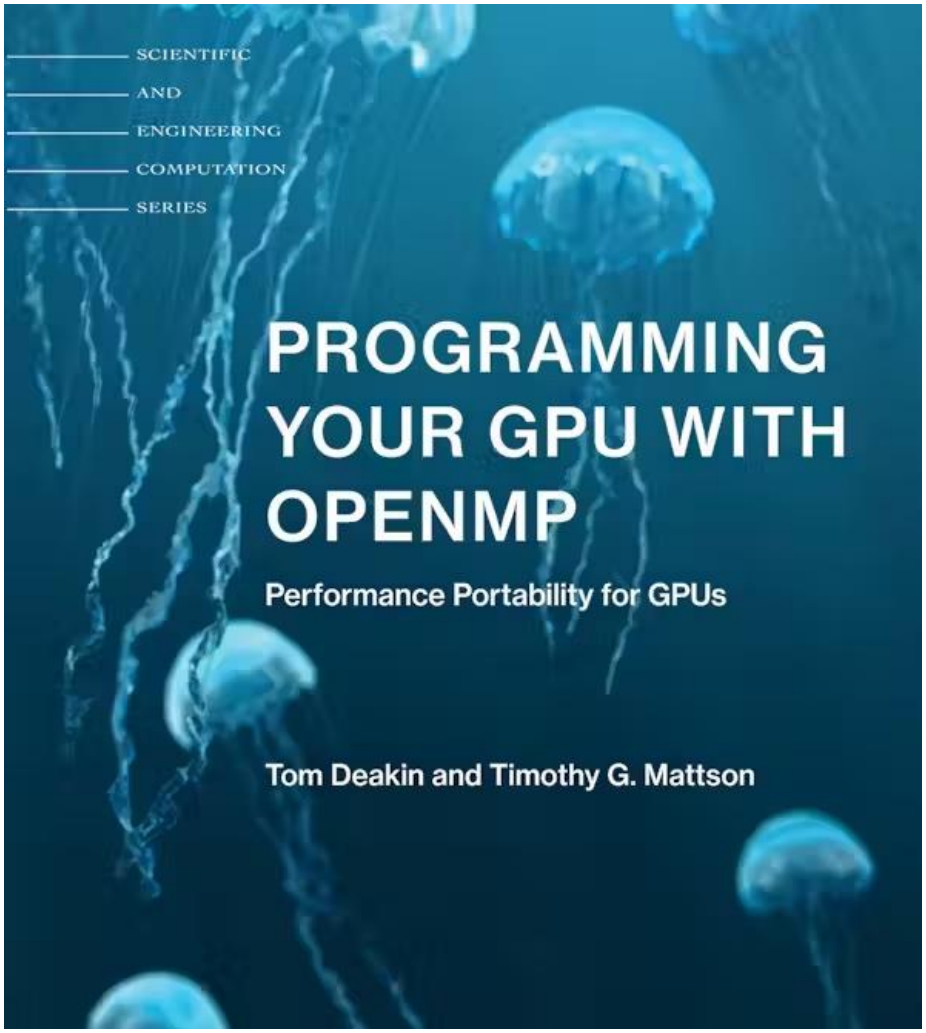
Learning objectives for week 3

- This week we focus on learning
  - Explain high-performance GPUs
  - Interplay of computer components like CPU, GPU, caches, and memory
  - Write parallel programs with OpenMP / Offloading
  - Write efficient programs for multi-core processors and many-core GPUs / Multi-GPUs
- Pre-requisites:
  - Profiency in C/C++
  - Access to a so-called CUDA enabled GPU (e.g. DCC or home-pc with NVIDIA CC.>=8.0).

Overview for week 3

- GPU computing – Mon
  - Introduction to GPU computing
  - OpenMP offload basics
  - OpenMP offload data mapping
- Performance tuning of GPU applications – Tue
  - Introduction to GPU tuning techniques
  - OpenMP offload memory accessing
  - OpenMP offload profiling tools / advanced topics
- Assignment 3 – Wed
  - Q&A

OpenMP book for week 3

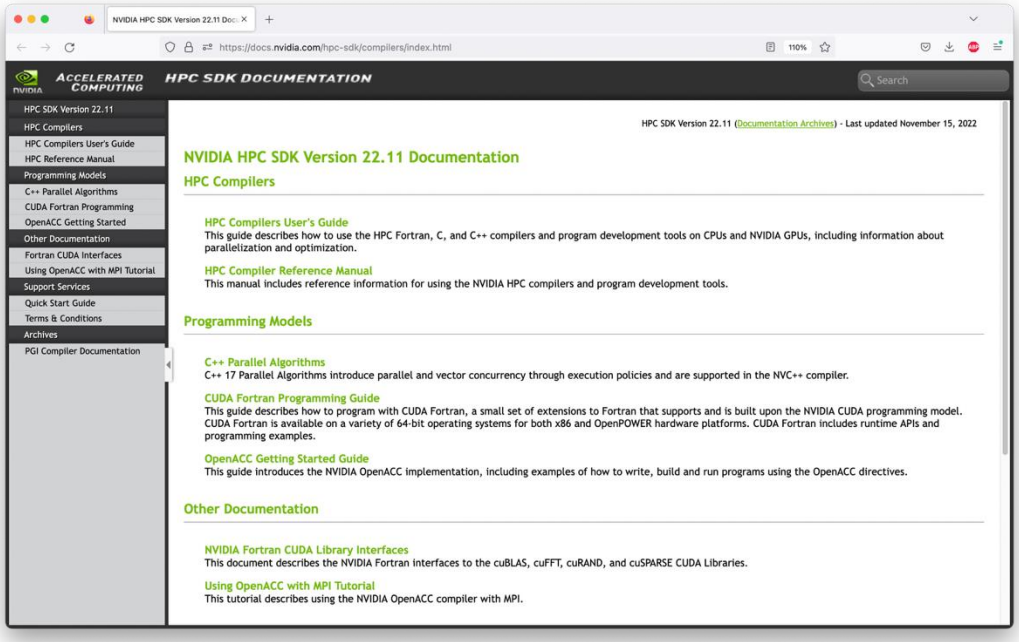


- We do not expect you to buy this or other books!
- Use slides and online documentation (see next slide)!
- All topics in week 3 are nicely covered in this MIT OpenMP series book (released 2023)!

Format	Bog, paperback • Engelsk • 336 sider
Indgår i serie	Scientific and Engineering Computation

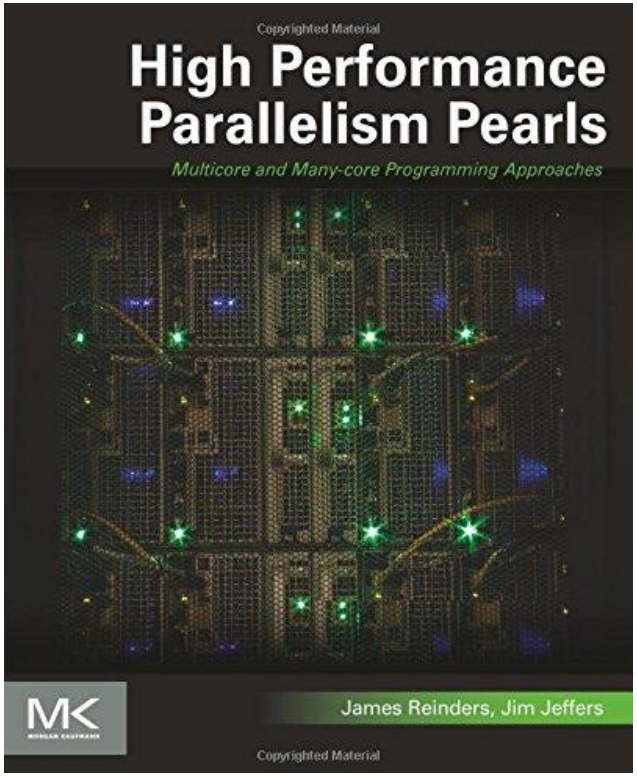
Normalpris: **kr. 1.154,95**

Our suggestion (if you get hooked)



https://docs.nvidia.com/hpc-sdk/compilers/index.html

+



Denne udgave Bog, paperback • Engelsk

Normalpris: **kr. 489,95**

https://www.openmp.org/spec-html/5.1/openmp.html#openmpch2.html

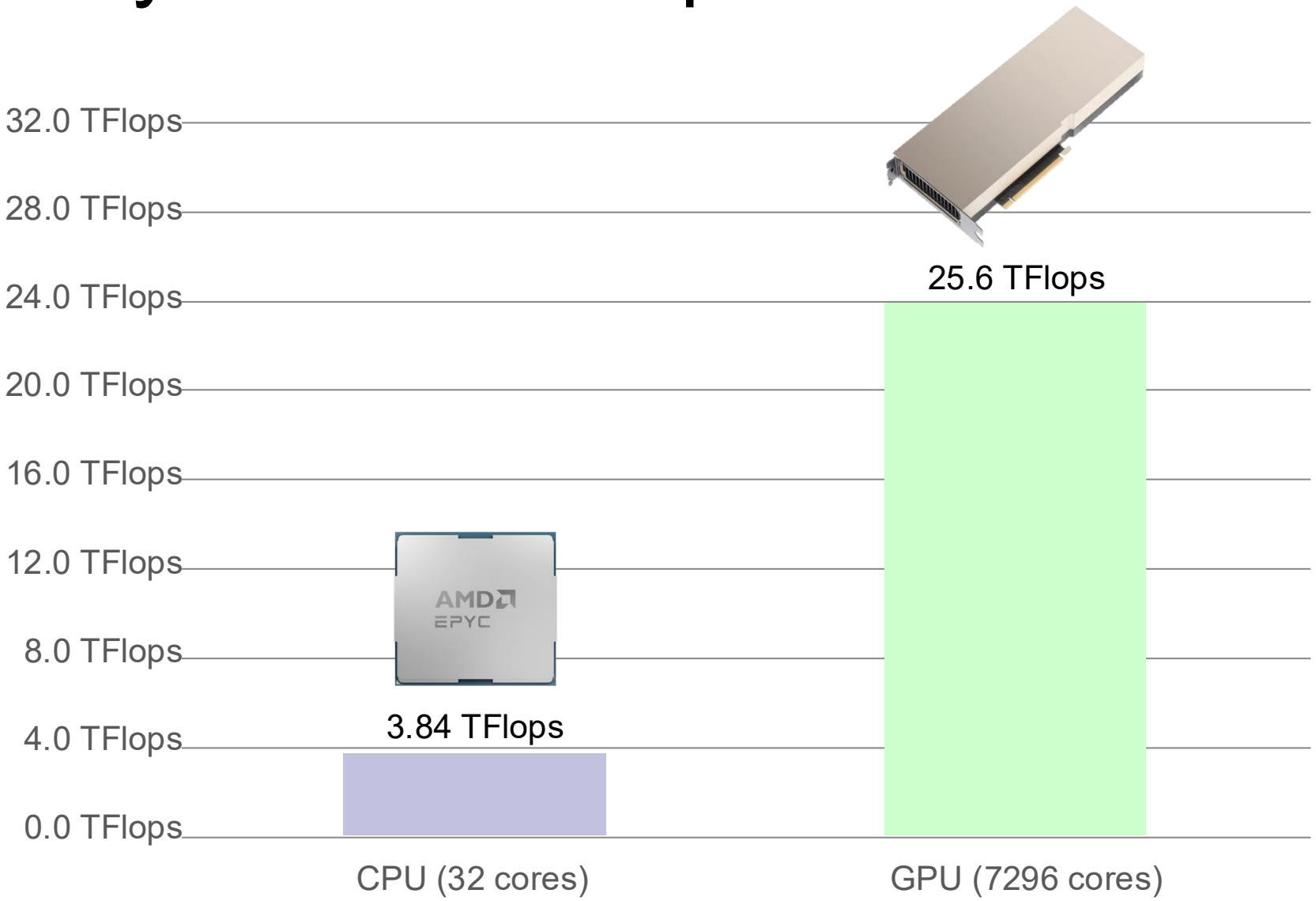
https://stackoverflow.com/questions/tagged/openmp+gpu

Introduction to GPU computing

Outline

- Why are GPUs important in HPC?
- Why are GPUs different from CPUs?
- GPUs as accelerators
- GPU hardware for 02614 course
- Motivation slides

Why are GPUs important in HPC?



Why are GPUs important in HPC?

2039 GB/s

460.8 GB/s

2000 GB/s
1750 GB/s
1500 GB/s
1250 GB/s
1000 GB/s
750 GB/s
500 GB/s
250 GB/s
0 GB/s

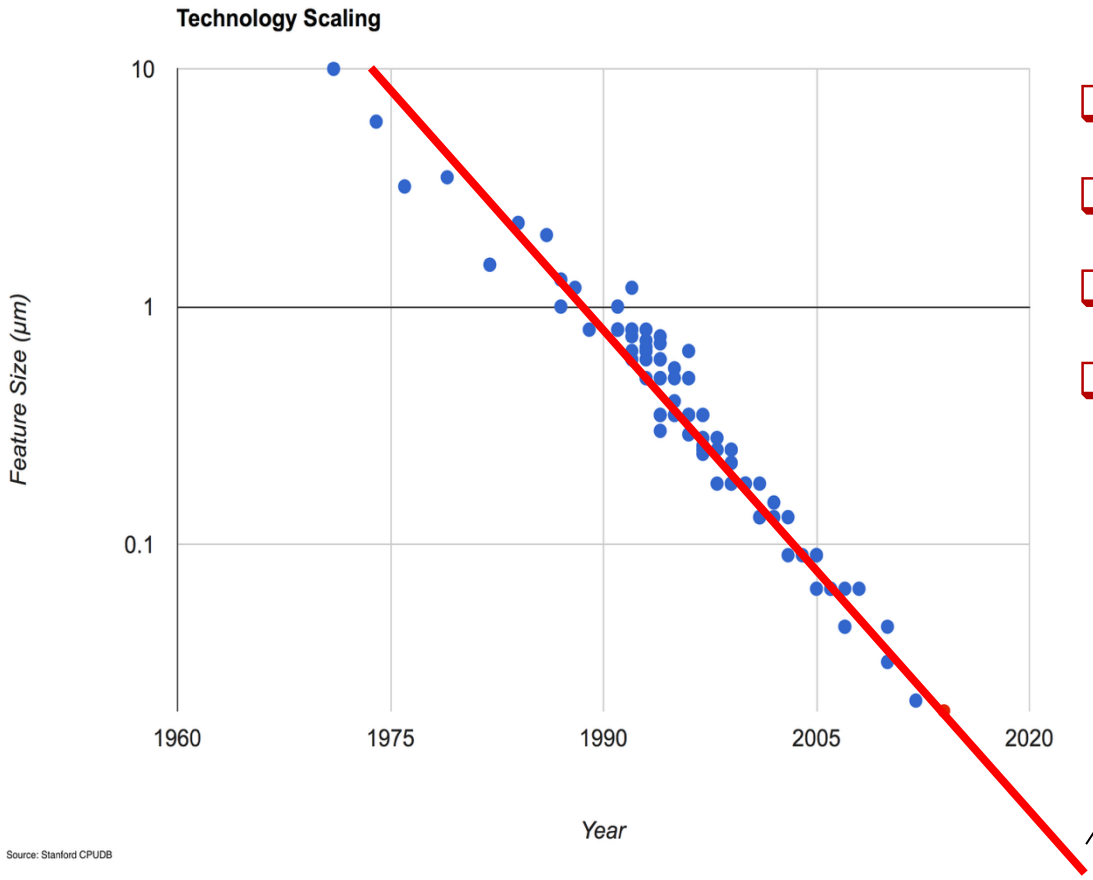
CPU (32 cores)

GPU (7296 cores)

Why are GPUs different from CPUs?

Recap from week 1: Good news

- Transistor size over the years; still decreases

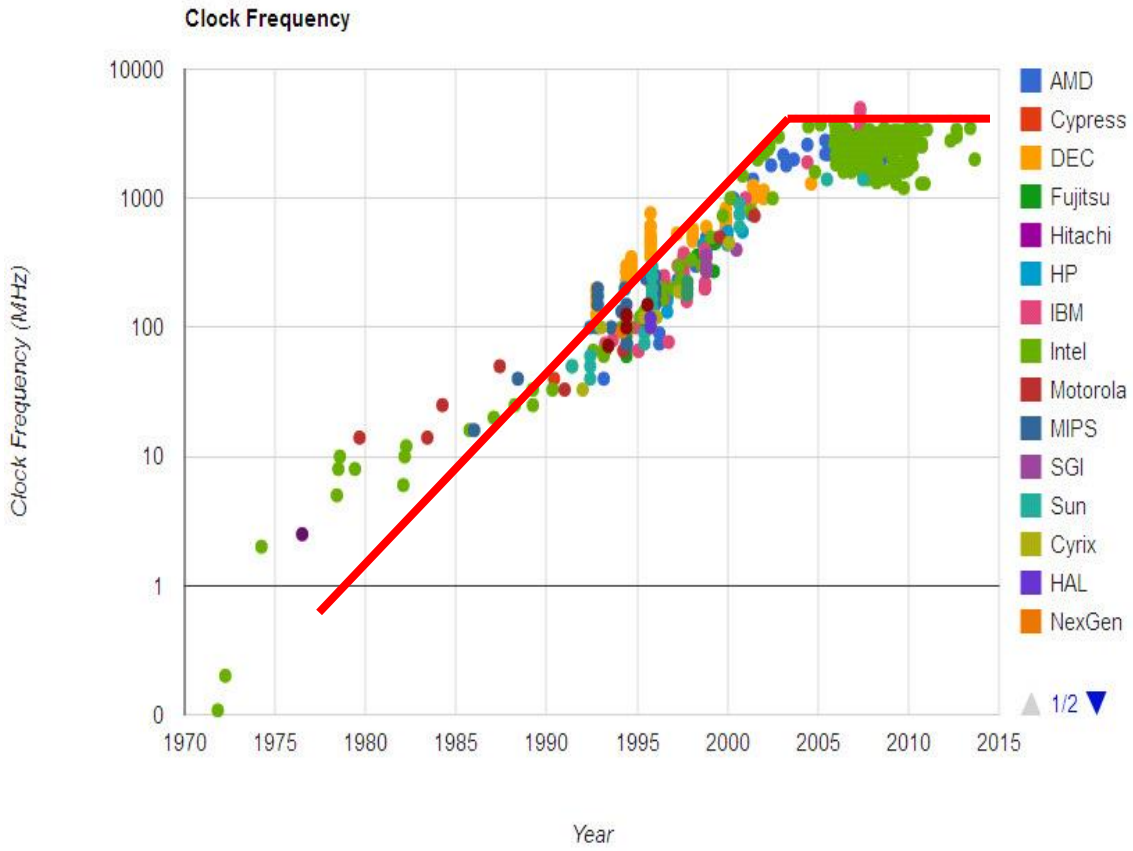


- Smaller
- Faster
- Less power
- More on chip

Currently best: 3 nm

Recap from week 1: Bad news

- Clock speed over the years; stagnated since 2005



- Increasing over many years
- However, over the last decade the clocks speeds have essentially remained constant

Why not increase clock speed?

- Power = Frequency x Voltage^2
- Heat produced depends on power

Why not increase clock speed?

- Power = Frequency x Voltage^2
- Heat produced depends on power
- “Nard scaling”
  - Reduce transistor voltage to counter higher clock speed
  - Broke down in 2005 because of weakened current in the wires (need to distinguish 0 and 1)

Why not increase clock speed?

- Power = Frequency x Voltage^2
- Heat produced depends on power
- “Nard scaling”
  - Reduce transistor voltage to counter higher clock speed
  - Broke down in 2005 because of weakened current in the wires (need to distinguish 0 and 1)
- What matters today: # operations per Watt!
- Trade-off favors slower simpler processors
  - More operations per watt
  - Frequency kept low

Building a modern processor

- What is the goal?

Building a modern processor

- What is the goal? Roughly two choices...

A.	B.
Latency (time to complete a task) [e.g. seconds]	Throughput (tasks completed per unit time) [e.g. jobs/hour]

Building a modern processor

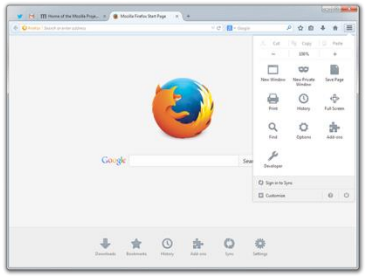
- What is the goal? Roughly two choices...

A.	B.
Latency (time to complete a task) [e.g. seconds]	Throughput (tasks completed per unit time) [e.g. jobs/hour]

- Unfortunately these goals are not always aligned

CPUs target latency (traditionally)

- Usual task of traditional CPUs
  - Desktop applications / OS
    - Lightly threaded
    - Lots of branches
    - Lots of (indirect) memory accesses



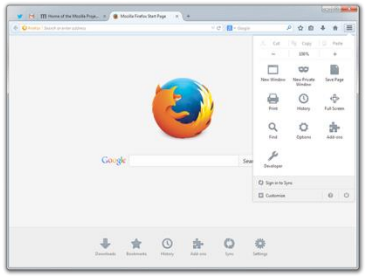
Mac OS

Windows 10

Microsoft Office

- CPUs try to minimize the time to complete a particular task – often to support user interaction!

CPUs target latency (traditionally)



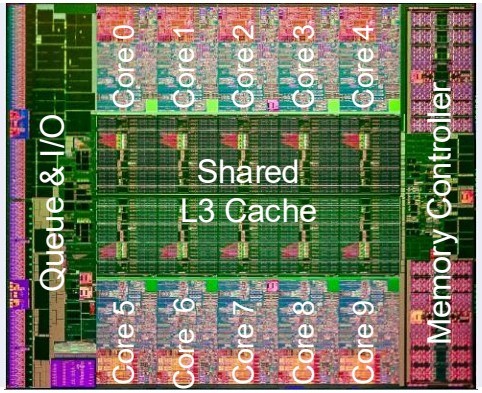
- Usual task of traditional CPUs
  - Desktop applications / OS
    - Lightly threaded
    - Lots of branches
    - Lots of (indirect) memory accesses

Mac OS

Windows 10

Microsoft Office

- CPUs try to minimize the time to complete a particular task – often to support user interaction!
- Complex control hardware
  - \+ Flexibility in performance
  - \+ Lightly parallel
  - − Expensive in terms of power



GPUs target throughput

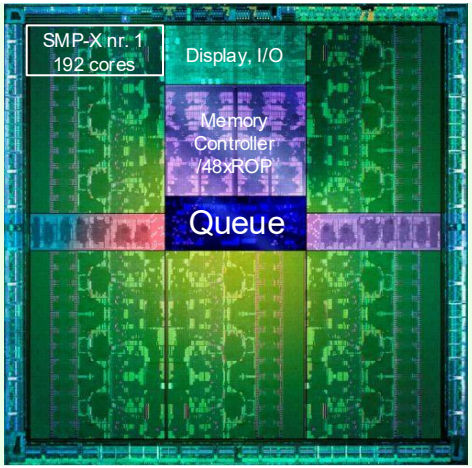
- GPUs are designed to compute pixels – fast!
  - Rendering video games in real-time
  - Play HD movies on smart phones
  - Render visual effects for movies…

- More concerned about the number of pixels per second than the latency of any particular pixel!

GPUs target throughput

- GPUs are designed to compute pixels – fast!
  - Rendering video games in real-time
  - Play HD movies on smart phones
  - Render visual effects for movies…

- More concerned about the number of pixels per second than the latency of any particular pixel!
- Simpler control hardware
  - + More transistors for computation
  - + Power efficient
  - − Highly parallel
  - − More restrictive in performance



Hardware hierarchy

- CPU (typical)
  - Processing Unit (L3 cache)
  - Core (L2 cache / L1 cache / Instr. cache)

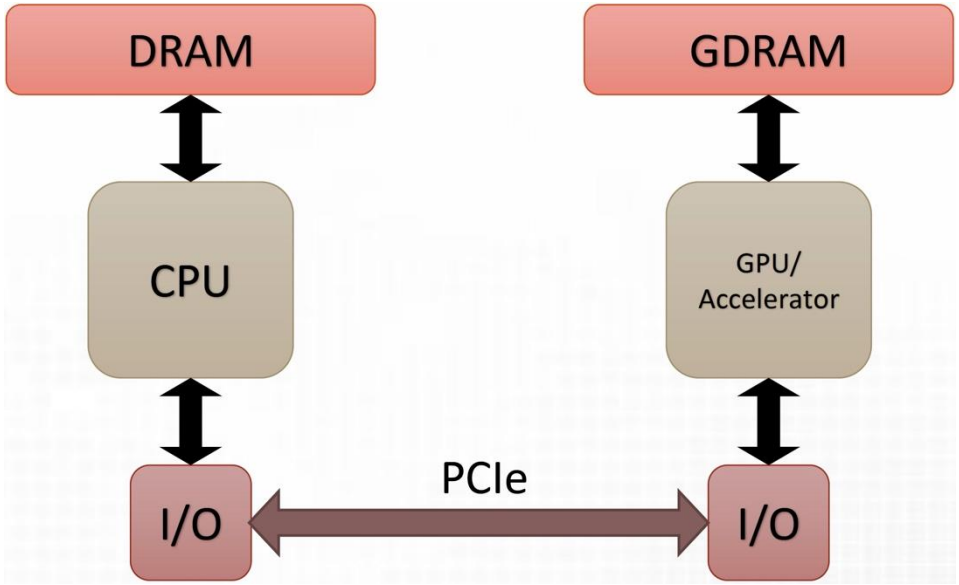
Hardware hierarchy

- CPU (typical)
  - Processing Unit (L3 cache)
  - Core (L2 cache / L1 cache / Instr. cache)
- GPU (Nvidia)
  - Processing Unit (L2 cache)
  - GPC – Graphics Processing Cluster
  - TPC – Texture Processing Cluster
  - SM – Streaming Multiprocessor (64 “cores” / L1 cache)
  - Processing block (Instr. cache)
  - “Core”

GPUs as accelerators

GPUs as accelerators

- Problem: Still require OS, I/O, and scheduling
- Solution: “Hybrid system”
  - CPU provides management
  - Accelerators such as GPUs provide compute power

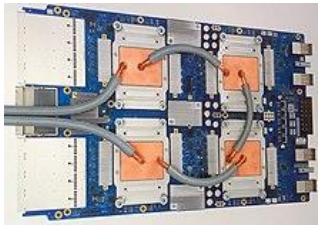


Types of accelerators

- GPUs
  - HPC high-end versions – Tesla branch
  - DP downgraded versions – Titan branch
  - Gamer versions – RTX branch

- Custom many-core processors
  - Japan / China

- TPUs (Tensor Processing Unit)
  - AI accelerator – application-specific integrated circuit (ASIC) developed by Google



Accelerators in Top500

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**El Capitan** - HPE Cray EX255a, AMD 4th Gen EPYC 24C 1.8GHz, AMD Instinct MI300A, Slingshot-11, TOSS, HPE DOE/NNSA/LLNL United States	11,340,000	1,809.00	2,821.10	29,685
2	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE Cray OS, HPE DOE/SC/Oak Ridge National Laboratory United States	9,066,176	1,353.00	2,055.72	24,607
3	**Aurora** - HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11, Intel DOE/SC/Argonne National Laboratory United States	9,264,128	1,012.00	1,980.01	38,698
4	**JUPITER Booster** - BullSequana XH3000, GH Superchip 72C 3GHz, NVIDIA GH200 Superchip, Quad-Rail NVIDIA InfiniBand NDR200, RedHat Enterprise Linux, EVIDEN EuroHPC/FZJ Germany	4,801,344	1,000.00	1,226.28	15,794
5	**Eagle** - Microsoft NDv5, Xeon Platinum 8480C 48C 2GHz, NVIDIA H100, NVIDIA Infiniband NDR, Microsoft Azure Microsoft Azure United States	2,073,600	561.20	846.84	

AMD Instinct (MI300A)

AMD Instinct (MI250X)

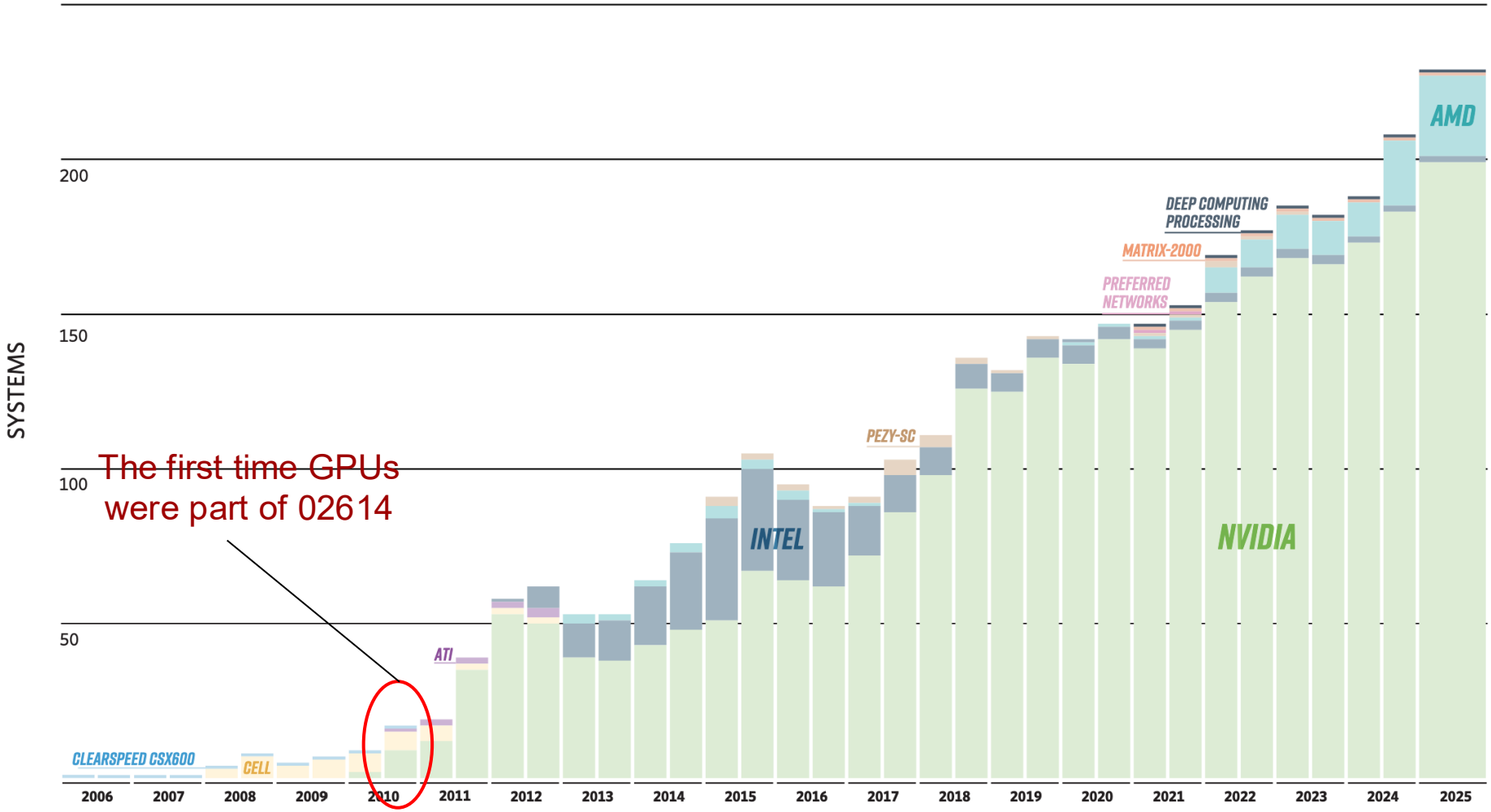
Intel Data Center GPU (Max)

Nvidia Tesla (Grace-Hopper)

Nvidia Tesla (Hopper)

Accelerators in Top500

ACCELERATORS/CO-PROCESSORS



Nvidia GPU architectures

- Five generations of high-performance GPUs:

	# GPUs	Name	Year	Architecture	CUDA cap.	CUDA cores	Clock MHz	Mem GiB	SP peak GFlops	DP peak GFlops	Peak GB/s
Kepler	5	Tesla K40c	2013	GK110B (Kepler)	3.5	2880	745 / 875	11.17	4291 / 5040	1430 / 1680	288
	8	Tesla K80c (dual)	2014	GK210 (Kepler)	3.7	2496	562 / 875	11.17	2796 / 4368	932 / 1456	240
Pascal	8	*TITAN X	2016	GP102 (Pascal)	6.1	3584	1417 / 1531	11.90	10157 / 10974	317.4 / 342.9	480
Volta	22	Tesla V100	2017	GV100 (Volta)	7.0	5120	1380	15.75	14131	7065	898
	12	Tesla V100-SXM2	2018	GV100 (Volta)	7.0	5120	1530	31.72	15667	7833	898
Ampere	6	Tesla A100-PCIE	2020	GA100 (Ampere)	8.0	6912	1410	39.59	19492	9746	1555
Hopper	-	Tesla H100-SXM5	2022	GH100 (Hopper)	9.0	8448	1650	79.18	38984	19492	3110

Source: http://www.hpc.dtu.dk/?page_id=2129

#Cores, Mem, and GB/s keep going up!

Clock freq. level off!

Peak increases 2x every ~3 years!

Trends in HPC due to GPUs

- Improvements at individual computer node level are greatest
  - Less data transfer
  - Heterogeneous computing
  - Avoiding MPI

Trends in HPC due to GPUs

- Improvements at individual computer node level are greatest
  - Less data transfer
  - Heterogeneous computing
  - Avoiding MPI
- "Super-nodes": Nvidia DGX-1
  - Eight tightly linked high-end GPUs
  - 40,960 cores / 960 TFlops in 1 node

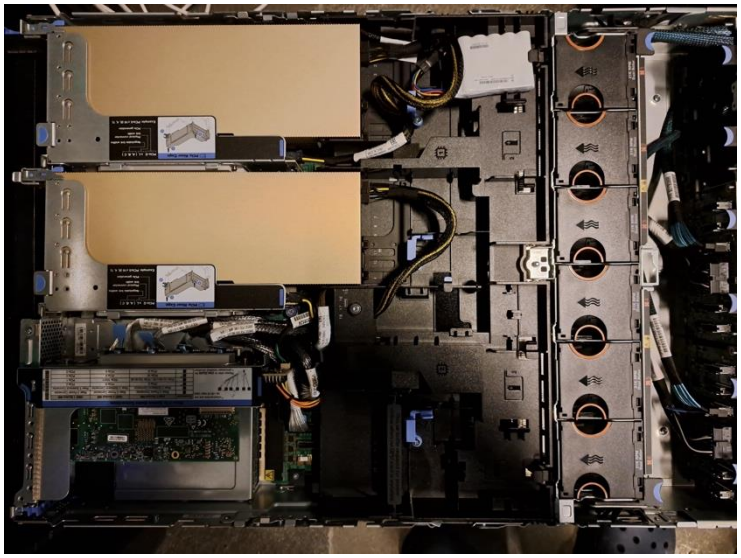
Trends in HPC due to GPUs

- Improvements at individual computer node level are greatest
  - Less data transfer
  - Heterogeneous computing
  - Avoiding MPI
- "Super-nodes": Nvidia DGX-1
  - Eight tightly linked high-end GPUs
  - 40,960 cores / 960 TFlops in 1 node
- Communication costs are increasing
  - Synchronization-reducing algorithms
  - Communication lower-bound algorithms

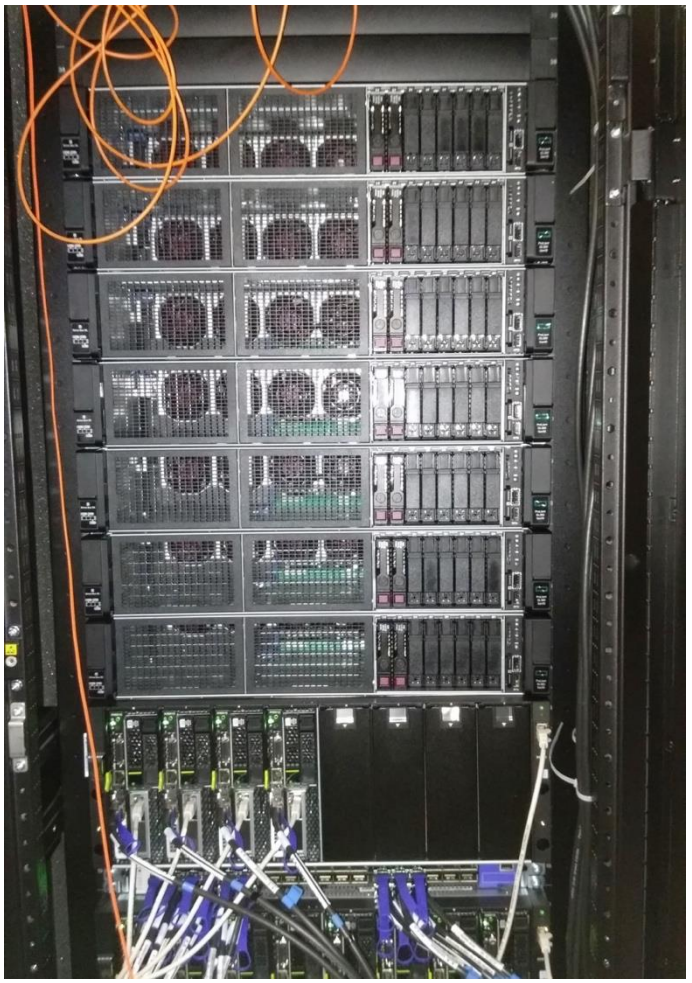
GPU hardware for this course

Hopper nodes for 02614

Delivered: November 23, 2023



1 HPE node:
2 x AMD EPYC 9354 CPU @ 3.25GHz (32 cores)
2 x Tesla H100-PCIE-80GB (7296 cores)
768 GB DDR5 @ 4800MHz

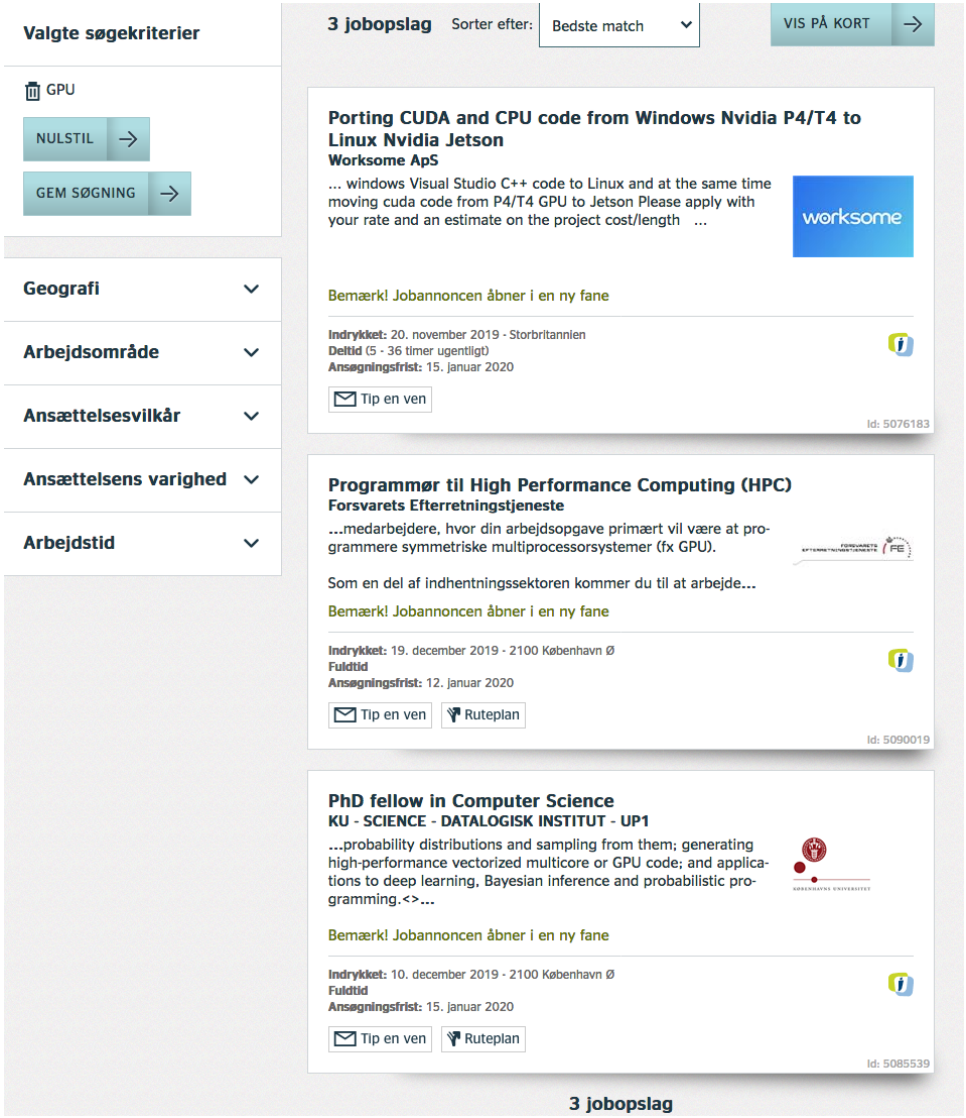


Hopper nodes for 02614

H100-PCIE-80GB

- 114 streaming multiprocessors (SM) – 80 GB memory
- Multi-user system: GPUs are in default mode (shared)
- Batch system: Script sets GPUs to exclusive mode

Motivation Jan 2020



Lukas Christian Høghøj, *Large-scale modelling on GPUs with OpenACC,* 2019

Patrick Møller Jensen and Julian Thomas Reckeweg Olsen, *GPU beamforming*, 2019

Konstantinos Gkanos, *Interactive, real-time room acoustic simulations*, 2019

Gandalf Saxe and Oisin D. Kiær, *Low Energy Transfer Orbits to Mars using Evolution Strategies*, 2019

Mia Sandra Nicole Siemon, *Comparison of GPU programming models*, 2019

Mathias Sorgenfri Lorenz, *Scaling analysis of a multi-GPU Poisson solver*, 2018

Tim Felle Olsen and Mathias Sorgenfri Lorenz, *Large-scale computations on modern GPUs*, 2018

Nick Clausen, *Leveraging the GPU architecture of embedded systems for model predictive control problems*, 2018

…

Motivation Jan 2024

PhD scholarship in Cardiac Vector Flow Ultrasound Imaging

Danmarks Tekniske Universitet, Kongens Lyngby

A 3-year PhD scholarship is available at the Center for Fast Ultrasound Imaging (CFU), Department of Health Technology (DTU Health Tech), DTU and at the Department of Cardiology, Rigshosptalet from March 2022. It is conducted in collaboration between the two departments, and the overall purpose of the project is to further develop fast vector flow-based imaging of the heart and make it applicable for clinical use

"Experience with GPU programming is desirable."

Tjek jobglæden:

★★★★★ 391 evalueringer INDRYKKET: 17-12-2021

International Workshop on OpenMP

↳ IWOMP 2022: **OpenMP in a Modern World: From Multi-device Support to Meta Progr**

Feasibility Studies in Multi-GPU Target Offloading

Anton Rydahl, Mathias Gammelmark & Sven Karlsson

Conference paper | First Online: 20 September 2022

178 Accesses

Part of the Lecture Notes in Computer Science book series (LNCS,volume 13527)

https://link.springer.com/chapter/10.1007/978-3-031-15922-0_6

Abstract

Many of the largest supercomputers are based on heterogeneous architectures with multiple general-purpose graphics processing units (GPGPUs) per compute node. While many APIs

Software Developer – with flair for mathematics and performance

Hillerød Se rejsetid FOSS

You will be a part of the team implementing non-trivial algorithms as software components to be used across the organization.

It would be a plus if you have some of the following qualifications:

- Experience with CUDA or similar GPU-based technologies.

Tjek jobglæden:

★★★★★ 105 evalueringer INDRYKKET: 22-12-2023

Denmark's new AI supercomputer Gefion ranked as 7th fastest storage systems in the world

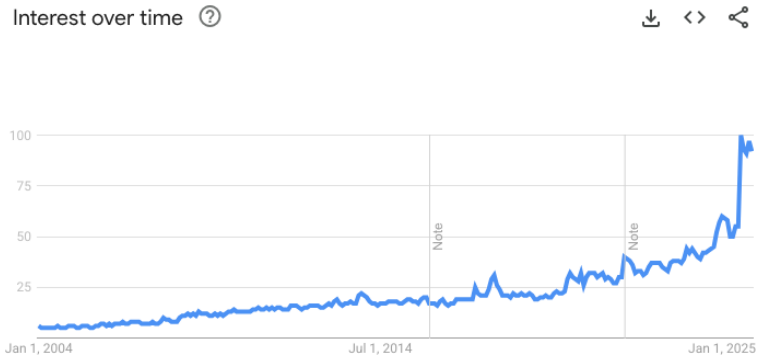
Motivation Jan 2026

Jobindex
5 jobannoncer

DTU

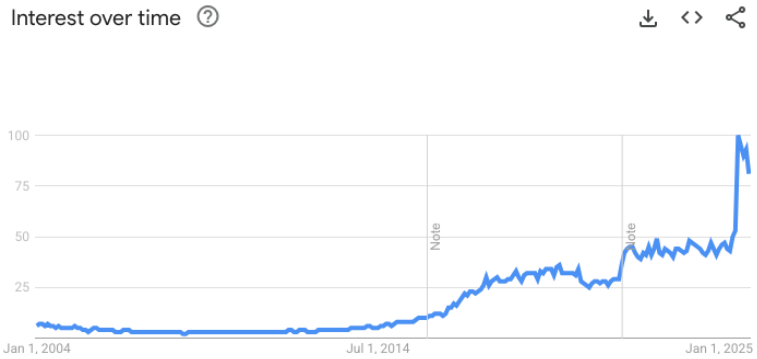
Google Trends

GPU



CPU

Machine learning



End of lecture