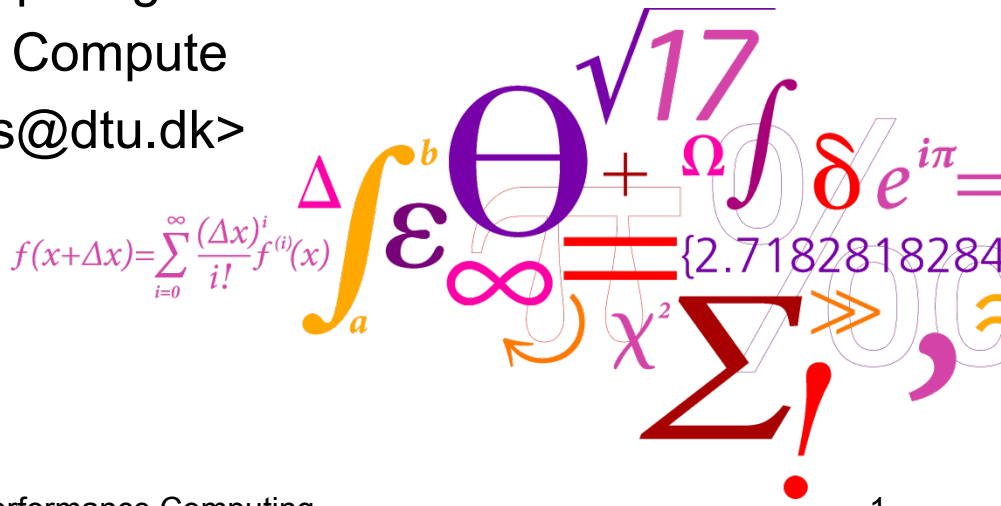
Introduction to GPU Computing

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



Practicalities

- Teacher(s) - Week 3
  - Hans Henrik B. Sørensen
  - Morten, Anders, Asbjørn & Marcus
- 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
  - Interplay of computer components like CPU, GPU, caches, and memory
  - Choose the optimal hardware platform for a given problem
  - Write parallel programs with OpenMP and CUDA
  - Write efficient programs for multi-core processors and many-core GPUs / Multi-GPUs
- Pre-requisites:
  - Profiency in C/C++
  - Access to a 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
  - CUDA programming model
  - CUDA memory model
- Performance tuning of CUDA applications – Tue
  - Basic tuning techniques
  - Memory access optimization
  - Advanced topics
  - Profiling tools / demo
- Assignment 3 – Wed

What is CUDA?

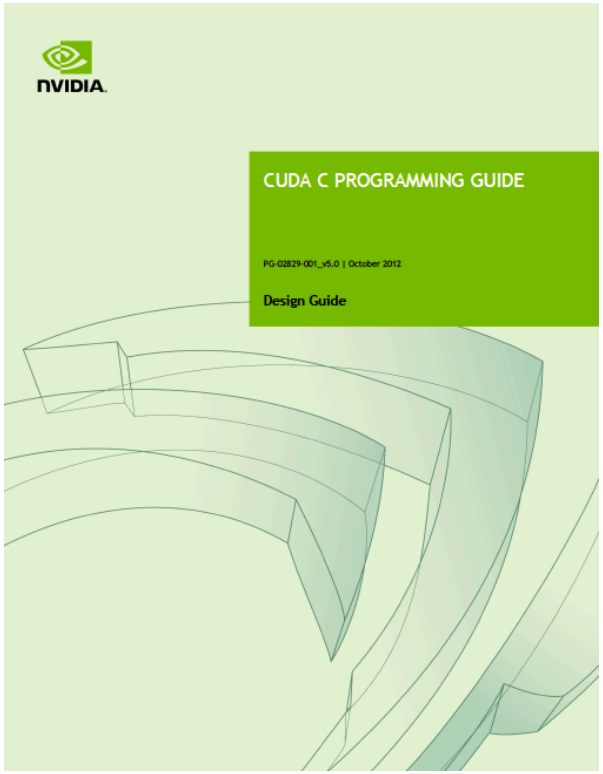
DTU

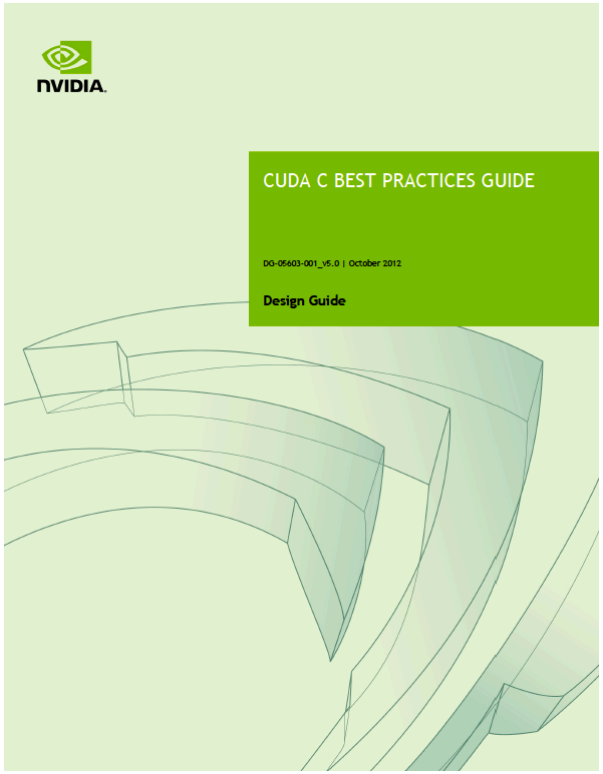
- [**C**ompute **U**nified **D**evice **A**rchitecture]
- A parallel computing standard and API proposed by Nvidia for general-purpose computations on CUDA-enabled GPUs
  - Priority #1: Make things easy (Sell GPUs)
  - Priority #2: Get performance
- Result: Simple to get started, but..
  - Requires expert knowledge to get best performance
- Scalable
- Well documented and free to use (!)

CUDA toolkit

- The CUDA toolkit comprises a full framework of compiler, profiler, low- and high-level APIs, and highly tuned GPU libraries, e.g.
  - cuBLAS – CUDA Basic Linear Algebra Subroutines library
  - cuFFT – CUDA Fast Fourier Transform library
  - cuRAND – CUDA Random Number Generation library
  - cuSOLVER – CUDA based collection of dense and sparse direct solvers
  - cuSPARSE – CUDA Sparse Matrix library
  - …
- Universities Teaching CUDA
  - More than 738 [in Denmark DTU, AU and KU].

Free CUDA material

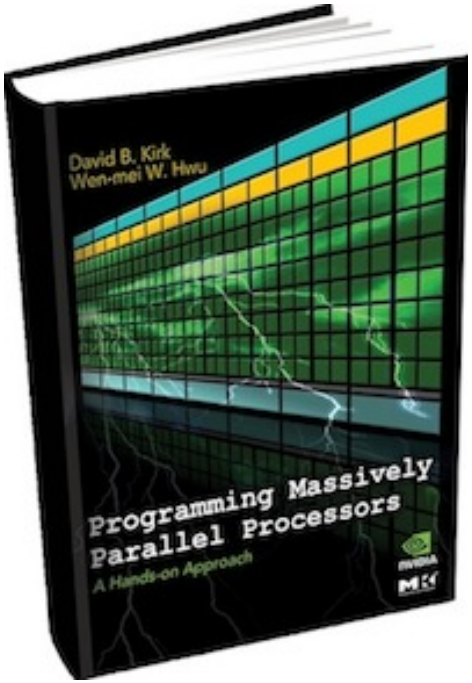


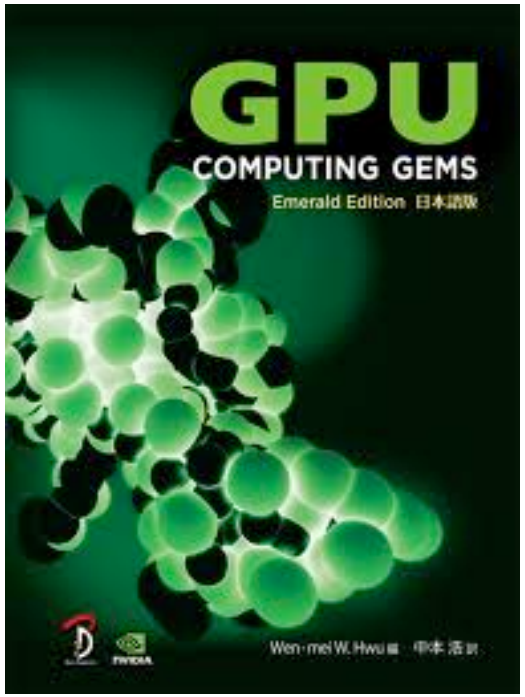


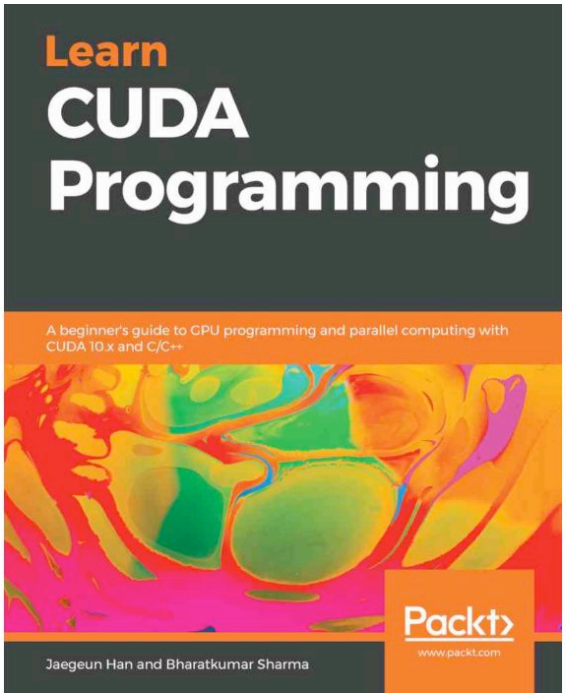
- Free online from Nvidia developer webpage
  - http://docs.nvidia.com/cuda/index.html
  - Pdf versions, see installation; /appl/cuda/11.1/doc/pdf/

Additional CUDA material

CUDA 10.0

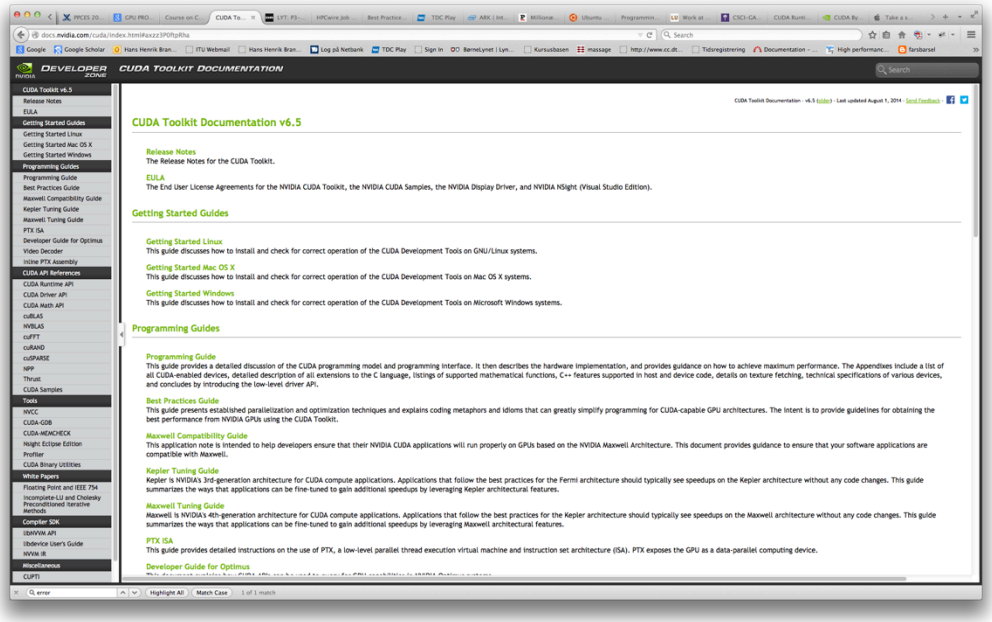




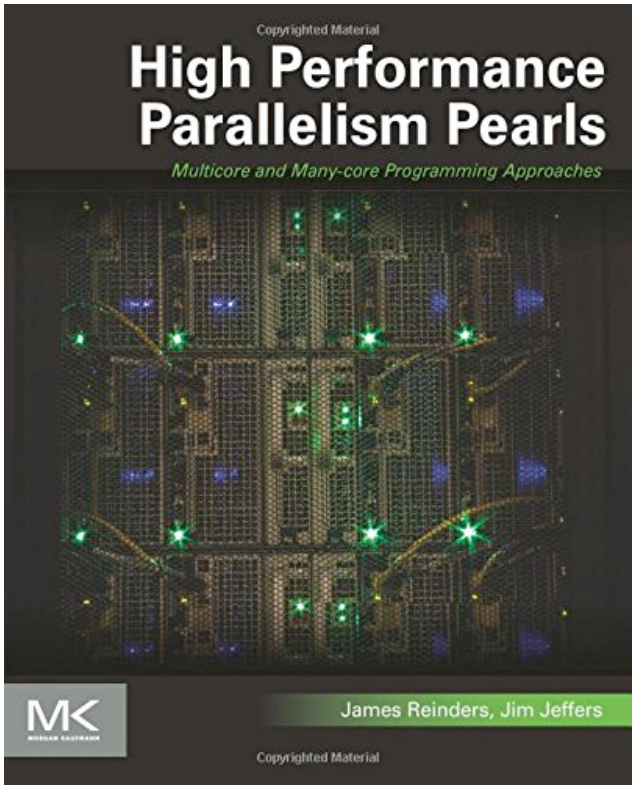


- These are currently the most widely used books
- ~~Feel free to come by my office building 324, room 280 to take a peek in these references~~

Our suggestion (if you get hooked)



+



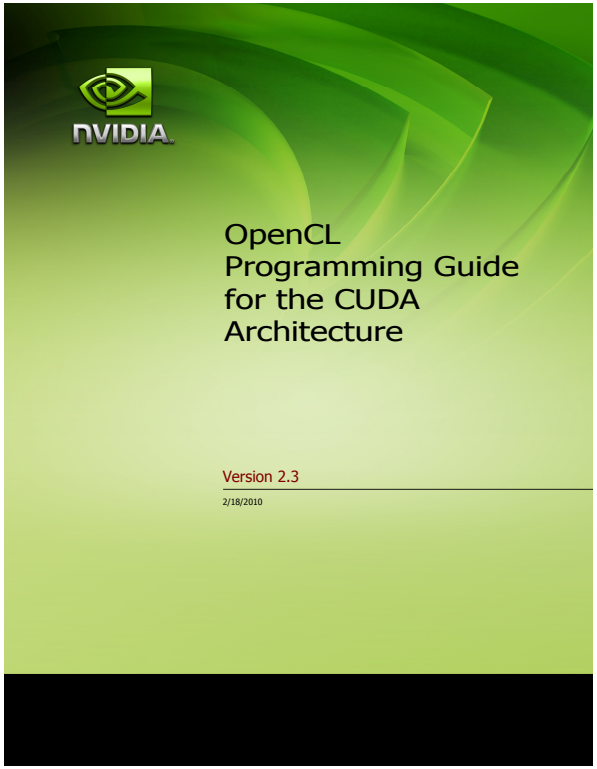
http://docs.nvidia.com/cuda/index.html

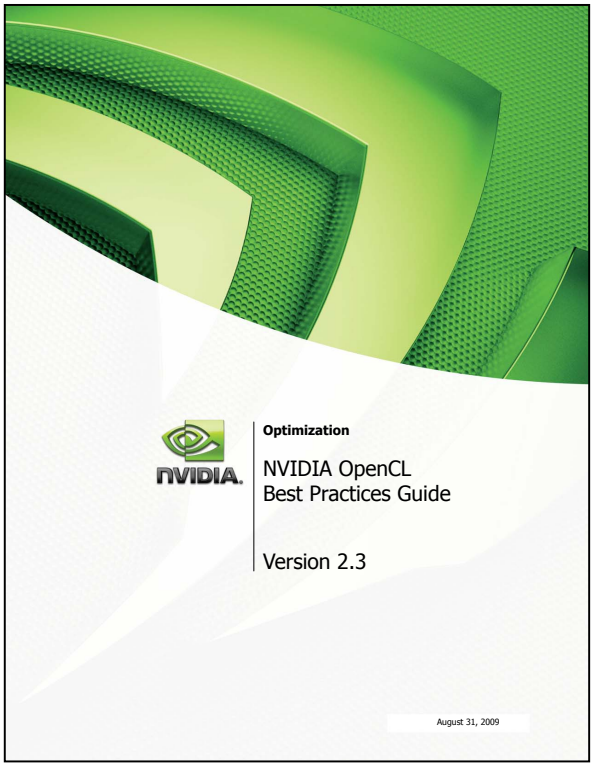
What is OpenCL?

- **Open C**omputing **L**anguage (current v3.0)
- Khronos group (non-profit organization):
  - "*OpenCL is an open, royalty-free standard for cross-platform, parallel programming of modern processors found in parallel computers, servers and handheld/embedded devices*."
- Open standard for heterogeneous computing
- Priority #1: Become the *industry-wide future standard* for heterogeneous computing
- Priority #2: Use all computational resources in the system efficiently
- Up to vendors to provide support!

KHRONOS GROUP

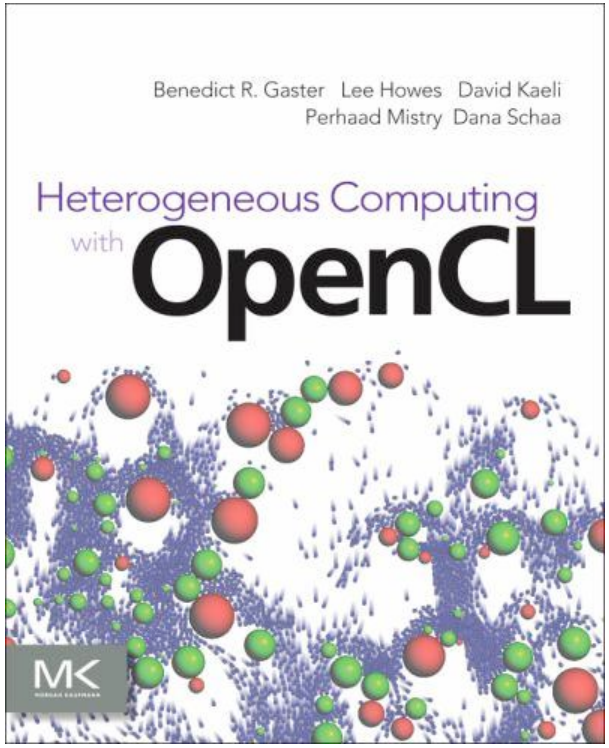
Free OpenCl material

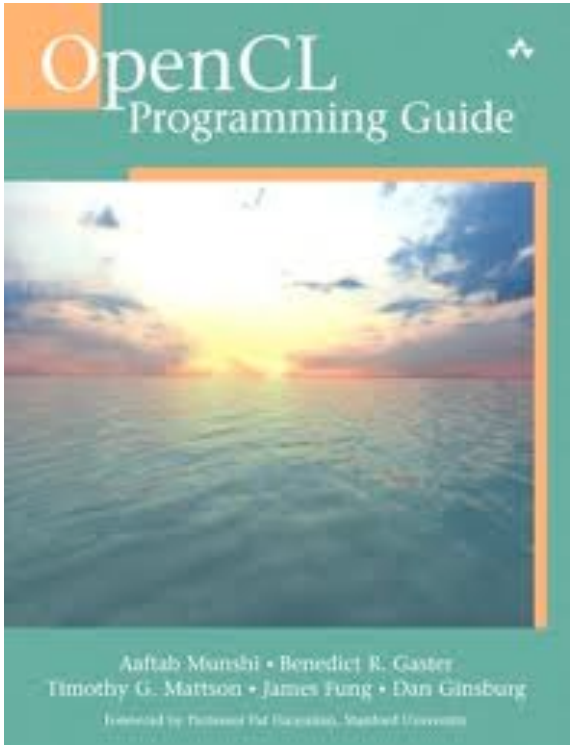


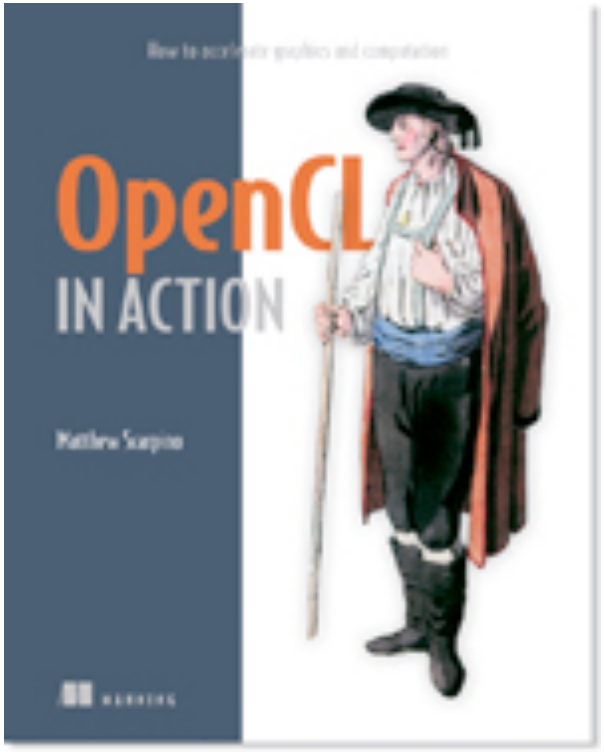


- Free from Nvidia and AMD developer webpages
  - https://developer.nvidia.com/opencl
  - Pdf versions, see installation; /usr/local/cuda/doc/pdf/

Additional OpenCL material







- ~~Feel free to come by 324-280 to take a peek~~

Why CUDA in this course?

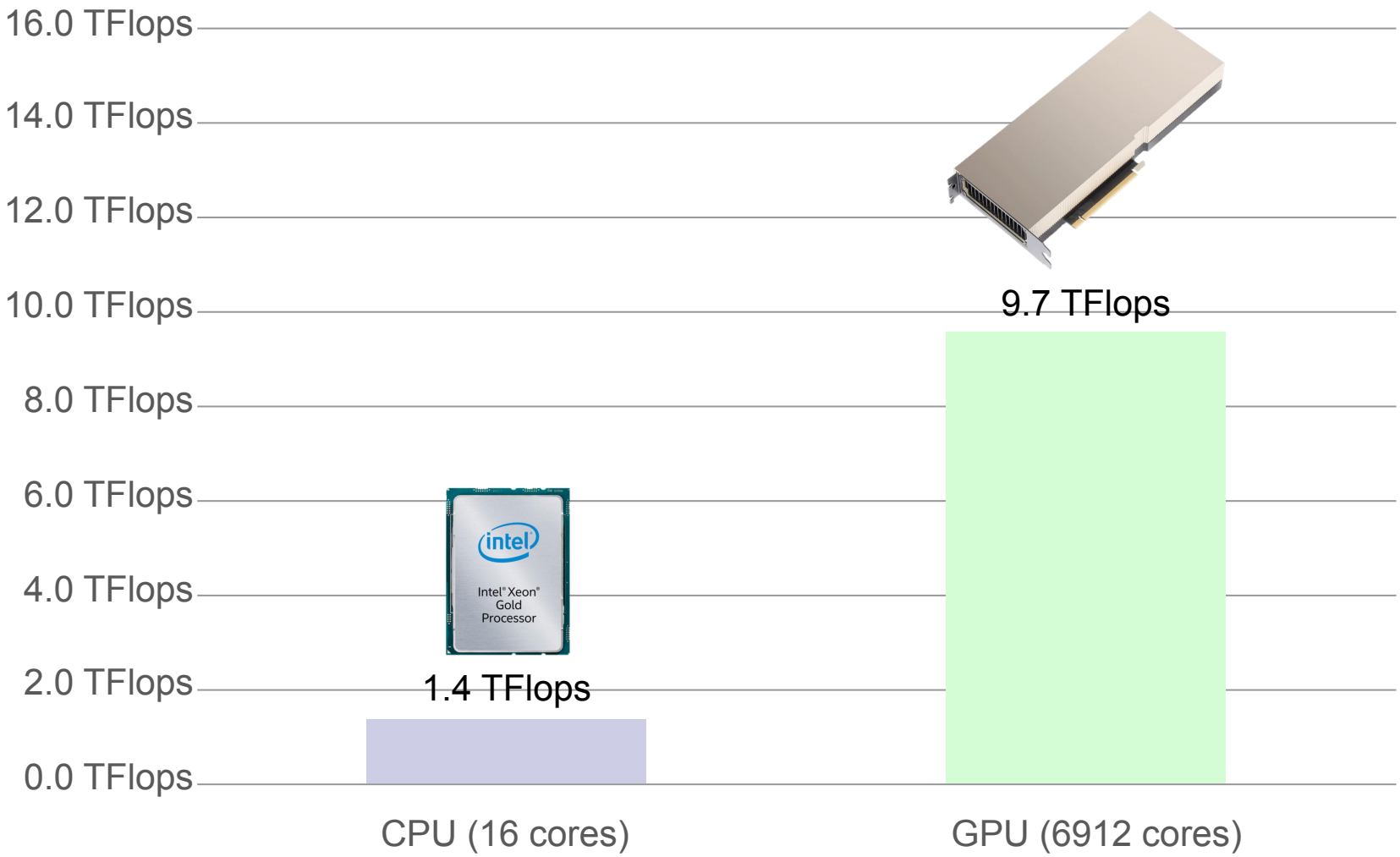
- CUDA and OpenCL are very similar
  - Most CUDA features map one-to-one to OpenCL features (only the syntax is different)
  - OpenCL provides more explicit functionality and is supported by a less mature software framework
- CUDA comes with tuned high-performance libs
  - OpenCL have less effort in this direction (currently)
- CUDA is well documented (by Nvidia)
- Nvidia products are widely used in HPC (>90%)
  - OpenCL still lags in performance for Nvidia products

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?

1555 GB/s

141 GB/s

1600 GB/s
1400 GB/s
1200 GB/s
1000 GB/s
800 GB/s
600 GB/s
400 GB/s
200 GB/s
0 GB/s

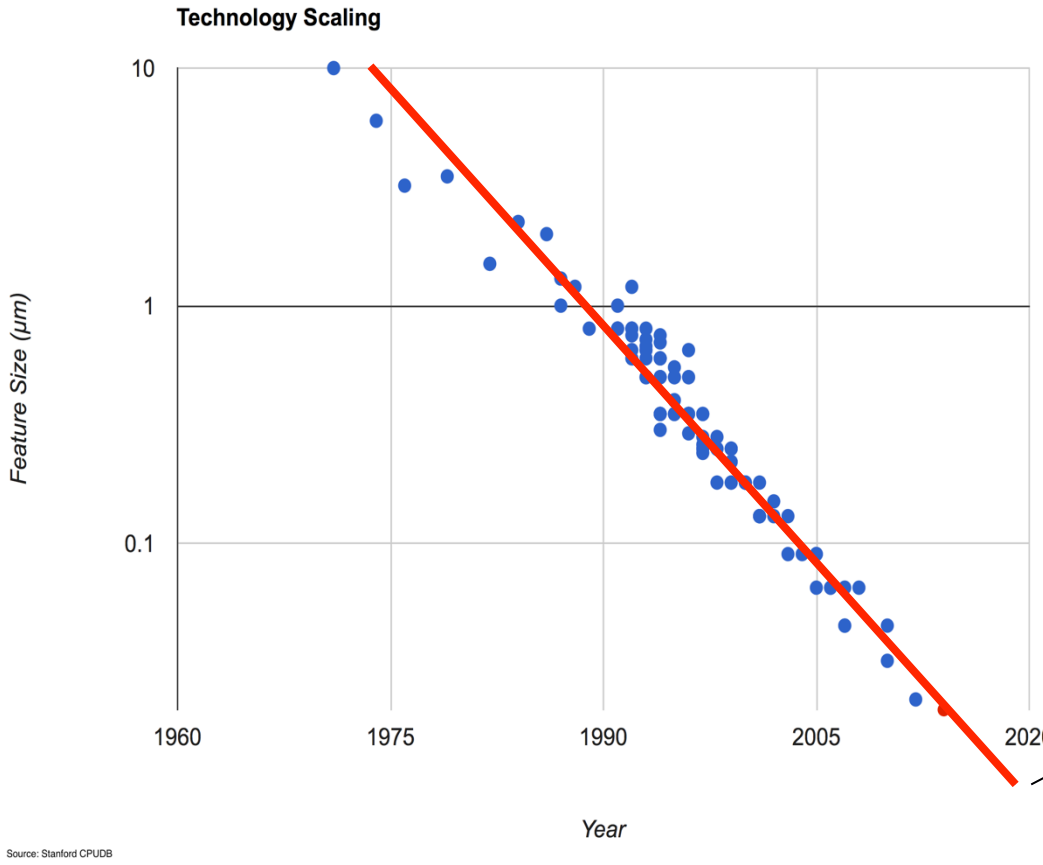
CPU (16 cores)

GPU (6912 cores)

Why are GPUs different from CPUs?

Recap from week 1: Good news

- Transistor size over the years; still decreases

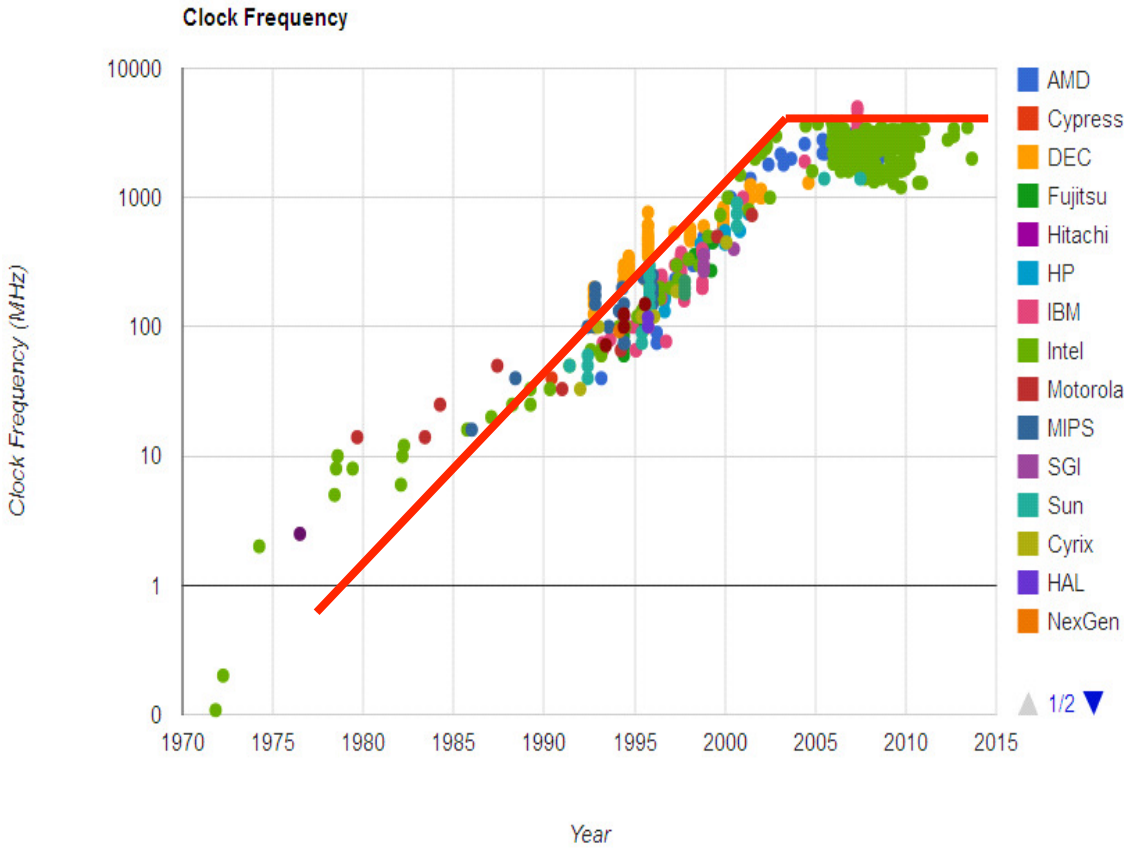


- Smaller
- Faster
- Less power
- More on chip

Currently best: 5 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

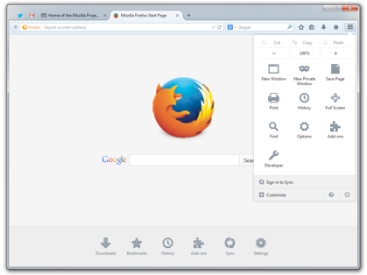
- What is the goal? Roughly two choices...

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



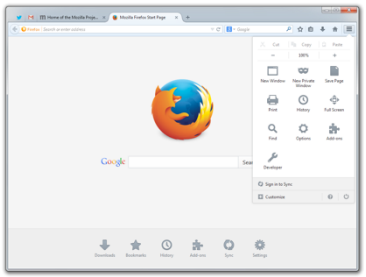
Mac OS

Windows 10

Microsoft Office

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

CPUs target latency (traditionally)



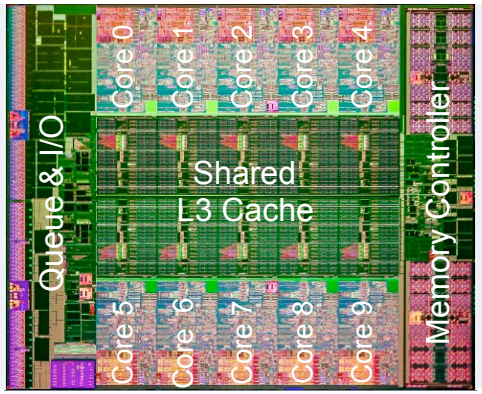
- Usual task of traditional CPUs
  - Desktop applications / OS
    - Lightly threaded
    - Lots of branches
    - Lots of (indirect) memory accesses

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

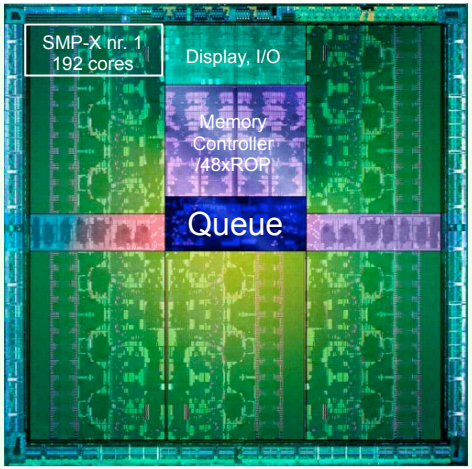
- GPUs are designed to compute pixels – fast!
  - Rendering video games in real-time
  - Play HD movies on smart phones
  - Render visual effects for movies…

- 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)

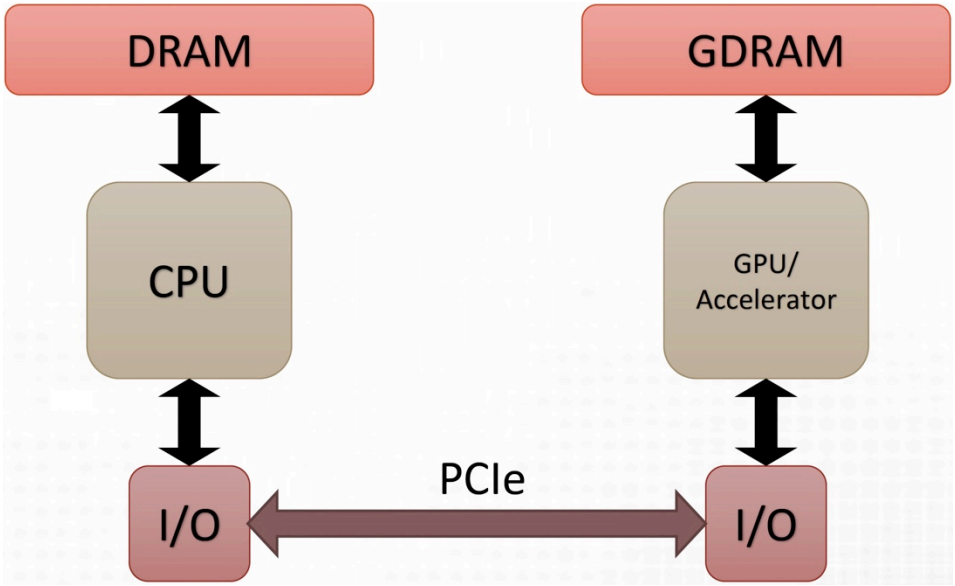
Hardware hierarchy

- 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

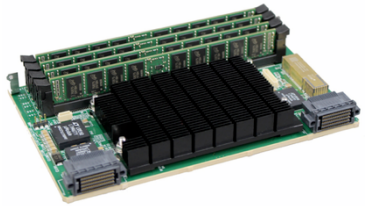


Types of accelerators

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

- Intel Xeon Phis
  - Many Integrated Cores (MIC) architecture
  - Based om Pentium 4 with wide vectors
  - Closer to traditional CPU / same compiler

- Custom many-core processors
  - Japan / China



Nvidia GPU architectures

- Four 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

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 ~4 years!

Accelerators in Top500

DTU

Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442,010.0	537,212.0	29,899
2	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148,600.0	200,794.9	10,096
3	**Sierra** - IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM / NVIDIA / Mellanox DOE/NNSA/LLNL United States	1,572,480	94,640.0	125,712.0	7,438
4	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway, NRCPC National Supercomputing Center in Wuxi China	10,649,600	93,014.6	125,435.9	15,371
5	**Selene** - NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100, Mellanox HDR Infiniband, Nvidia NVIDIA Corporation United States	555,520	63,460.0	79,215.0	2,646

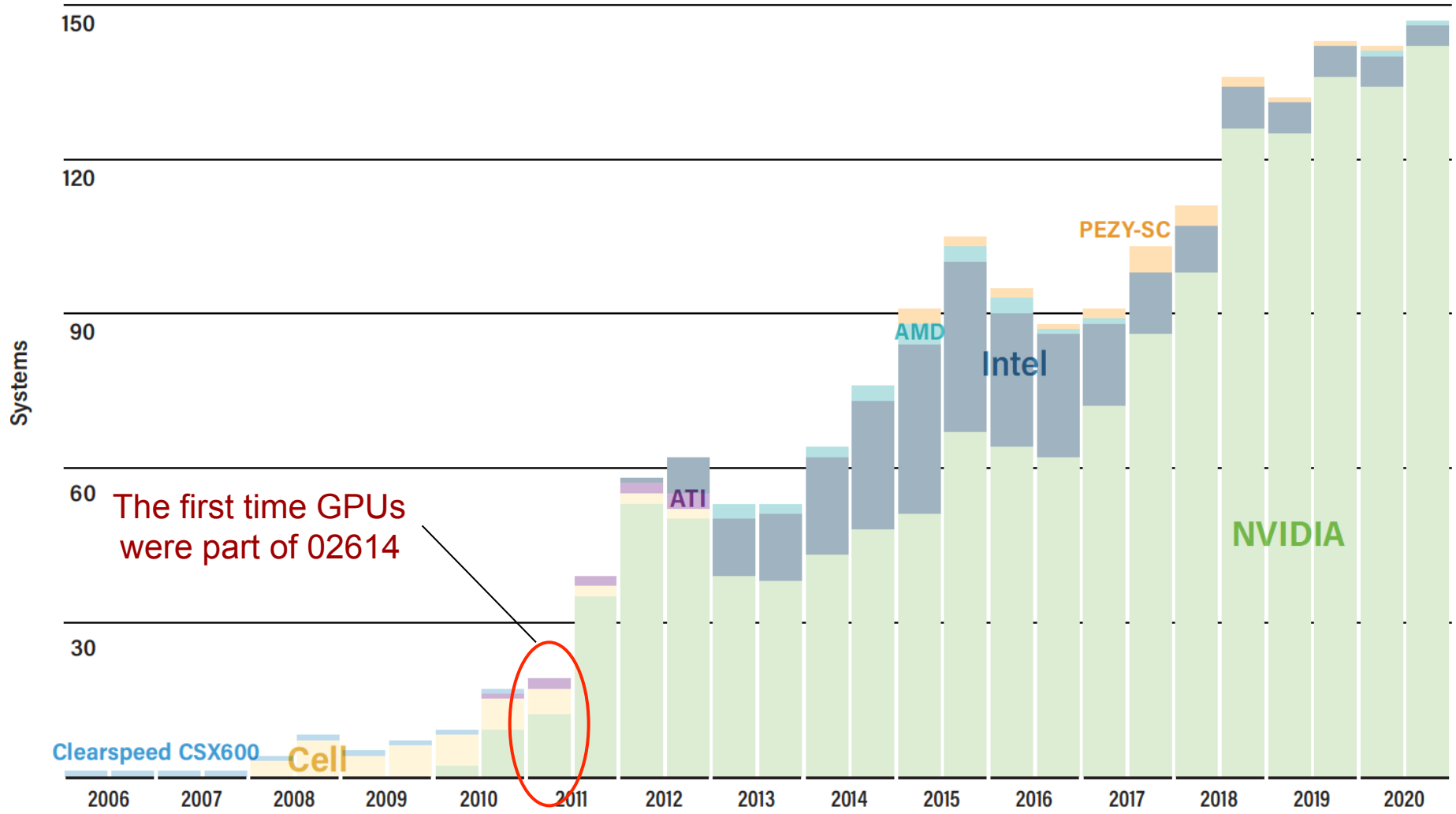
Nvidia Tesla (Volta)

Nvidia Tesla (Volta)

RISC processor with 260 cores

Nvidia Tesla (Ampere)

Accelerators in Top500

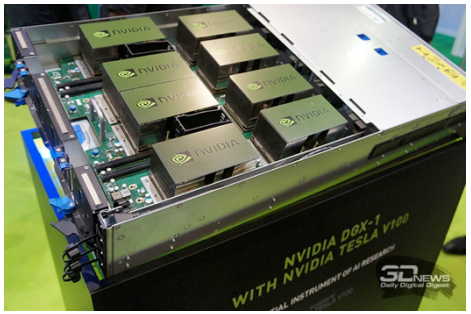


Trends in HPC due to GPUs

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

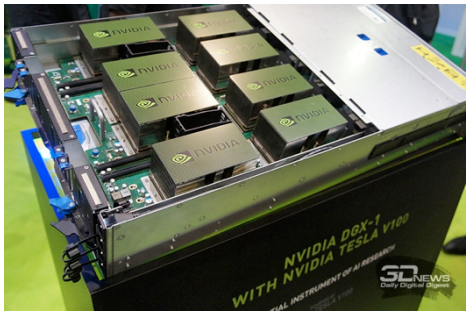
Trends in HPC due to GPUs

- 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

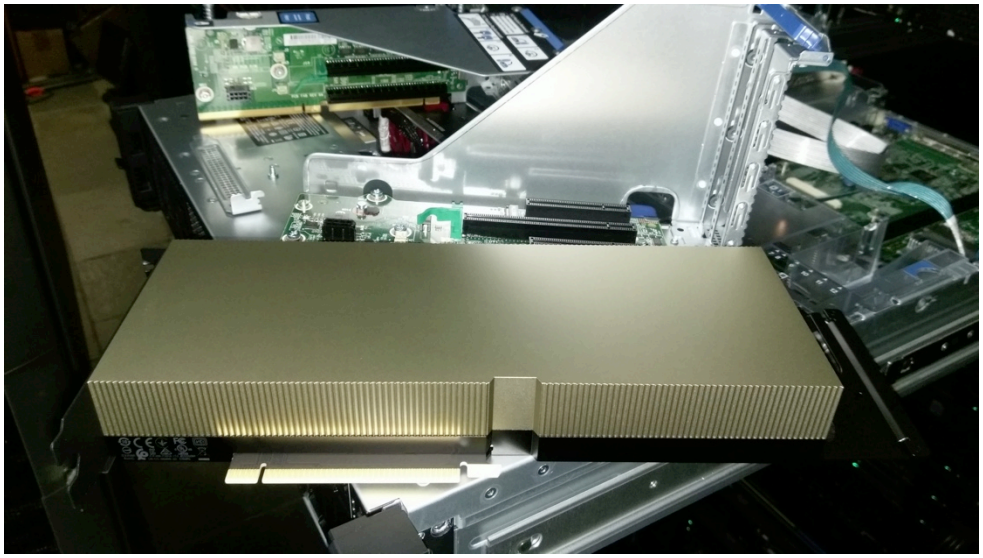
- Improvements at individual computer node level are greatest
  - Less data transfer
  - Heterogeneous computing
  - Avoiding MPI
- "Super-nodes": Nvidia DGX-1
  - Eight tightly linked high-end GPUs
  - 40,960 cores / 960 TFlops in 1 node
- Communication costs are increasing
  - Synchronization-reducing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



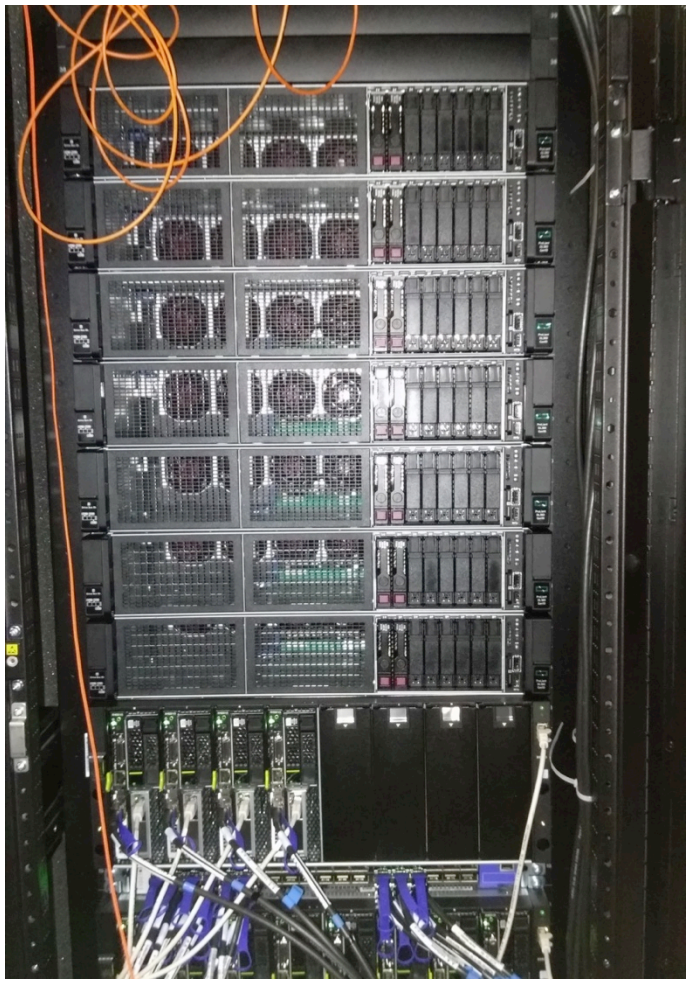
GPU hardware for this course

Ampere cluster for 02614

Delivered: December 23, 2020



1 HPE node:
2 x Intel Xeon Gold 6126 CPU @ 2.90GHz (16 cores)
2 x Tesla A100-PCIE-40GB (6912 cores)
768 GB DDR4 @ 2666MHz

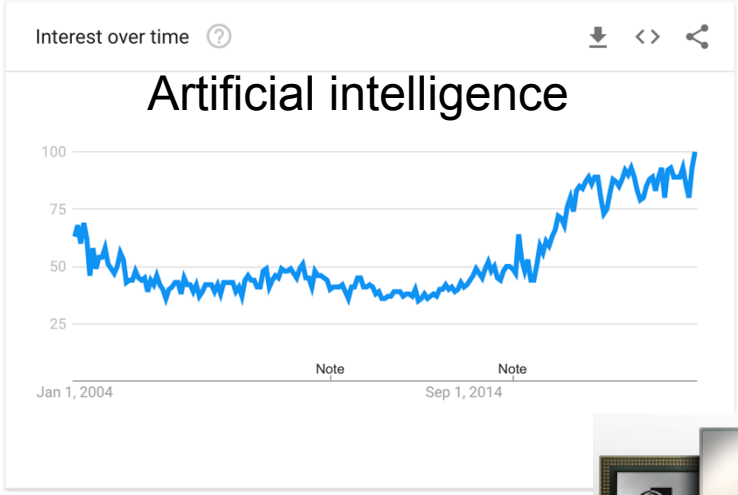


6 HPE nodes:
Theoretical peak: 133,2 TFlops (FP64)

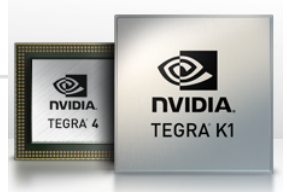
Google trends: GPUs still going up

THE WORLD'S FIRST AI SUPERCOMPUTER IN A BOX

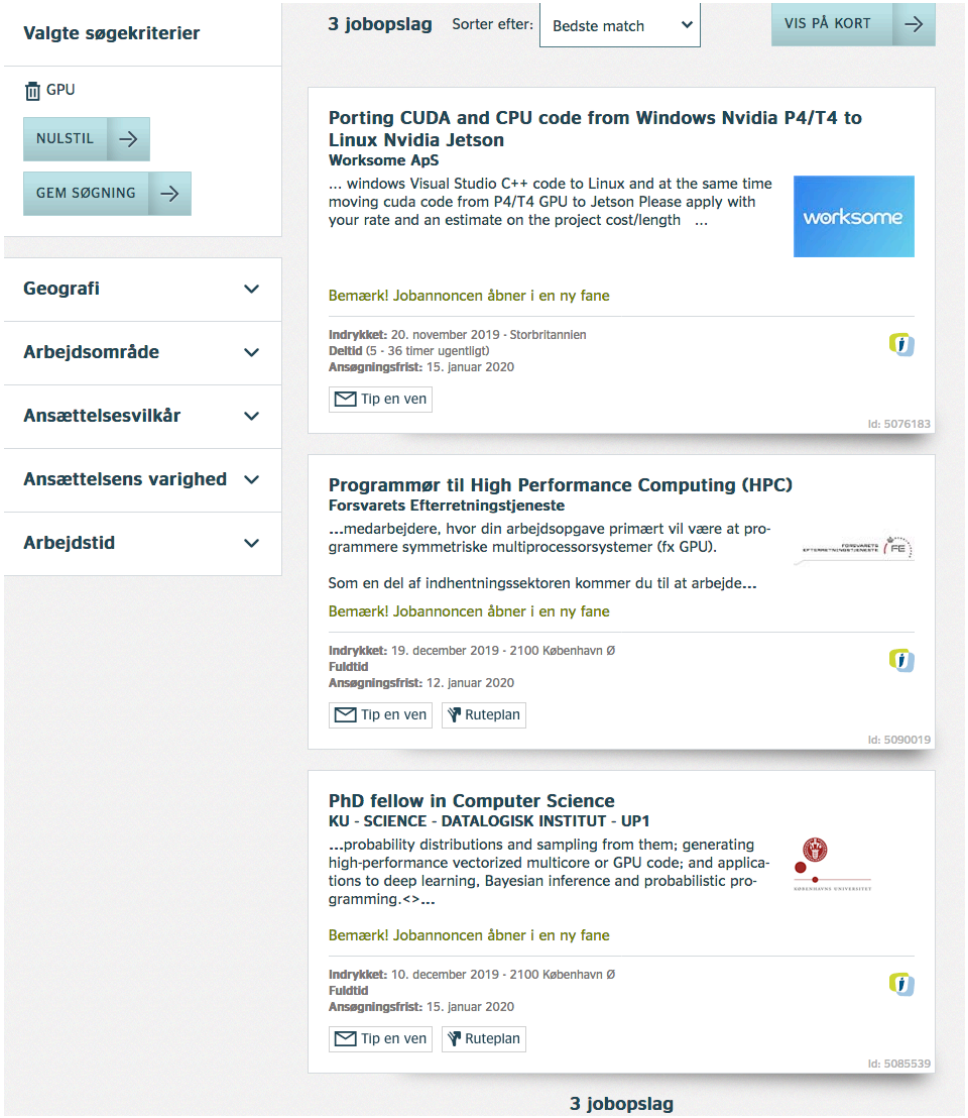
Get faster training, larger models, and more accurate results on deep learning with the NVIDIA® DGX-1™. This is the world's first purpose-built system for deep learning and AI accelerated analytics,



TEGRA K1—THE WORLD'S MOST ADVANCED MOBILE PROCESSOR



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 2021

This January six jobs available!

Experienced Machine Learning Software Developer

salling group

Salling Group, Brabrand

We're looking for an ambitious and experienced software developer to come help us achieve our vision of building new smart services leveraging machine learning technologies to help our customers.

You will build our machine learning consumer-facing services that support our websites, apps, partners and 3rd party offerings and continue the work we've already started.

You will be part of a well-functioning team of 4 ML developers and a product owner. The team is responsible for AI and machine learning initiatives in Salling Group and helps get machine learning services to market and expose them through our API Management Platform. The product owner will help set the direction and scope for our machine learning services.

Din søgning

× GPU

Opret Jobagent

Filtrér din søgning

TELEDYNE MARINE Everywhereyoulook™

C# / WPF .NET Software Developer, Teledyne Marine product lines

HEGSØ Search & Consulting recruiting on behalf of Teledyne Marine

C# / WPF / .NET developer to support us making software for our customers and industry leading product portfolio. Joining our software team focusing on the UI development side, working in a global team with a mix of new and experienced developers on our software technology platform. We support all sensor product lines from a shared software framework, delivering multiple UI systems based on same platform. We work with logic/math execution in highly efficient C/C++ code, and we use .NET/WPF on the UI side.

MR Physicist, part time, Nuklearmedicin og PET, Aarhus Universitetshospital

midt regionmidtjylland

The Department of Nuclear Medicine & PET at Aarhus University Hospital invites applications for a position as part- time MR physicist with...

Region Midtjylland, Aarhus N

Region Midtjylland

☆ Gem job | J Følg | Del | Indrykket 18. december 2020

Om virksomheden ★★★☆☆ (513) | QuickApply

Software Engineer, Portfolio Analytics

SimCorp A/S, Copenhagen

This is an unique opportunity to join our Portfolio Analytics and Reporting team who are responsible for delivering Risk and Performance analytics to our end users. Would you like to work with a highly complex product in the best possible culture or just considering new opportunities within .NET , MS Azure and High-Performance-Computing (HPC), then we would love to hear from you!

As our new Software Engineer, you will join SimCorp's Product Division, a full-scale agile organization following SAFe. Here, you become part of a cross functional agile team based in central Copenhagen. Your team is responsible for delivering functionality within a specific module of our product SimCorp Dimension. You apply your talents to all stages of the development lifecycle, including creation and review of user stories, development, design, testing, coding, code reviews, writing automated test, and support.

SimCorp

Two postdoc positions in Machine Learning Models for all Atom Simulations of Enzymatic Catalysts and Electrochemical Battery Interfaces

Som erhvervsorienteret universitet er målet at levere forskning på et højt internationalt niveau, der tager udgangspunkt i teori- og modelopbygning og empiri.

Danmarks Tekniske Universitet, København

Danmarks Tekniske Universitet

☆ Gem job | J Følg | Del | Indrykket 14. december 2020

Om virksomheden

End of lecture