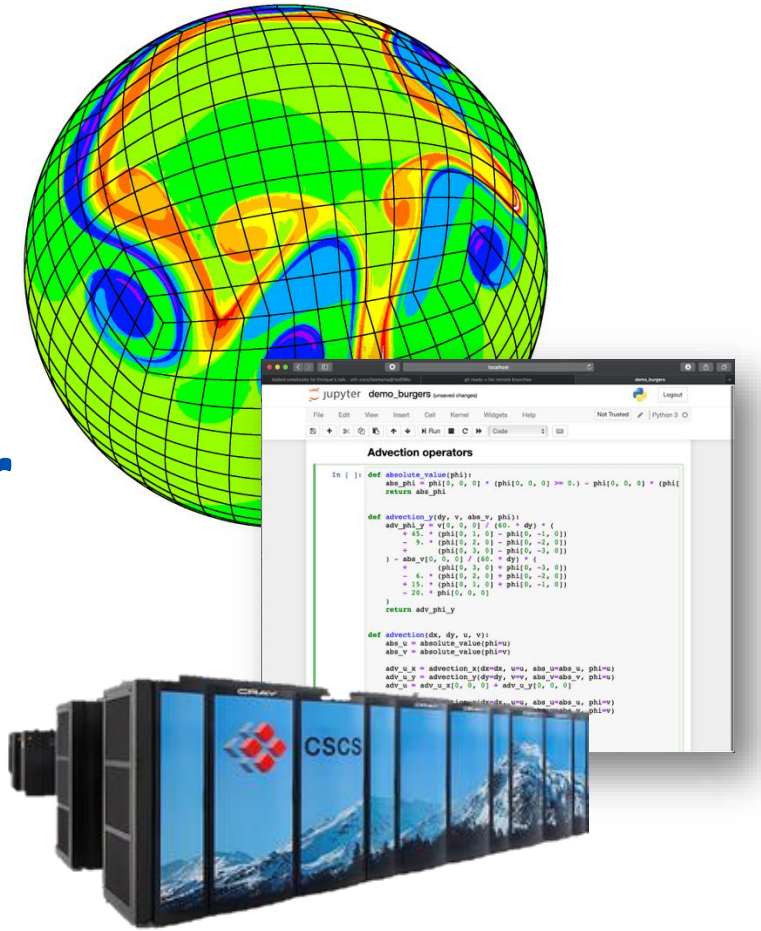
High Performance Computing for Weather and Climate (HPC4WC)

Content: Graphics Processing Units
Lecturer: Simon Adamov, Oliver Fuhrer
Block course 701-1270-00L
Summer 2025



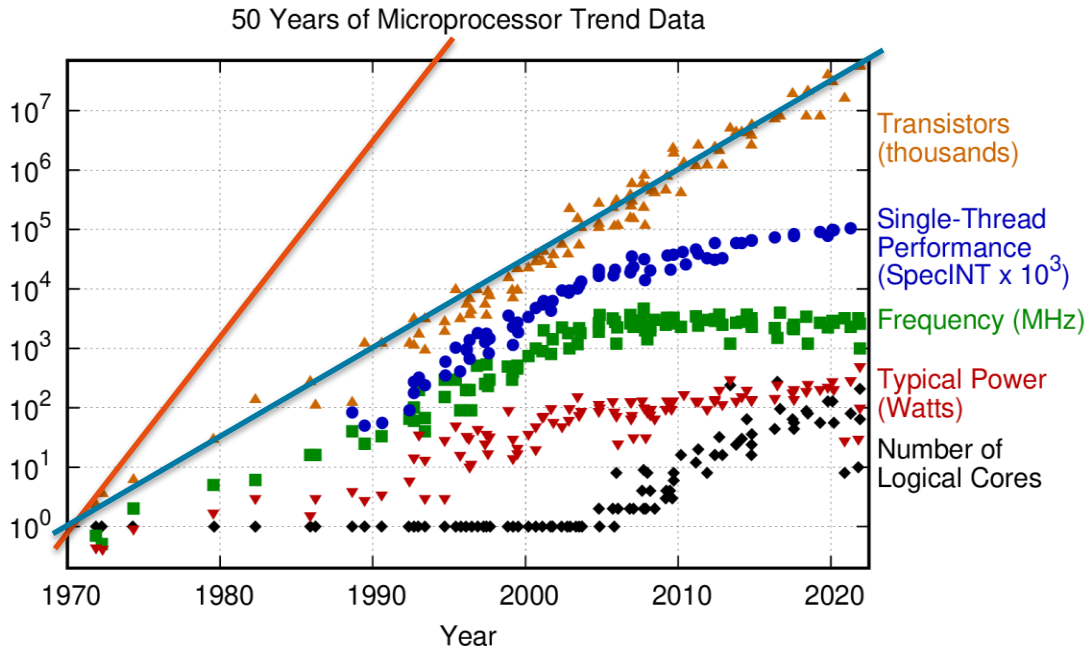
Learning Goals

- Understand why specialized hardware such as GPUs has become the new norm
- Learn how to program a GPU using a high-level programming language
- Grasp the potential and difficulties of GPU computing

Moore's "Law" (1965)

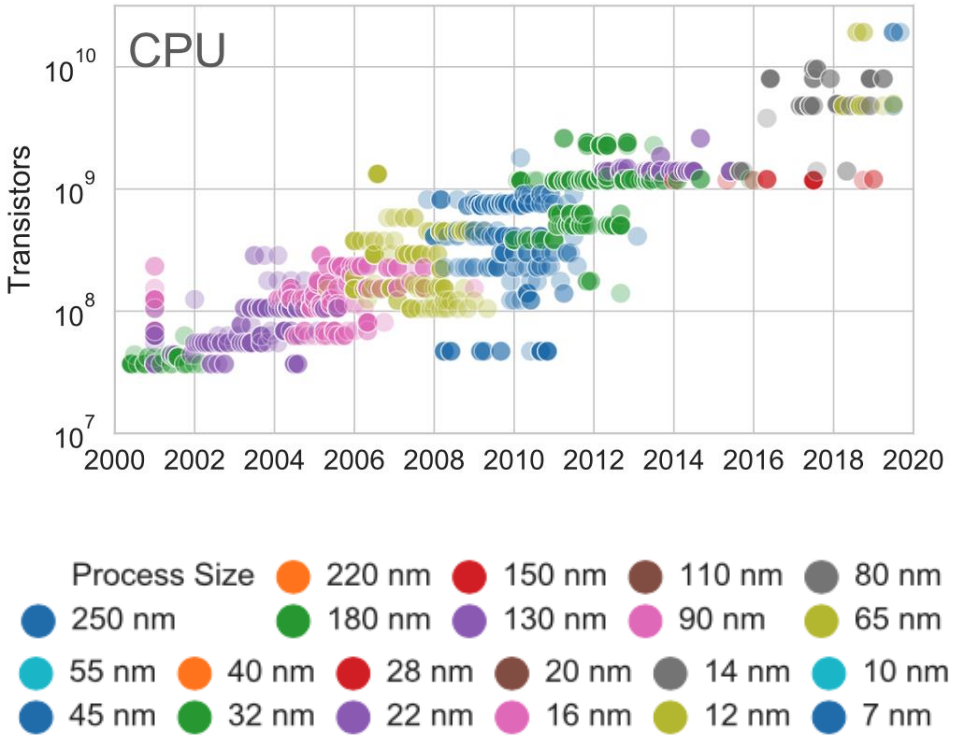
- "The number of transistors in a dense integrated circuit will double every two years"

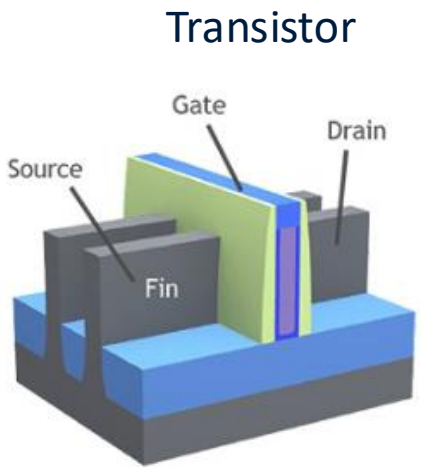
--- Yearly doubling
--- Biannual doubling



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2021 by K. Rupp

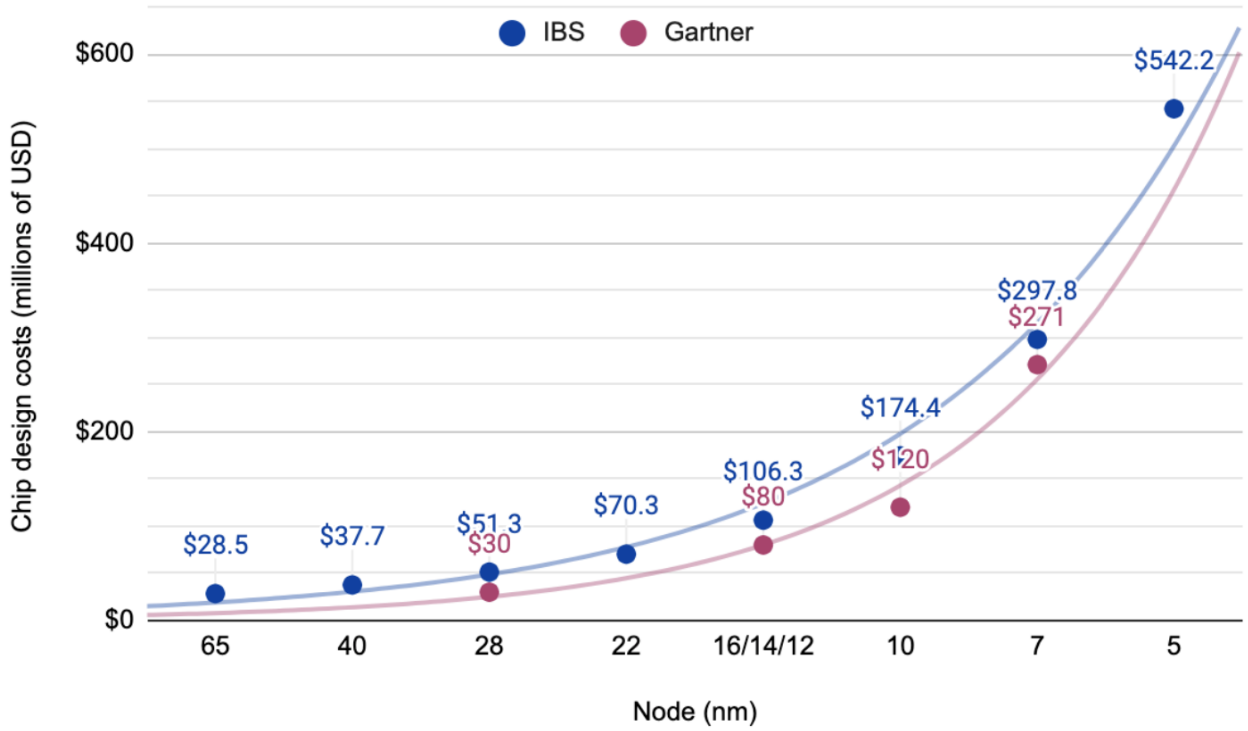
The End of General Purpose Computing?





Distance between Si-atoms is 0.21 nm!
Currently 3nm designs are available.

Chip Design Costs Increase



Source: top500.org

How Do the Most Powerful Systems Perform?

--- Yearly doubling

So why are we (still) ok?

Source: top500.org

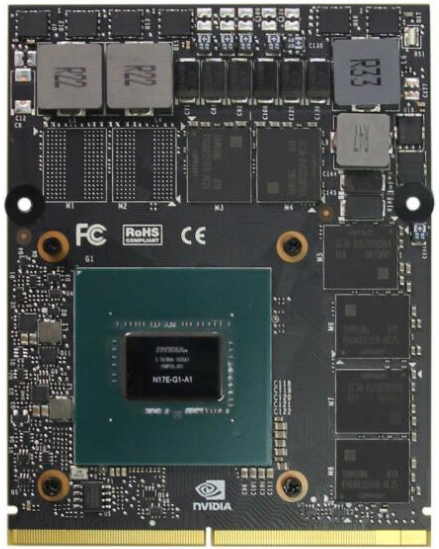
Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)	
1	**El Capitan** - HPE Cray EX255a, AMD 4th Gen EPYC 24C 1.8GHz, AMD Instinct MI300A, Slingshot-11, TOSS, **HPE** DOE/NNSA/LLNL United States	11,039,616	1,742.00	2,746.38	29,581	AMD GPU
2	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation 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	AMD GPU
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	INTEL GPU
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	793.40	930.00	13,088	NVIDIA GPU
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		NVIDIA GPU

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)	
6	**HPC6** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, RHEL 8.9, **HPE** Eni S.p.A. Italy	3,143,520	477.90	606.97	8,461	AMD GPU
7	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, **Fujitsu** RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899	
8	**Alps** - HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200 Superchip, Slingshot-11, HPE Cray OS, **HPE** Swiss National Supercomputing Centre (CSCS) Switzerland	2,121,600	434.90	574.84	7,124	NVIDIA GPU
9	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, **HPE** EuroHPC/CSC Finland	2,752,704	379.70	531.51	7,107	AMD GPU
10	**Leonardo** - BullSequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 SXM4 64 GB, Quad-rail NVIDIA HDR100 Infiniband, **EVIDEN** EuroHPC/CINECA Italy	1,824,768	241.20	306.31	7,494	NVIDIA GPU

Specialized Chips are on the Rise!

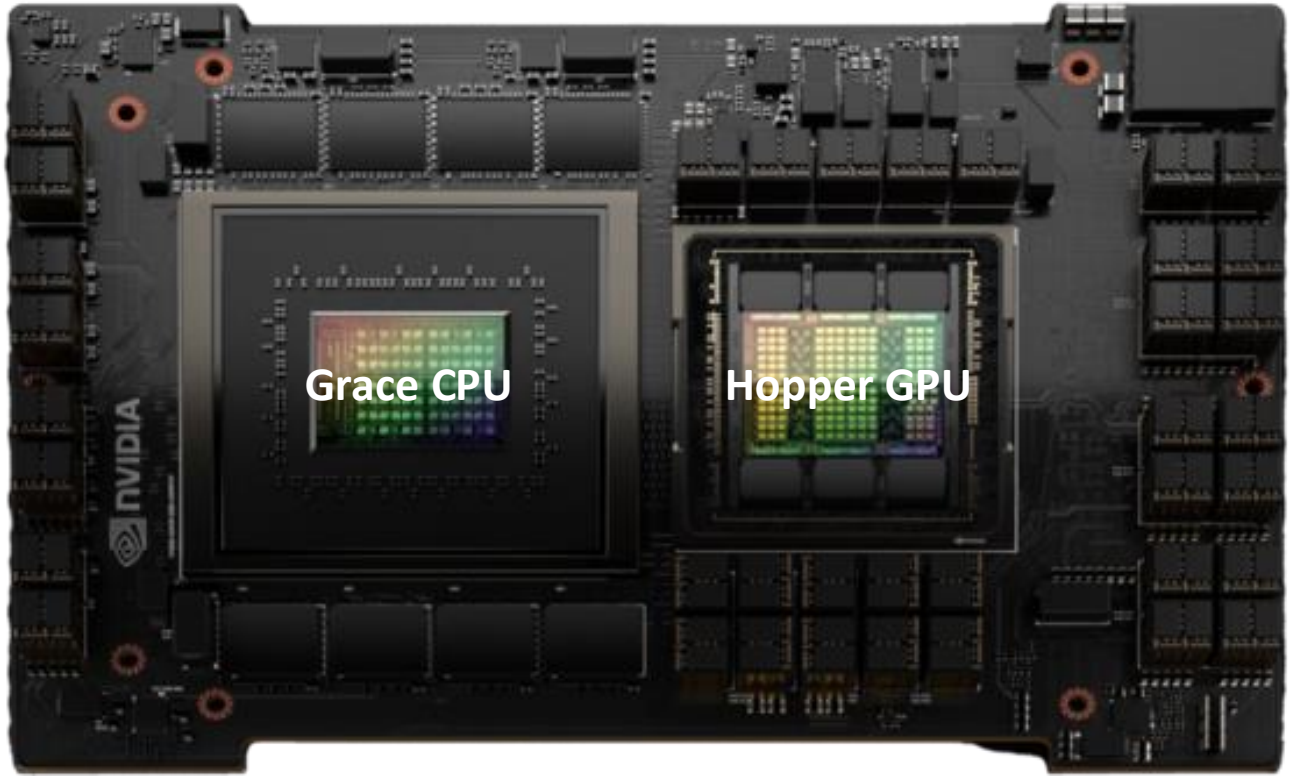
Google's TPU
(e.g. machine learning)

FPGA
(e.g. bitcoin mining)



GPU
(e.g. gaming)

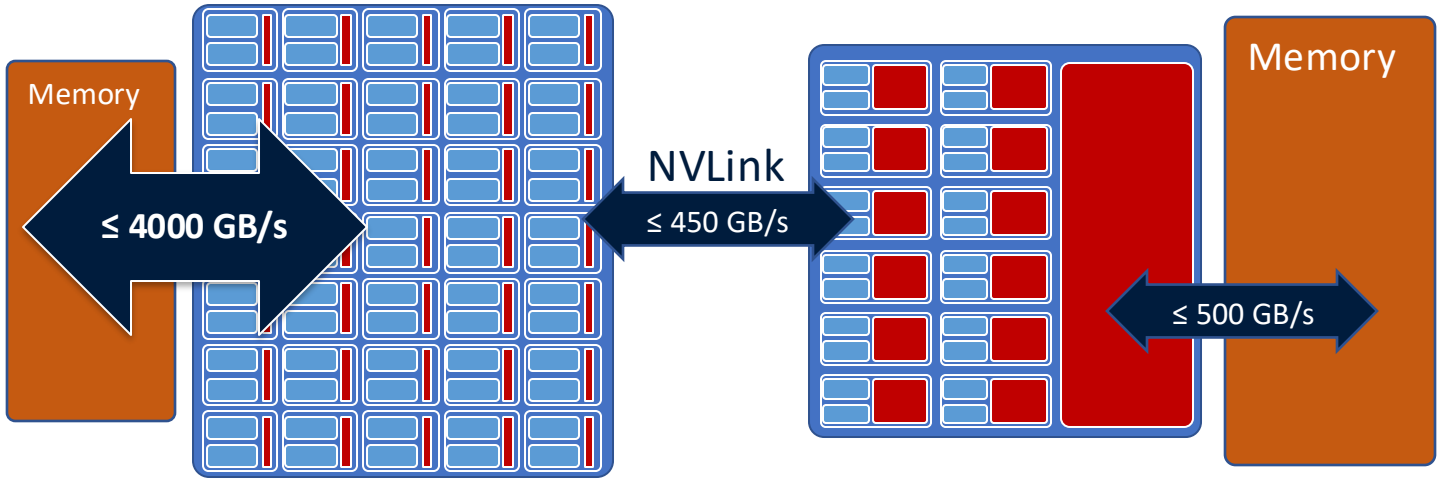
The NVIDIA GraceHopper 200



~ 200 W ≤ 2.5 TFLOP/s ≤ 500 GB/s

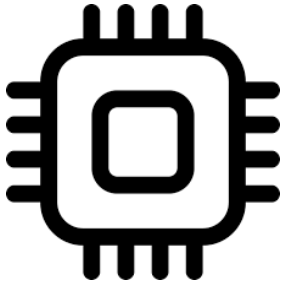
~ 700 W ≤ 34 TFLOP/s ≤ 4000 GB/s

Node Architecture

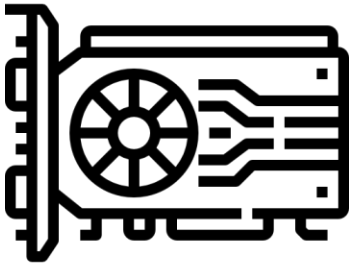


Important to minimize memory transfers between CPU and GPU!

CPU vs. GPU



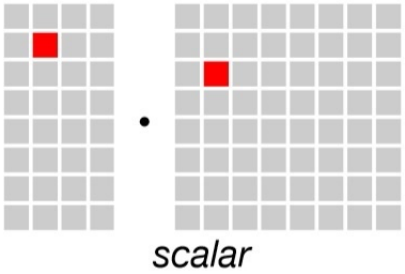
Architecture



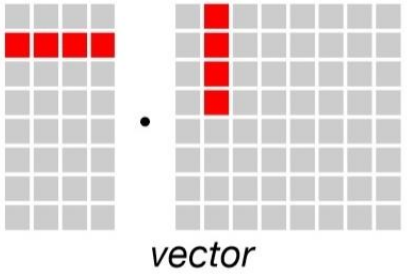
Latency

Optimization

Bandwidth



Compute Primitive

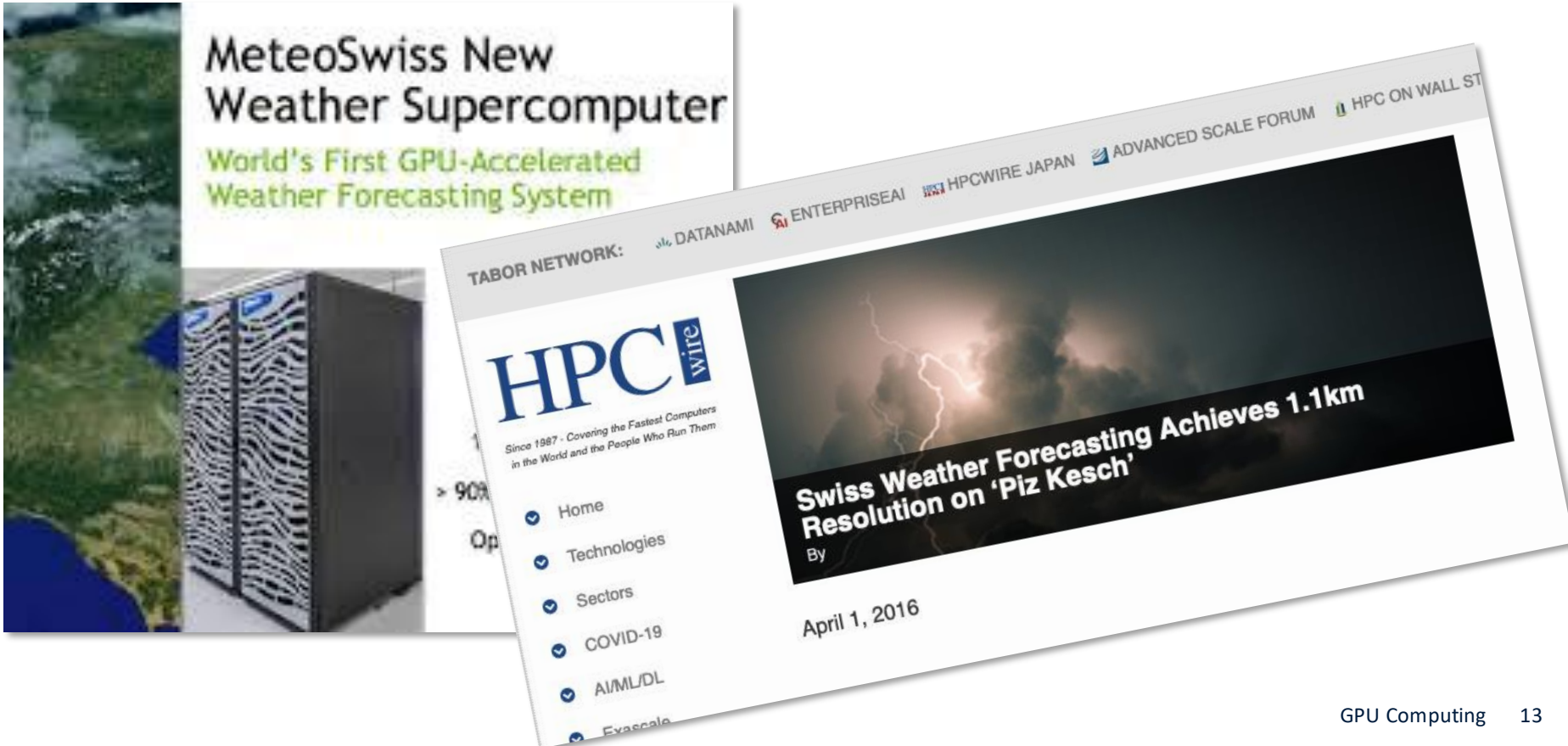


Hybrid Supercomputer - ALPS

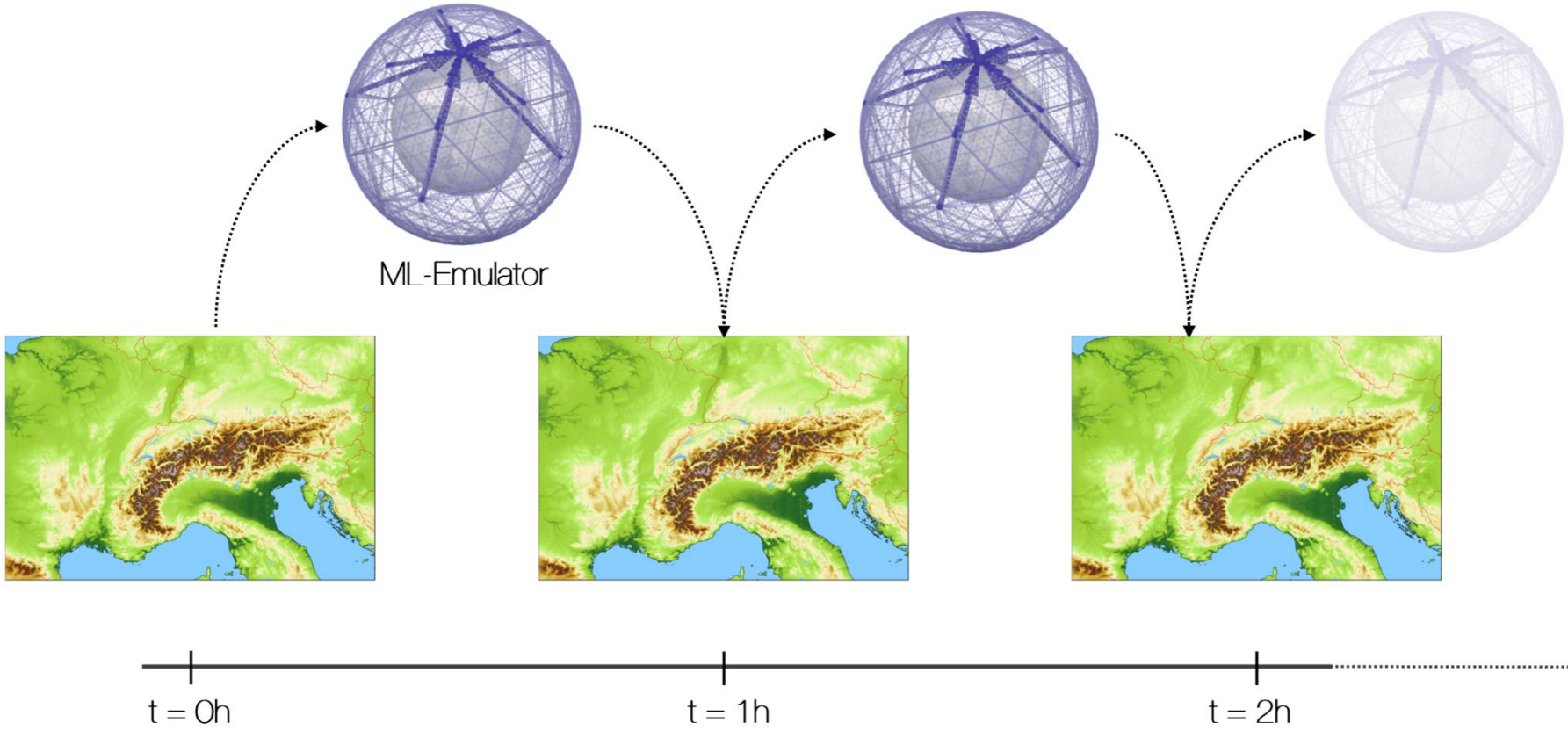
~ 90% BW

~ 80% FLOP/s

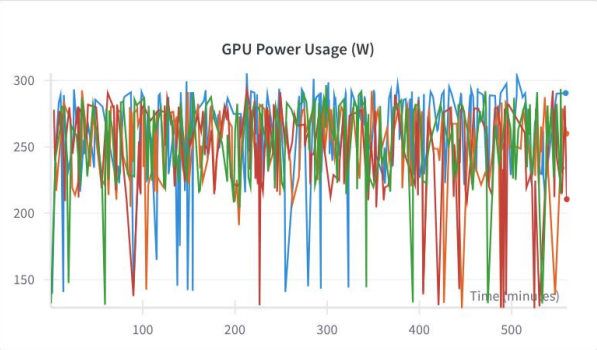
Weather and Climate on GPUs

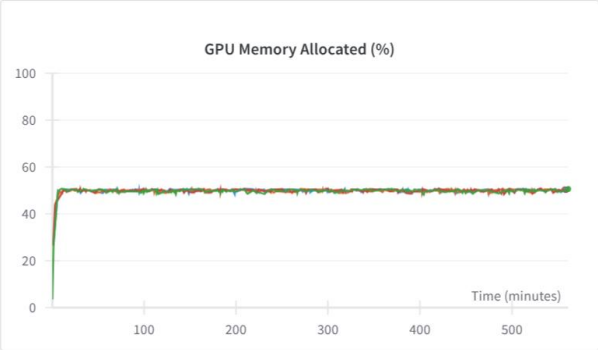


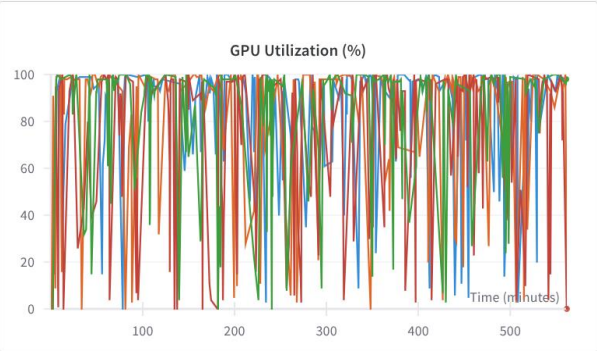
ML-Weather Modelling

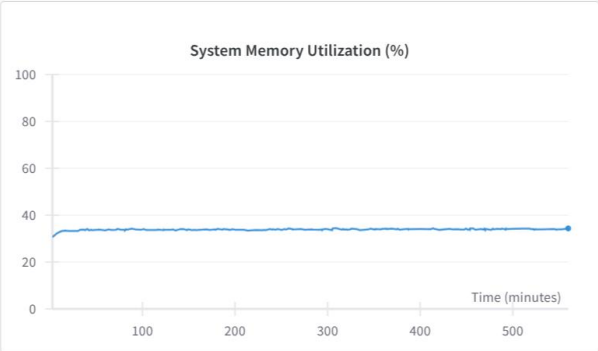


Pitfalls and Challenges









Programming GPUs

Libraries

CuPy

Directives

OpenACC

Directives for Accelerators

Programming Models

NVIDIA

CUDA

Lab Exercises

01-GPU-programming-cupy.ipynb

- Introduction to GPU programming using a high-level programming language