VLAAMS SUPERCOMPUTER CENTRUM

Vlaanderen is supercomputing

Artificial Intelligence and GPUs

VSC Lunch Session on A.I. (Oct 7th 2021)

Kenneth Hoste (VSC, HPC-UGent)

VLAAMS SUPERCOMPUTER CENTRUM

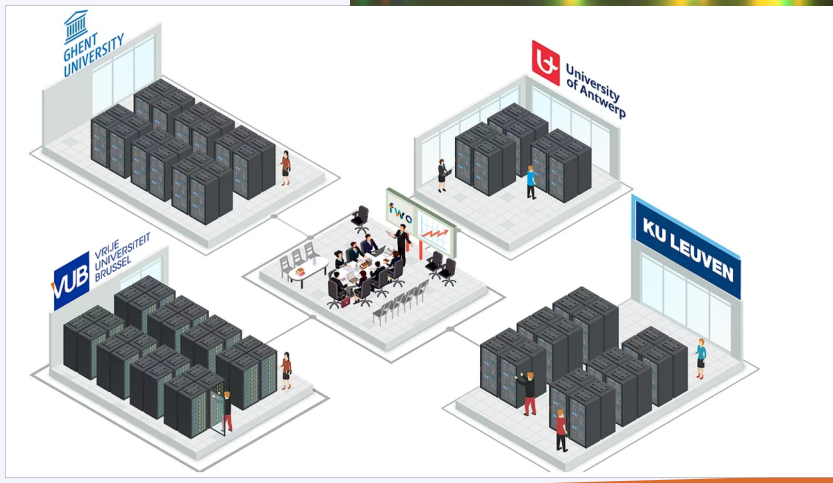
Vlaanderen is supercomputing

fwo

Vlaams Supercomputer Centrum (VSC)

https://www.vscentrum.be - https://docs.vscentrum.be

- **Partnership between 5 Flemish university associations**, managed & co-funded by FWO
- Virtual center - infrastructure in 4 hubs (KU Leuven, UAntwerpen, UGent, VUB)
- Provides **infrastructure, support, training** for scientific and technical computing
- Tier-2 + Tier-1 compute clusters, VSC cloud, VSC data component
- **One VSC account** to access all VSC infrastructure
- Access for both **academic researchers** (free of charge),
  research institutes and industry (pay what you use, free exploratory)



Kenneth Hoste

GHENT UNIVERSITY

@kehoste

- Masters + PhD in Computer Science from Ghent University (2010)
- Dissertation topic: *"Analysis, Estimation and Optimization of Computer System Performance Using Machine Learning"*
- **HPC system administrator at Ghent University + VSC since 2010**
- Main tasks: user support, software installations, training
- Lead developer of EasyBuild (tool to install scientific software on HPC systems)

easybuild

https://easybuild.io

AGENDA

Artificial Intelligence and GPUs

1. The Deep Learning Revolution

4. Programming Models for GPUs

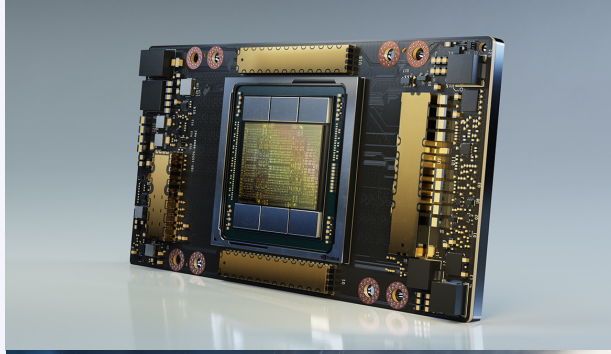
2. GPUs for Computational Science

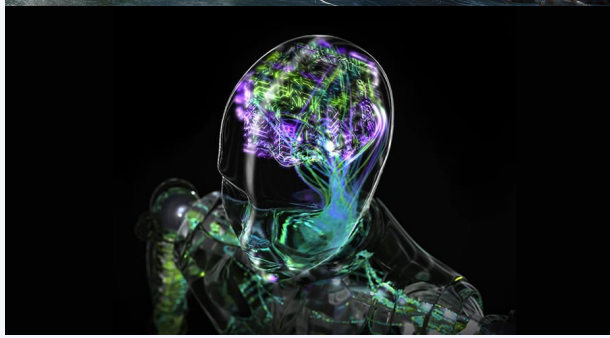
5. Practical guidelines for using GPUs

3. GPU resources in the VSC

6. Example use case: AlphaFold 2

What is a GPU?





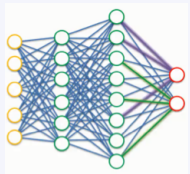
- **G**raphical **P**rocessing **U**nit, “video card” that can be used as a co-processor
- Originally created for Computer Graphics and Video Processing
- *Huge* market thanks to movie and video gaming industry
- Well suited for **processing large blocks of data in parallel**
- Also used as **accelerator** for (some) **scientific computations**
  - GPGPU: General Purpose Computing on GPU
  - Order(s) of magnitude speedup (10x - 100x, or more) is *possible*
- Recent top-of-the-line GPUs support **massively parallel computations**

The Deep Learning Revolution

Deep Learning in a nutshell

- Type of **machine learning** algorithm
- Multi-layer neural network (DNN)
- Inspired by biological systems
- Wide variety of **Artificial Intelligence (AI) applications**: game playing, natural language processing, bioinformatics, …

2009 — "Big Bang" of Deep Learning: **orders of magnitude speedup** for training deep neural networks on **GPUs**

2011 — 

Superhuman performance in visual pattern recognition contest using Convolutional Neural Networks (CNNs)

2012 — **Multi-task deep neural networks** used to win "Merck Molecular Activity Challenge" to predict the biomolecular target of one drug

2015 — AlphaGo

AlphaGo beats professional human player at board game Go (link)

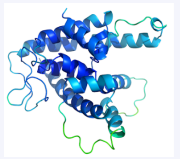
2017 — Release **of TensorFlow v1.0.0** by Google Brain team, with support for running on (multiple) GPUs

2018 — Release of **PyTorch v1.0.0**

TensorFlow is among most popular projects on GitHub, and is the most popular machine learning project

2019 — **Deepfake** technology goes mainstream: used by online communities and companies to replace faces in video content, later also applied to voices

Turing Award ("Nobel Prize" in Computer Science) for conceptual and engineering breakthroughs that have made deep neural networks a critical component of computing

2020 — 

AlphaFold 2 blows away competition in CASP14 competition by achieving high accuracy in **protein structure prediction** (link)

2021 — **More accurate rain prediction** using DeepMind's DGMR tool (link)

What Enabled the Deep Learning Revolution?

Big Data + Data Science

- Key to success for Deep Learning: lots of “examples”
- More (good) data ⟶ Better trained models
- Desire to extract information from tsunami of data (sensors, …)
- More data is sometimes easy to generate (AlphaGo)
- Wealth of data can enable scientific breakthroughs (AlphaFold)

GPUs

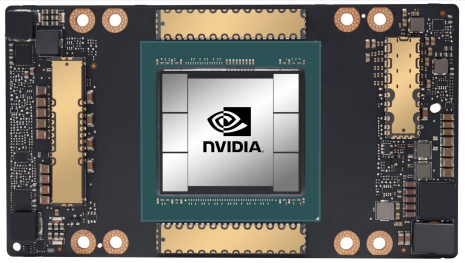
- GPUs are well suited for Deep Learning algorithms
- Large performance speedup compared to only using CPUs
- Rise of GPGPU computing fueled Deep Learning hype
- Deep Learning applications fueled GPGPU hype

Also:

algorithms, open source software (TensorFlow, PyTorch), …

GPUs for computational science

- In 2008 Nvidia started selling GPUs that were **well suited for scientific computations**
  - Double-precision floating-point support (FP64)
  - More and faster GPU memory
  - Fast interconnect between GPUs (NVLink)
  - Most recently: Nvidia Ampere series, incl. A100 with up to 80GB GPU memory (HBM2)



- Sparked the Deep Learning revolution
- **Order-of-magnitude speedup** is **possible** for specific types of computations (incl. deep learning & AI)
- Big impact on HPC systems infrastructure
- Porting traditional scientific software applications to leverage GPU resources took some time…

VLAAMS SUPERCOMPUTER CENTRUM

Vlaanderen is supercomputing

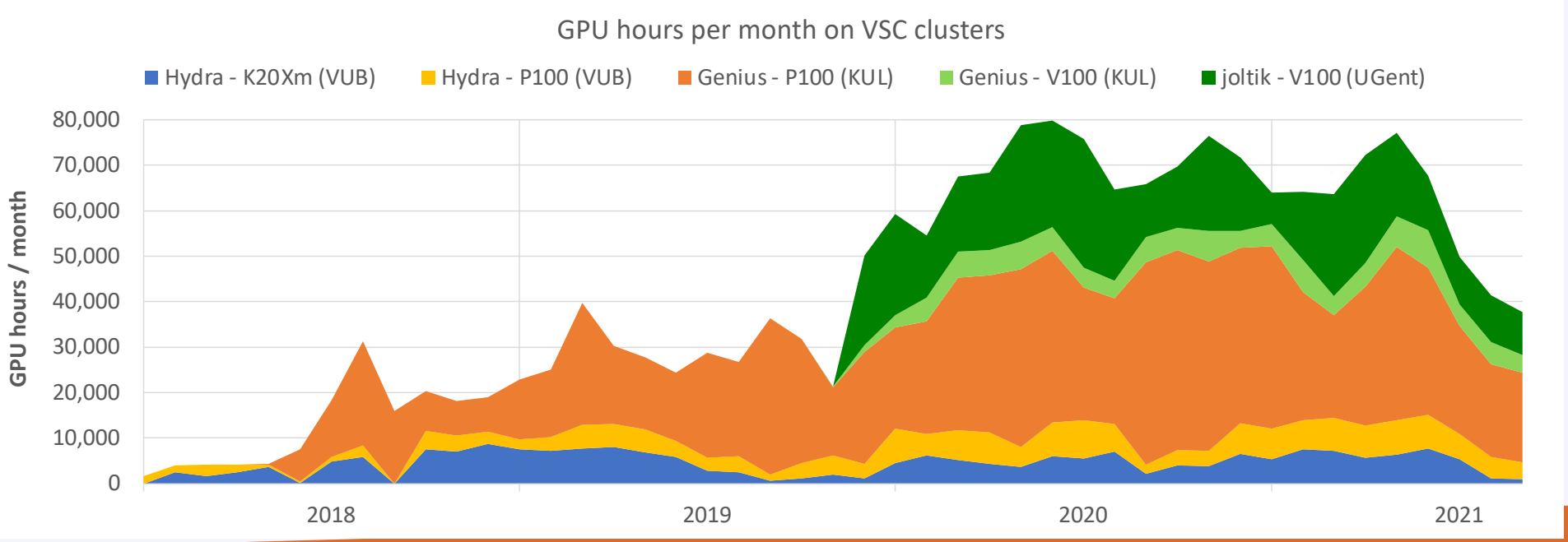
Currently available GPGPU resources at VSC

https://docs.vscentrum.be/en/latest/hardware.html

- **Hydra Tier-2 cluster at VUB** https://hpc.vub.be/docs/faq/basic/#how-can-i-use-gpus-in-my-jobs
  - **6 nodes** with **2x NVIDIA K20Xm GPUs** (2x 10-core Intel Ivy Bridge, 128GB RAM, **6GB GPU mem.**), QDR IB – since 2013
  - **4 nodes** with **2x NVIDIA P100 GPUs** (2x 12-core Intel Broadwell, 256GB RAM, **16GB GPU mem.**), 10Gbps – since 2017
- **Genius Tier-2 cluster at KUL** https://docs.vscentrum.be/en/latest/leuven/genius_quick_start.html#submit-to-genius-gpu-node
  - **20 nodes** with **4x NVIDIA P100 GPUs** (2x 18-core Intel Skylake, 192GB RAM, **16GB GPU mem.**), EDR IB – since 2018
  - **2 nodes** with **8x NVIDIA V100 GPUs** (2x 18-core Intel Cascade Lake, 768GB RAM, **32GB GPU mem.**), EDR IB – since 2019
- **Leibniz Tier-2 cluster at UAntwerpen** https://docs.vscentrum.be/en/latest/antwerp/gpu_computing_uantwerp.html
  - **2 nodes** with **2x NVIDIA P100 GPUs** (2x 18-core Intel Broadwell, 128GB RAM, **16GB GPU mem.**), EDR IB – since 2017
- **joltik Tier-2 cluster at UGent** https://www.ugent.be/hpc/en/support/documentation.htm (see Chapter 21)
  - **10 nodes** with **4x NVIDIA V100 GPUs** (2x 16-core Intel Cascade Lake, 256GB RAM, **32GB GPU mem.**), HDR-100 IB – since 2020

Usage of VSC GPU resources

- **AI (image analysis, deep learning, …): PyTorch, TensorFlow**
- Bioinformatics: Beagle, BEAST, **AlphaFold**
- Computational chemistry: GROMACS, VASP, OpenMM
- Custom code: C++ and CUDA, Python (scipy, **scikit-learn**, scikit-image, …), …

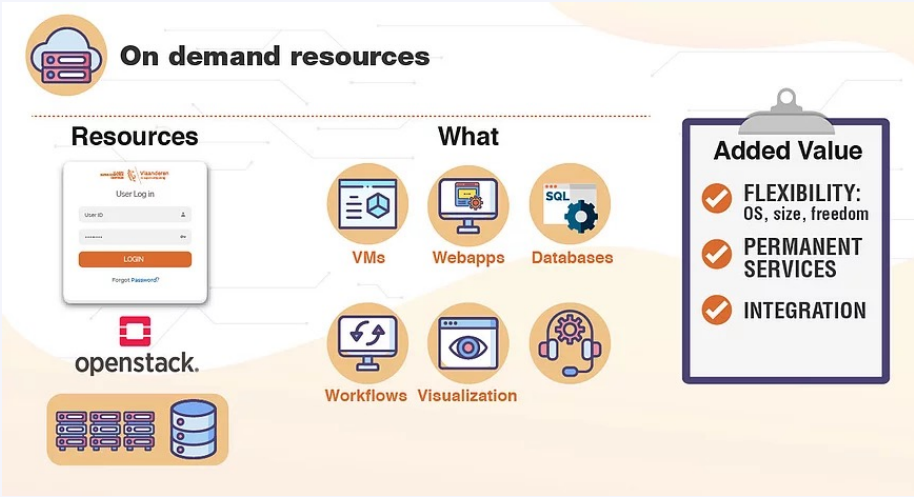


VLAAMS SUPERCOMPUTER CENTRUM

Vlaanderen is supercomputing

VSC Tier-1 Cloud (incl. limited GPU resources)

- **Self-managed virtual machines** (VMs): you install the software you need!
- Golden VM images available, can be modified, or use your own images
- Project based access (project proposal due by specific cut-off date)
- **Several GPU instance types, each with 1 virtual GPU (vGPU)**
  - **1 vGPU = 1/4th of an NVIDIA T4 GPU => 4GB GPU memory**
  - GPU driver + CUDA already installed in available GPU images
  - Can be used for small GPGPU compute workloads (like AI inference) or visualization (with some limitations)
  - Tweaked instance types can be considered (on request)
- VSC network access can be made available (on request)



Documentation: https://www.vscentrum.be/cloud

For more information: cloud@vscentrum.be

VLAAMS SUPERCOMPUTER CENTRUM

Vlaanderen is supercomputing

Coming soon at VSC: NVIDIA A100 GPUs

https://www.nvidia.com/en-us/data-center/a100

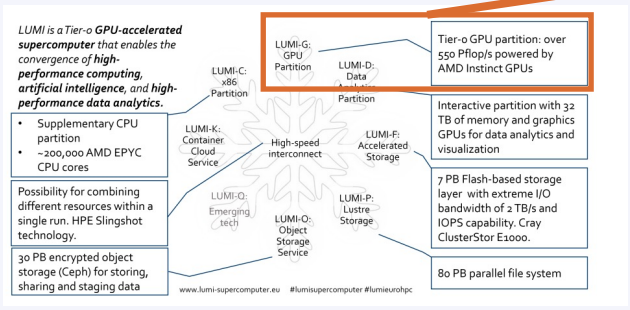
- **Tier-2 @ UGent: new `accelgor` cluster (currently in pilot)**
  - **9 nodes** with **4x NVIDIA A100 GPUs** (2x 24-core AMD Milan, 500GB RAM, **80GB GPU mem.**), HDR IB
- **Tier-2 @ VUB: additional GPGPU nodes in Hydra cluster (currently in pilot)**
  - **6 nodes** with **2x NVIDIA A100 GPUs** (2x 24-core AMD Rome, 256GB RAM, **40GB GPU mem.**), EDR IB
- **Tier-1 Hortense (at UGent): GPU partition (coming soon…)**
  - **20 nodes** with **4x NVIDIA A100 GPUs** (2x 24-core AMD Rome, 256GB RAM, **40GB GPU mem.**), dual HDR IB
- *Multi-Instance GPU (MIG)* support to allow running multiple jobs on a single GPU at once

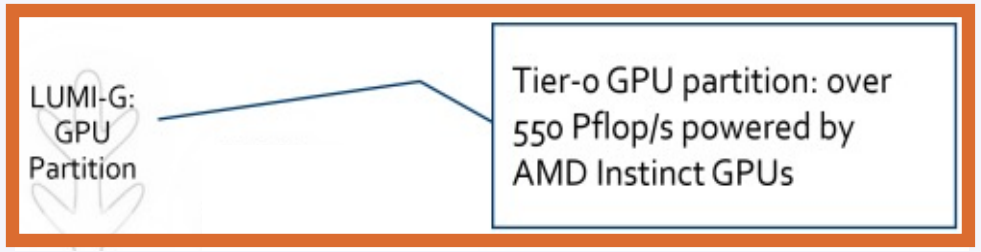
Beyond the VSC: GPU resources in LUMI

- LUMI-G: large GPU partition with next generation AMD Instinct GPUs
- **Will also be accessible for Belgian researchers through LUMI consortium!** more info here
- **Different software platform:** *AMD Radeon Open Compute platform (ROCm)*
  - **CUDA-only software is not supported!**
- Deep Learning software like TensorFlow and PyTorch already have ROCm support
  https://rocmdocs.amd.com/en/latest/Deep_learning/Deep-learning.html
- More detailed info (and specialized training sessions) soon…

LUMI

www.lumi-supercomputer.eu





https://www.lumi-supercomputer.eu/may-we-introduce-lumi

For more information:

lumi-be-support@enccb.be

https://www.enccb.be/LUMI

GPU software

Programming Models, Libraries, Applications for GPU platforms

Low-level (programming languages)

NVIDIA CUDA

(only for NVIDIA GPUs!)

OpenCL

HIP

OpenACC

OpenMP (offload)

Higher-level libraries

cuBLAS
cuFFT
cuML
cuGraph
cuDF
…

(only for NVIDIA GPUs!)

RAPIDS
https://rapids.ai

Numba

Kokkos

CuPy

C++ python

Deep Learning frameworks

TensorFlow

PyTorch

Keras

mxnet

MIOpen

Applications

AlphaFold

GROMACS

COSMO

Racon

BEAST

Guppy

GPU software installations on VSC compute infrastructure

In order of preference:

- **Central installations provided by VSC support teams (RECOMMENDED)**
  - Accessible via environment module system (example: `module avail TensorFlow`)
  - Usually **installed from source and optimized for specific cluster hardware** via easybuild
- **NVIDIA NCG containers** (https://developer.nvidia.com/ai-hpc-containers)
  - Can be used on VSC infrastructure via **Singularity** container tool
  - Alternative: container image provided by 3rd parties, for example on Docker Hub (be careful!)
- Manage your own software stack via **conda**
  - Some care should be taken here: polluting of your home directory, not compatible with env. modules, …
  - It's possible that performance is not optimal (software is often not tuned for specific hardware)

VLAAMS SUPERCOMPUTER CENTRUM

Vlaanderen is supercomputing

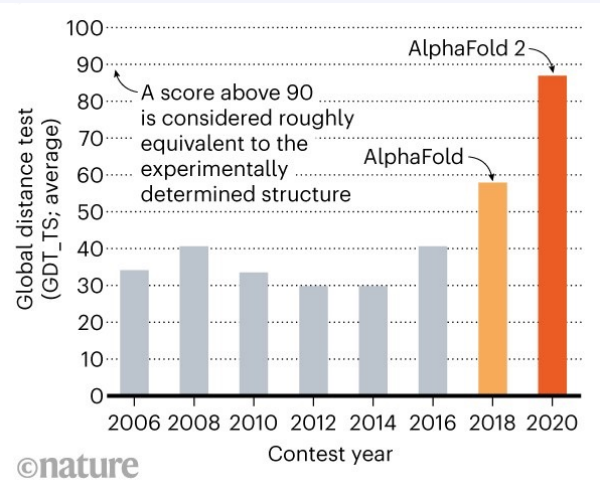
Practical guidelines for using GPUs

- **Check for significant performance speedup** on GPU for the software you are using (compare with CPU only)
- Try to make sure you **make maximal use of the GPU capabilities** (if not, consider using a different type of GPU)
- **Start with using a single GPU** (if possible) before evaluating additional speedup with multiple GPUs (if any…)
- **Try to avoid wasting GPU resources** for workloads where only some parts are running CPU-only
- **Take into account how portable your workload is to other types of GPUs** (CUDA vs HIP vs OpenCL vs …)
- Don’t hesitate to reach out for help: https://www.vscentrum.be/getintouch

EXAMPLE USE CASE: AlphaFold 2

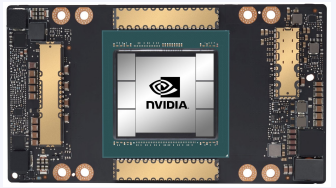
Accurately predicting protein structure with AlphaFold

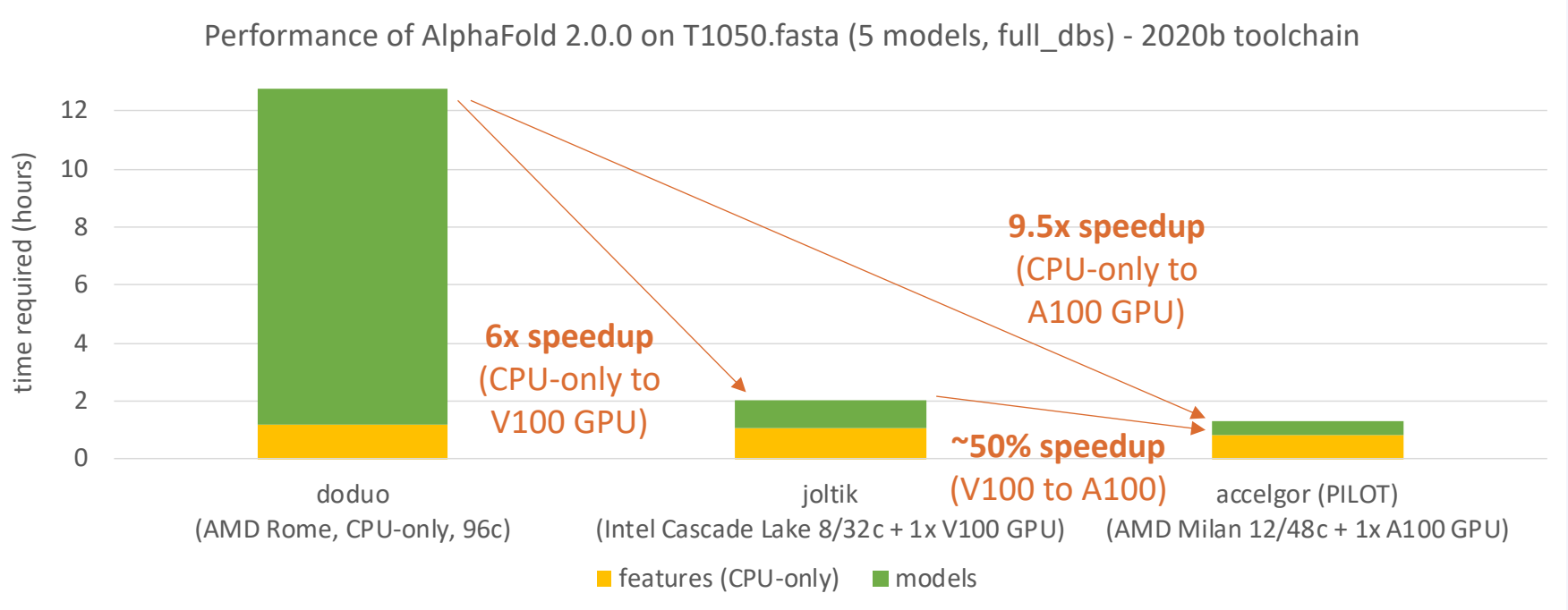
- AlphaFold 2 AI algorithm (by Google's DeepMind) **"solves a 50-year-old grand challenge in biology"**
- Clear winner of recent *Critical Assessment of protein Structure Prediction* (CASP14 - 2020)
- **Accurately predicts what 3D shapes proteins "fold" into**, based on 1D sequence of amino acids
- Relies on large dataset of examples (~2.5TB!) + feature selection and Deep Learning techniques
- **Replaces time-consuming experiments that used to take several weeks/months!**
- **Significant speedup by running AlphaFold on sufficiently capable GPUs**
- Applications: drug discovery, better understanding diseases, etc.



EXAMPLE USE CASE: AlphaFold 2

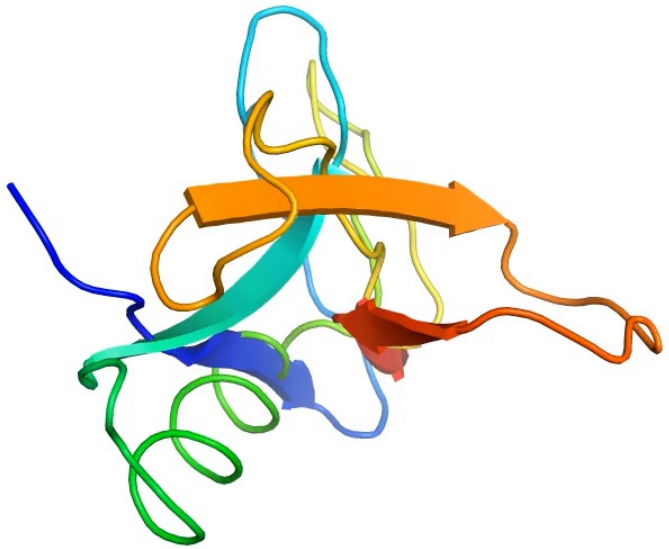
CPU vs GPU performance of AlphaFold





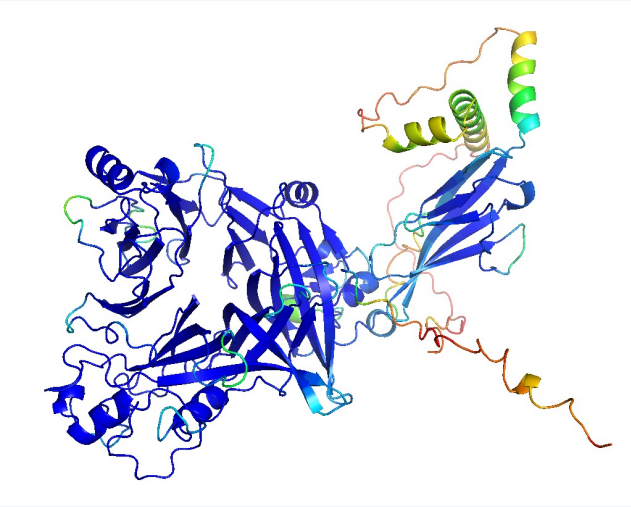
- Little performance tuning done, some room for potential improvement for both CPU-only and GPU runs
- Modest example input (sequence of 779 amino acids)
- More significant differences observed with longer input sequences, or using larger dataset (like CASP14)
- More GPU memory opens the door for more challenging (longer) input sequences

Iterative refinement of predicted structure + confidence level



Recycling iteration 0, block 01
Secondary structure assigned from the final prediction

AlphaFold iteratively refines prediction of folded protein



- Confidence level for prediction is also available
- Blue = very confident
- Yellow/green = unsure in various degrees
- Red = no idea…
- **Very valuable for domain scientists to interpret result!**

More information on AlphaFold

- Protein folding and AlphaFold explained in 2 minutes: https://youtu.be/KpedmJdrTpY
- AlphaFold: The making of a scientific breakthrough: https://youtu.be/gg7WjuFs8F4
- Blog posts:
  - https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle
  - https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology
  - https://deepmind.com/blog/article/putting-the-power-of-alphafold-into-the-worlds-hands
- Paper in Nature: https://www.nature.com/articles/s41586-021-03819-2
- Source code on GitHub: https://github.com/deepmind/alphafold
- Prediction of multi-chain protein complexes with AlphaFold-Multimer: https://deepmind.com/research/publications/2021/protein-complex-prediction-with-alphafold-multimer

Acknowledgements

Thanks to:

- Various VSC colleagues for their input:
  - **Alex Domingo (VUB)**
  - **Álvaro Simón García (UGent)**
  - **Balázs Hajgató (UGent)**
  - **Tim Jaenen (FWO)**
  - **Kurt Lust (UAntwerpen)**
  - **Jan Ooghe (KU Leuven)**
  - **Ewald Pauwels (UGent)**
  - **Ward Poelmans (VUB)**
- **Jasper Zuallaert (VIB-UGent)** for his feedback regarding the AlphaFold use case

More information on VSC and GPUs

- Vlaams Supercomputing Centrum (VSC)
  - Website: https://www.vscentrum.be
  - Available hardware resources: https://docs.vscentrum.be/en/latest/hardware.html
- NVIDIA GPUs and Deep Learning:
  - https://developer.nvidia.com/hpc
  - Deep Learning Institute: https://www.nvidia.com/en-gb/training
  - NVIDIA A100: https://www.nvidia.com/en-us/data-center/a100
  - NVIDIA A100 in depth: https://developer.nvidia.com/blog/nvidia-ampere-architecture-in-depth
- Tutorials on GPU programming: https://www.gpuhackathons.org/index.php/technical-resources