LIGATE

Speaker name:
Luigi Crisci

IWOCL & SYCLcon 2022

Towards a Portable Drug Discovery Pipeline with SYCL 2020

Luigi Crisci, Majid Salimi Beni, Biagio Cosenza – University of Salerno
Nicolò Scipione, Davide Gadioli, Emanuele Vitali, Gianluca Palermo – Politecnico di Milano
Andrea Rosario Beccari – Dompé Farmaceutici S.p.A.

EuroHPC Joint Undertaking

European Commission

This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 956137. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Italy, Sweden, Austria, Czech Republic, Switzerland.

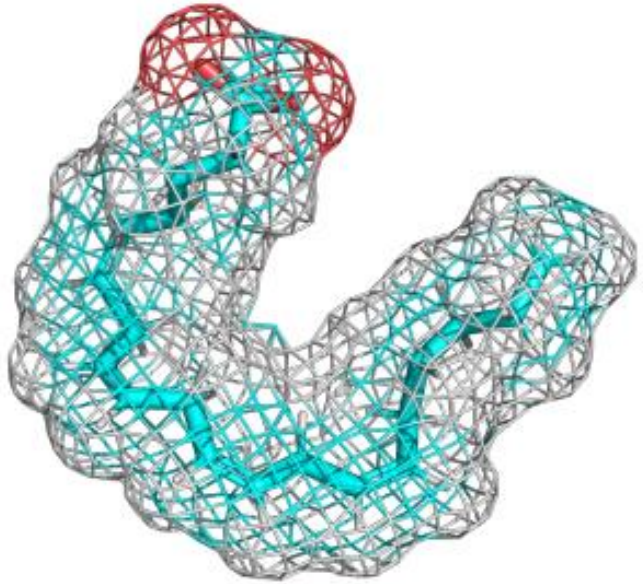
LIGATE

Outline

- Introduction to LiGen
  - Virtual screening
  - The LiGen drug discovery processing pipeline
- Towards a portable pipeline
  - Porting from CUDA to SYCL 2020
  - Exploiting SYCL 2020 features
- Experimental evaluation
  - Application performance
  - USM vs accessors
  - Performance portability (NVIDIA V100, AMD M100, Intel Xeon)
- Conclusion

Virtual Screening

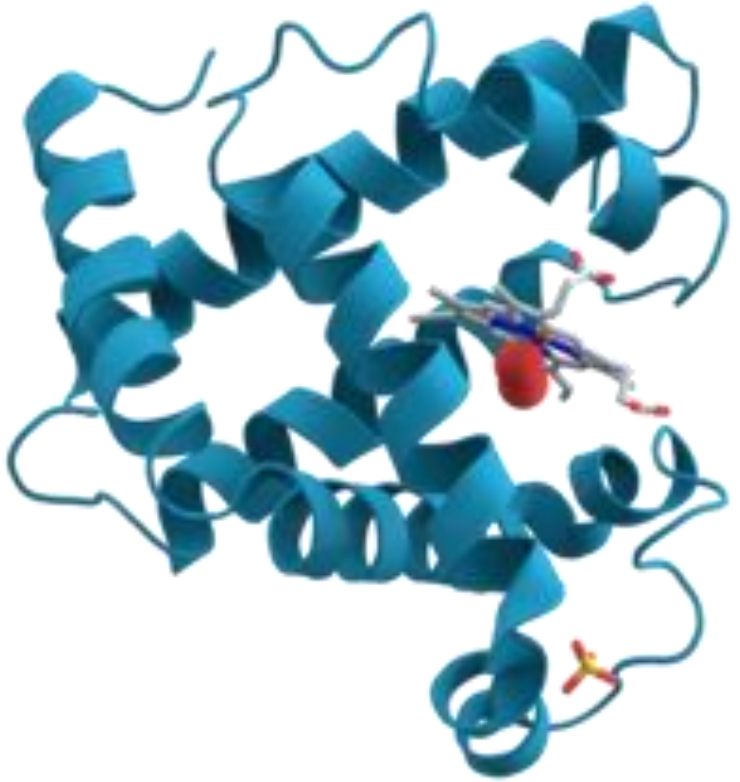
- Initial stages of the drug discovery process
- Ligand: small molecule with usually less than 100 atoms
- Can be view as a graph
  - Atoms as vertices
  - Bond as edges
- Can be generated by chemical reactions
- **Drug Candidate**



Arachidonic Acid ligand example

Virtual Screening

- A protein is a molecule that is composed of one or more chains of amino acid residues
  - Tens of thousand of atoms
- Proteins perform a vast array of functions within organisms:
  - Catalysing metabolic reactions,
  - DNA replication,
  - Providing structure to cells and organisms
  - Transporting molecules from one location to another.
- **Drug's target**

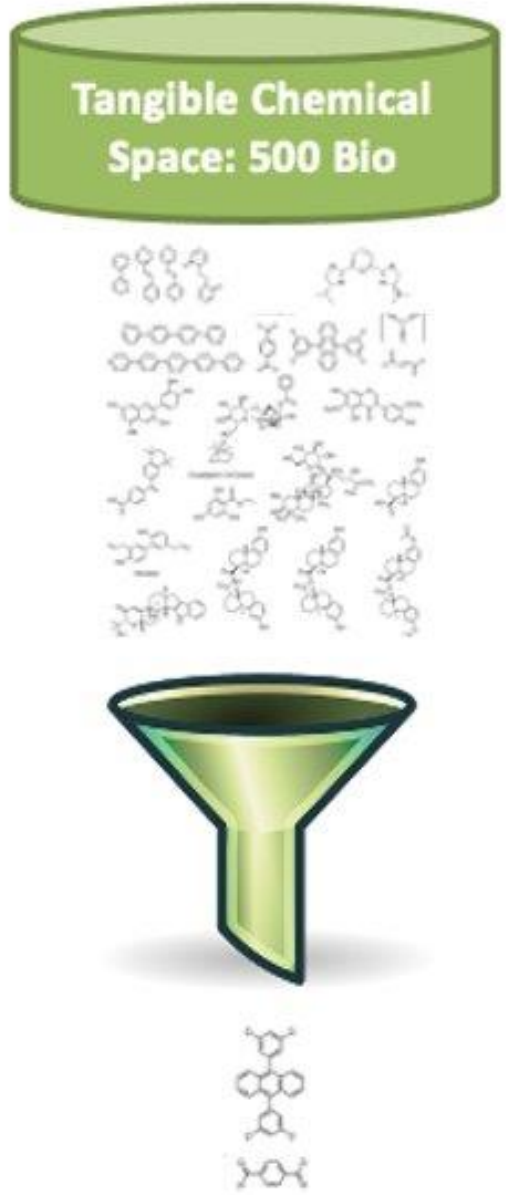


A representation of the 3D structure of the protein myoglobin

LIGATE

Virtual Screening

- A ligand bound into empty spaces of a protein associated with the target disease can change its behaviour, and hence the outcome of disease.
- Virtual screening aims at selecting the most promising ligands from a huge set of possible candidates
  - To forward to the next stages of drug discovery

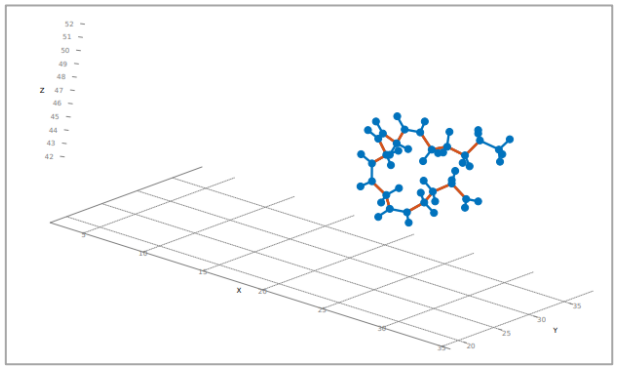


LIGATE

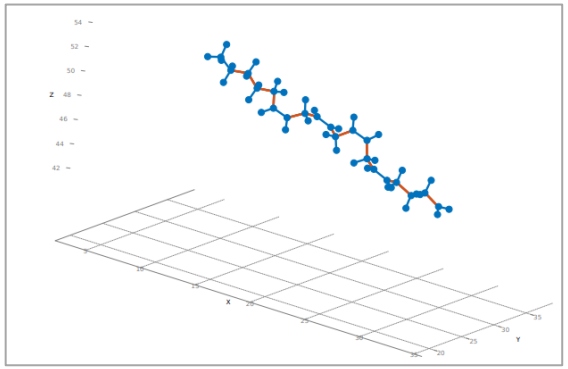
Virtual Screening

Challenges:

- Evaluate as much ligand candidates as possible
- The bond strength depends on the ligand's atoms 3D displacement when it in teracts the with target protein
    - We need to "dock" before "score" it
- A ligand can be change its shape
    - Multiple way to dock the same ligand

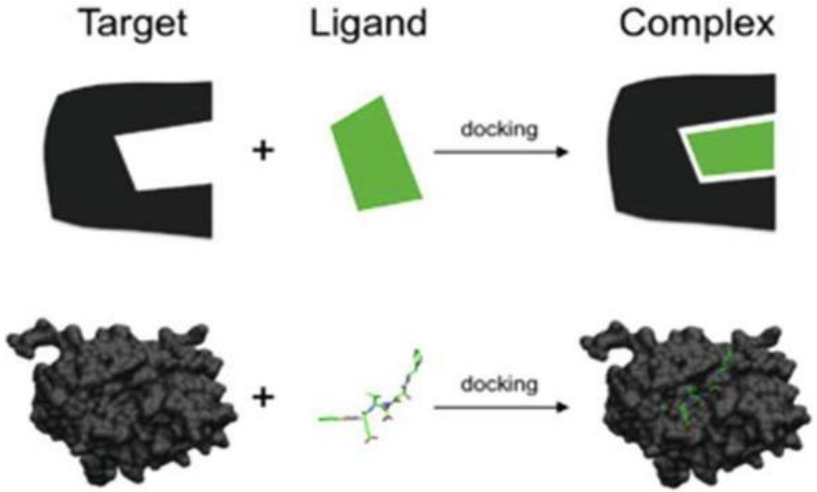


Reference



Unfolded

Two conformations of the Arachidonic Acid.

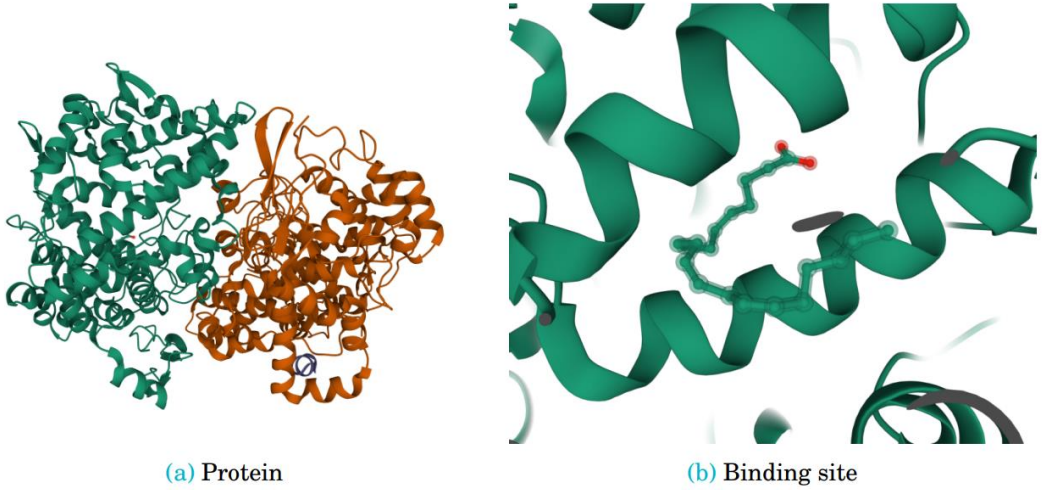


LIGATE

Dock & Score

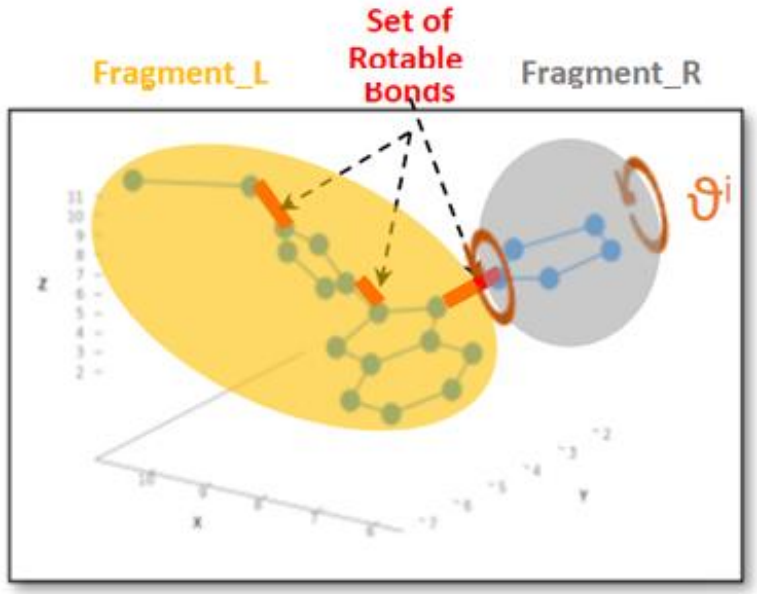
Molecular Docking as "Relaxed" lock and key model:

- Protein considered a rigid body
- Ligand considered a flexible body
- A ligand is docked on target protein's binding sites
- Multiple pose for each ligand

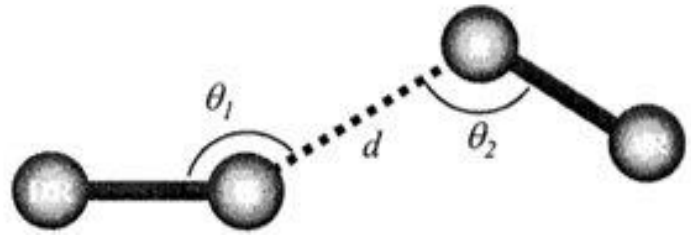


(a) Protein (b) Binding site

Arachidonic Acid bound to
the Cyclooxygenase active site of COX-2



$\vartheta=[\vartheta^1,\ldots,\vartheta^N]$

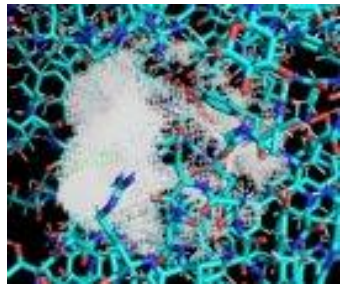


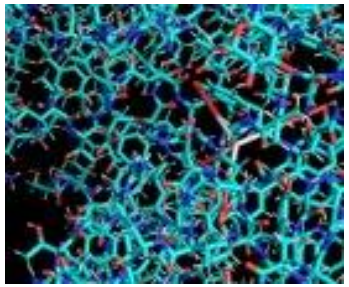
LIGATE

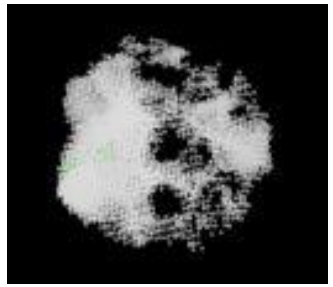
Dock & Score

Additive Chemical score

- Used to evaluate likeness of ligand (pose)-pocket bond
- Takes into account Atomic interactions
    - VdW,
    - Lennard-Jones
    - Metal bonds
    - Hydrogen bonds
    - Solvation effects;







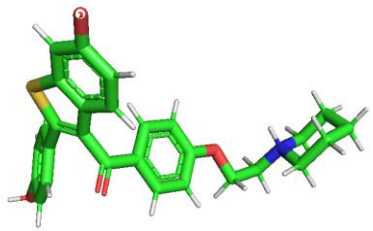
LiGen: Ligand Generator Platform

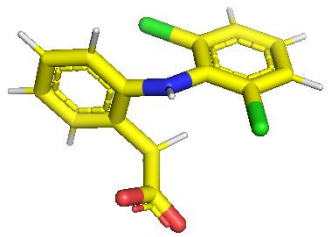
The LiGen drug discovery framework

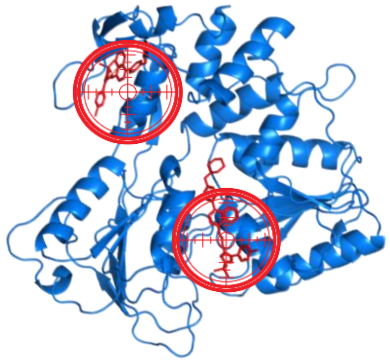
- Owned by Dompe Farmaceutici and co-developed by PoliMi and CINECA
- used for the discovery of drug against different pathogens
    - Zika [ANTAREX], SARS-CoV-2 [EXSCALATE4COV]

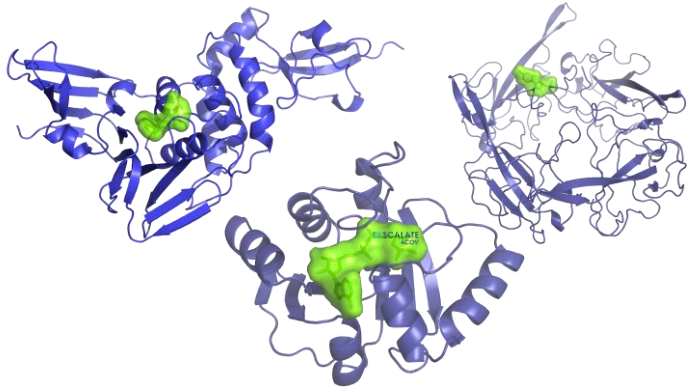
LiGen processing pipeline

- Dock & score as an embarrassing parallel task
- Process each molecule in a pocket, which is a region of interest inside a protein.
- A protein can have multiple pockets and they may have different conformational states, abstracted as different pockets.



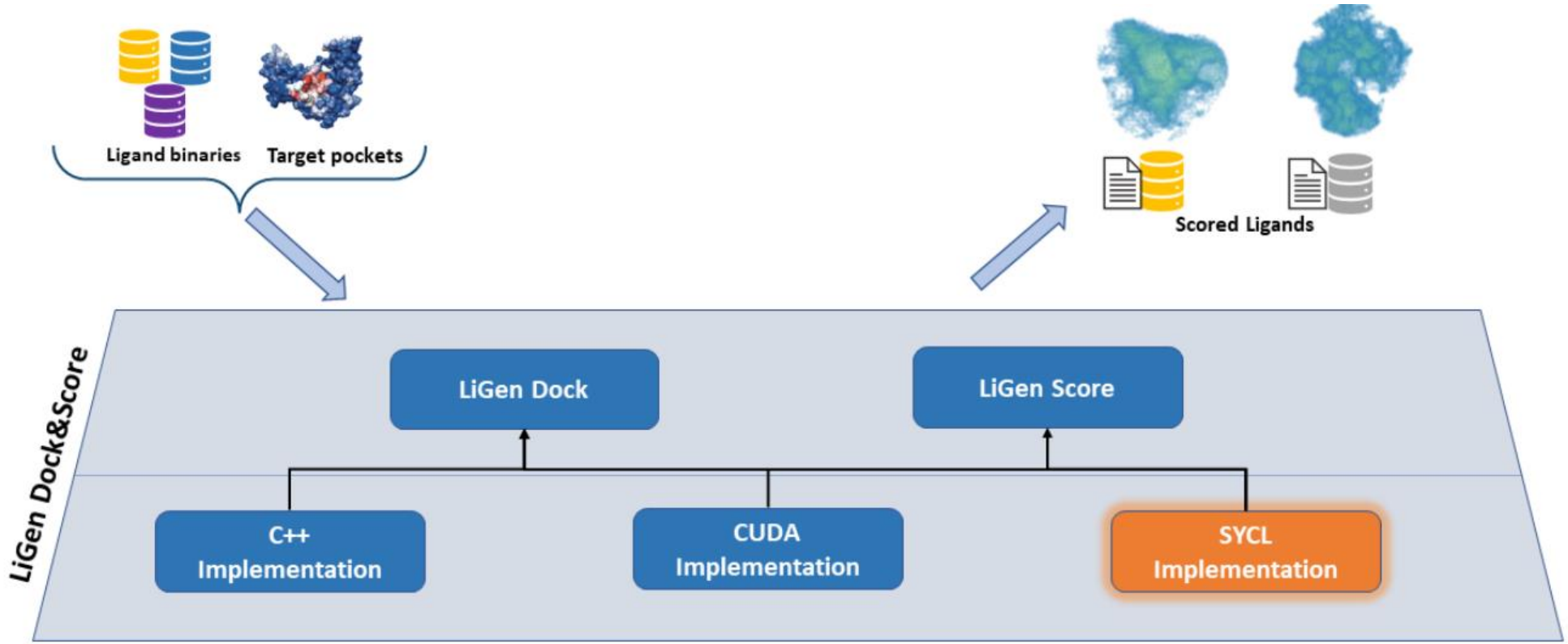






LIGATE

The LiGen Drug Discovery Pipeline



LIGATE

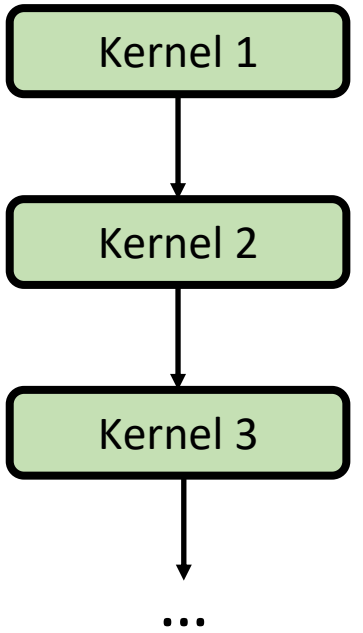
Towards a Portable Pipeline

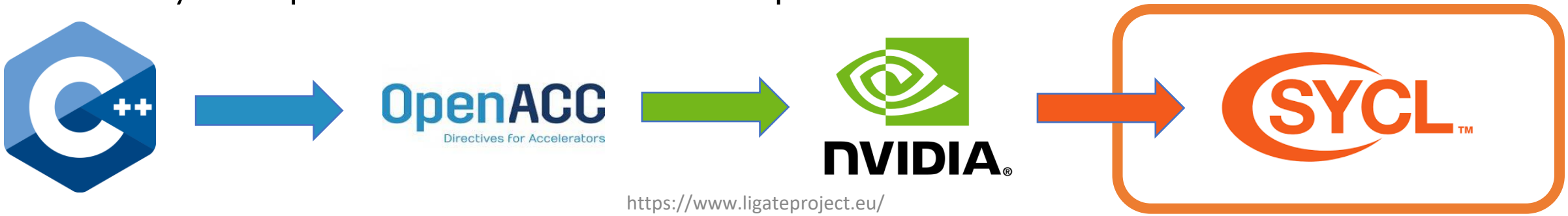
- The LIGATE project aims at building a portable drug discovery pipeline
  - Funded from the European High-Performance Computing Joint Undertaking Joint Undertaking (JU)
- HPC is heading toward specialization and extreme heterogeneity
- SYCL enables to write platform independent code, while keeping native-comparable performance

LIGATE

Porting from CUDA to SYCL2020

- Originally implemented in C++, then ported to OpenACC and CUDA
- 17 CUDA kernels, dozens of device functions
  - Most kernels have dependency from the previous one
  - Minimal overlapping
- Porting features from CUDA
  - Warp-level primitives -> SYCL 2020 subgroups
  - Custom reductions -> SYCL 2020 reductions
  - Custom group algorithms (all_of, any_of,etc.) -> SYCL2020 group algorithms
  - Dynamic parallelism -> removed without performance difference





Exploiting SYCL 2020 Features

- Supporting both Unified Shared Memory (USM) and accessor-based memory management
- Sub-group shuffles and collectives allow us to write more efficient and concise code
- Group algorithms raise efficiency and improve performance portability among architectures
- Reduction native support avoid boilerplate code and improve performance

*By using **SYCL 2020 reduction and group algorithms we removed more than 430 lines of code***

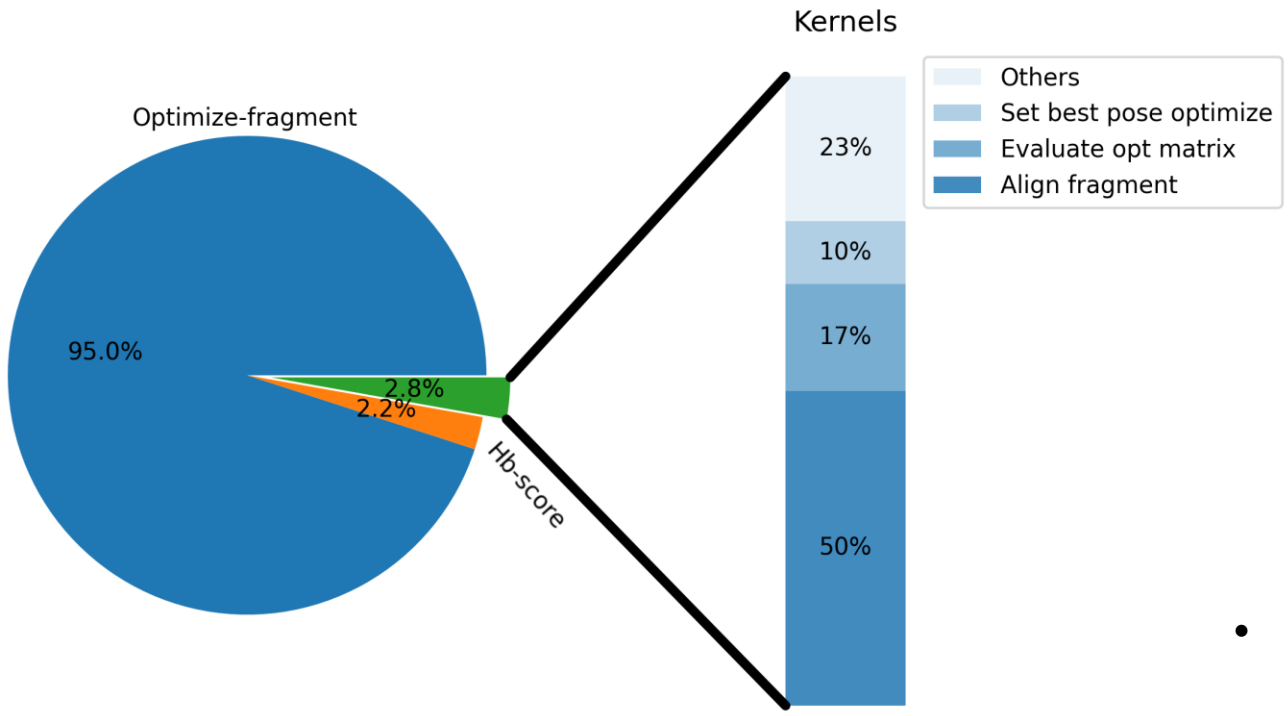
Performance Tuning

- Application tuning
  - Block size and number of iteration of the heaviest kernels depending on workload type
  - Up to the scheduler
- Kernel tuning
  - Limited tuning due to high register pressure (e.g., unrolling not effective)
  - Workgroup size
  - Blocking, local memory usage
- Accessors vs USM
  - We developed two version of LiGen, using either Accessor or USM

Experimental Setup

- SYCL versions
  - hipSYCL (sha 1046a78777a23ca75c6ea4e92291f1dbe36169ca)
  - DPC++ (release 2021-12)
- Hardware
  - Nvidia Setup
    - IBM POWER9 AC922 at 2.6(3.1) GHz
    - NVIDIA Volta V100 GPU
    - Cuda 11.0
  - AMD Setup
    - Intel Xeon Gold 5218 CPU @ 2.30GHz
    - AMD MI100 GPU

Application Performance Evaluation



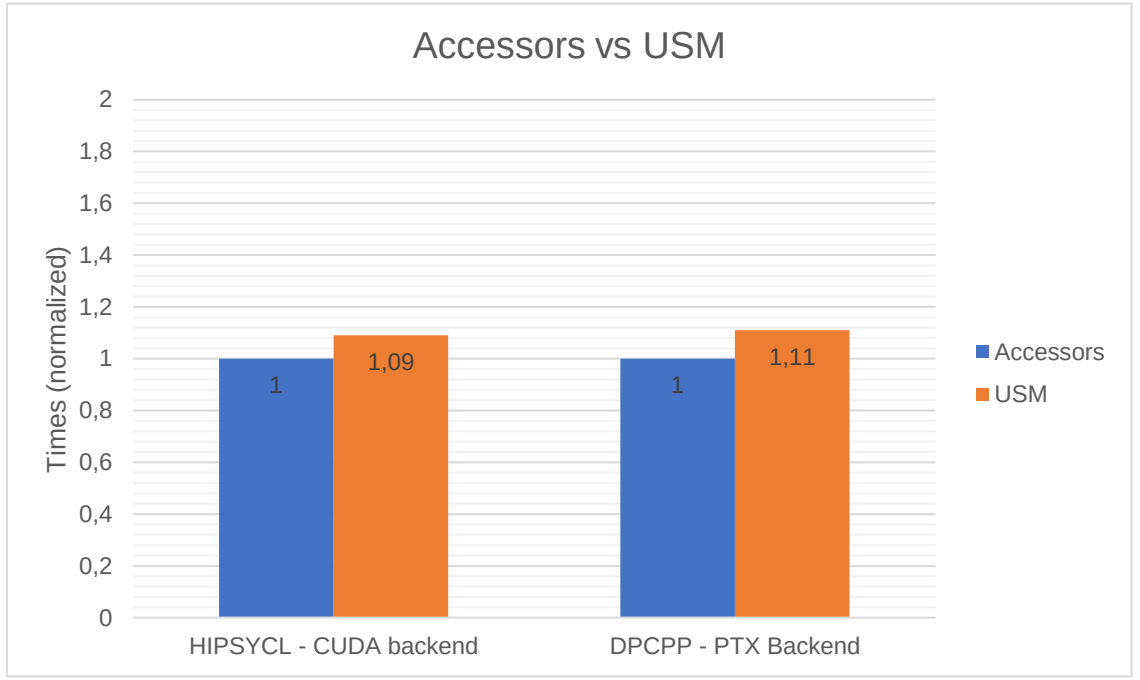
Runtime per Kernel

- Optimize-fragment kernel
  - Takes up to 95% of the device time
    - Called iteratively
    - Subgroup reduction, then reduction over group, then reduction across groups
    - Subgroup shifting and reductions
    - About 420 lines of code
    - High register pressure
- Hb-score kernel
  - takes 2.2% of the device time
    - Mostly due to group sorting

LIGATE

Accessor vs USM Evaluation

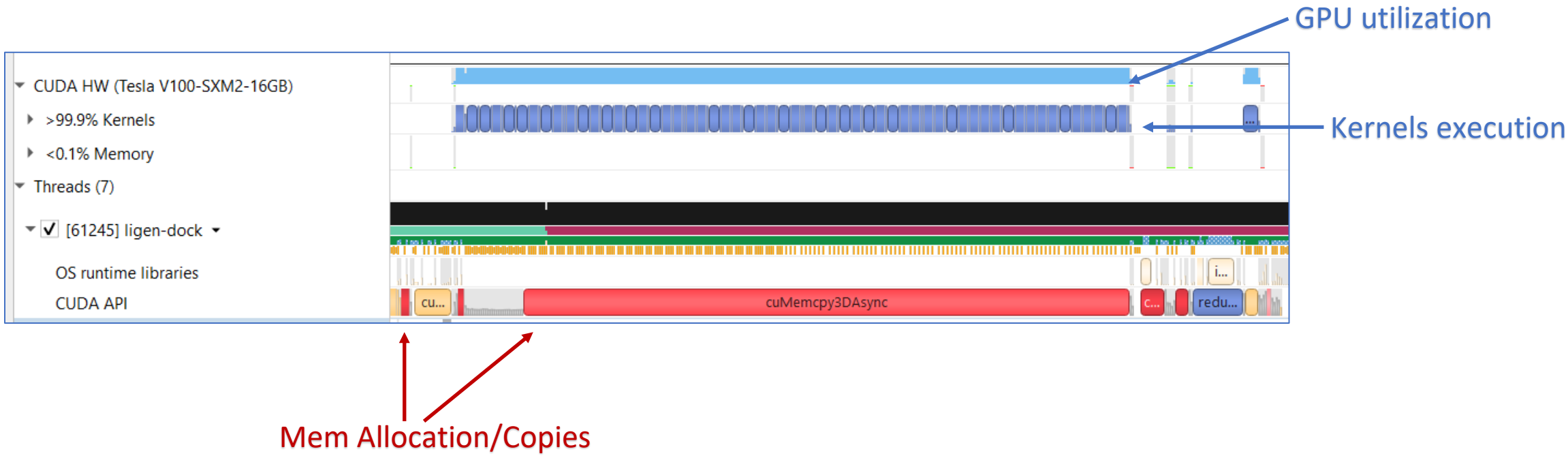
- USM version
  - memory management is similar to the CUDA implementation
  - Buffers allocated and initialized at the start
  - Deallocated all together at the end
- Accessor
  - defines each buffer at the start
  - Initialization is managed by the runtime
- No significant differences in the kernel implementations
- Accessor version similar performance compared to the USM version



LIGATE

Accessors vs USM Evaluation

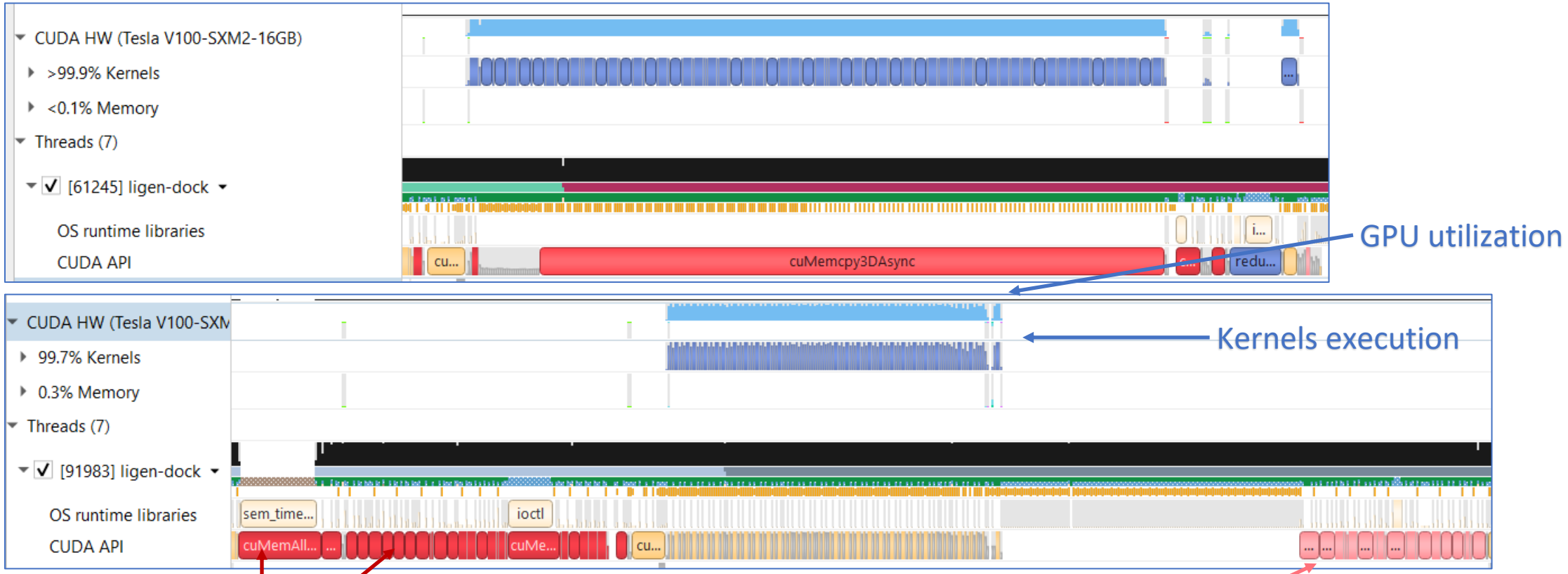
- Example with DPC++ / PTX backend, on a Tesla V100
- Accessor version (top): memory allocation and free happening across application workflow



LIGATE

Accessors vs USM Evaluation

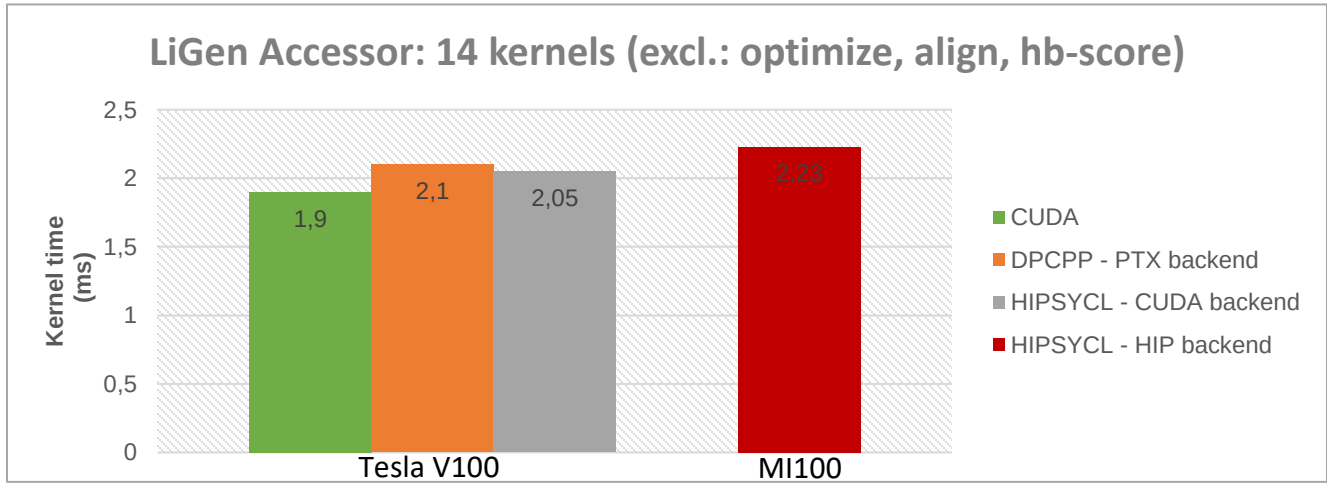
- Example with DPC++ / PTX backend, on a Tesla V100
- Accessor version (top): memory allocation and free happening across application workflow
- USM version (bottom): allocation at start, deallocation at end



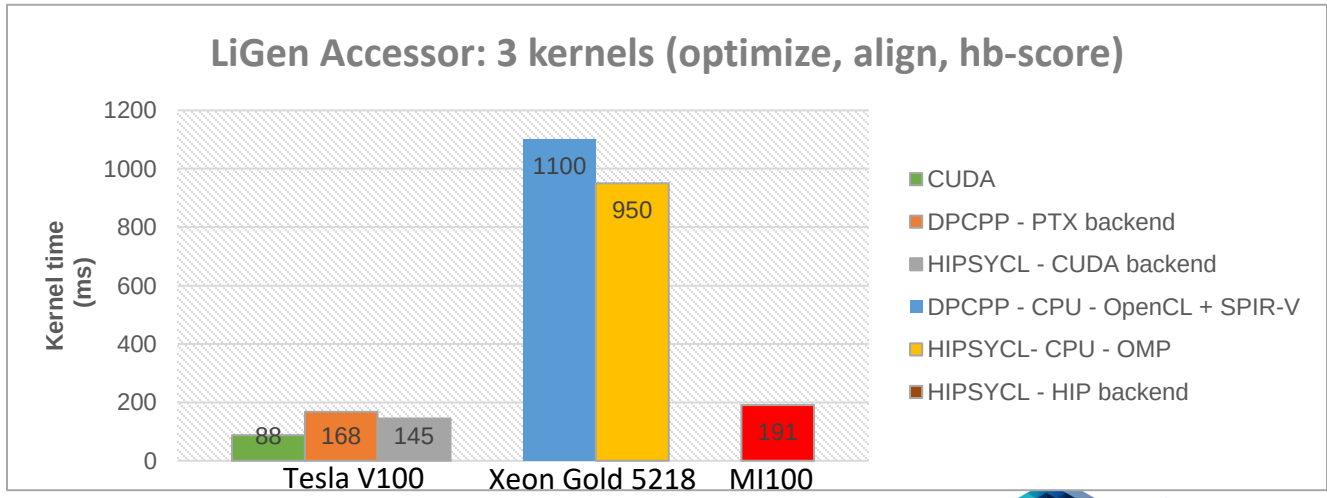
LIGATE

Performance Portability Evaluation

- 14 kernels show **excellent performance portability**
  - On NVIDIA V100: portable SYCL version vs manually optimized CUDA **only 7% faster**
  - High performance on both V100 and M100



- Three kernels have **lower performance portability**
  - Optimize and align kernel
    - high register pressure
  - Hb-score kernel
    - sorting kernel



LIGATE

Conclusion

- LiGen 3.0 relies on SYCL to provide performance portability
- All CUDA features ported to SYCL
  - successfully runs on a wide range of hardware
- SYCL 2020 features reduced LiGen code complexity
  - reduction and group algorithms allowed us to remove more than 430 lines of code
  - accessors and USM version, very similar in performance
- Performance portability study
  - excellent performance portability on 14 out of 17 kernels
    - manually-tuned CUDA only 7% faster than portable SYCL on V100
  - issues with three kernels because of
    - high register pressure (new Ligen 4.0 pipeline will attack this problem)
    - customized sorting function

LIGATE

www.ligateproject.com

Dompé
POLITECNICO MILANO 1863
CINECA
KTH VETENSKAP OCH KONST
CHELONIA Applied Science
tofmotion depth sensing and imaging
UNIVERSITÀ DEGLI STUDI DI SALERNO
E4 COMPUTER ENGINEERING
universität innsbruck
VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER
University of Basel

EuroHPC Joint Undertaking

European Commission

This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 956137. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Italy, Sweden, Austria, Czech Republic, Switzerland.

LIGATE