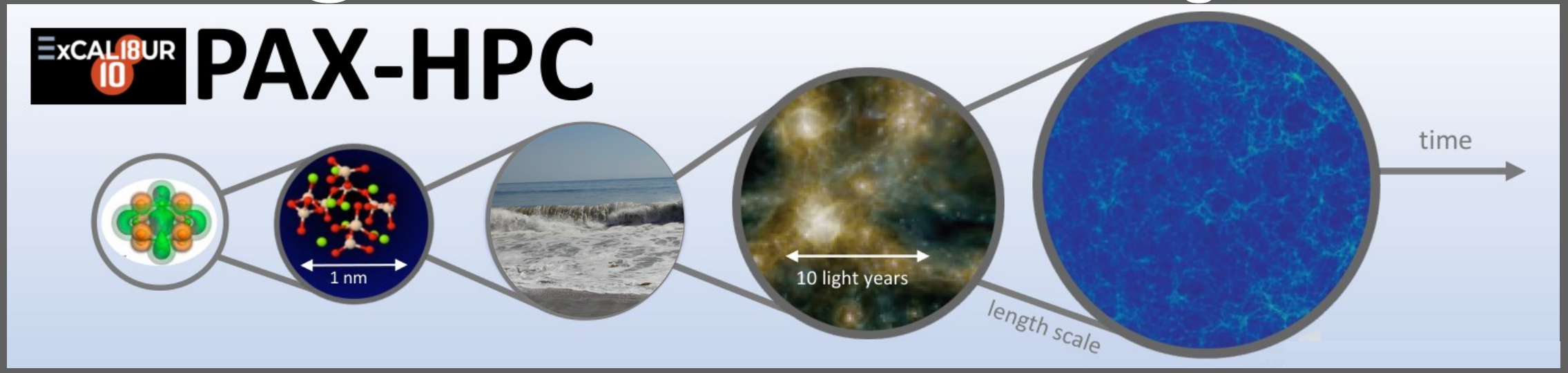
PAX: Modelling particles on GPUs

Abouzied Nasar
University of Manchester

Phil Hasnip
University of York

Computing Insight UK 2023, 7-8th December 2023

Particles @ Exascale: From atoms to galaxies



Particle-based methods in Science & Engineering

- Very wide range of applications; Material modelling, Engineering/CFD, Galaxy formation
- Two applications covered in this talk:

1- Astrophysics: SWIFT for galaxy formation and planetary collision ➔ GPU acceleration whilst maintaining scalability

2- Quantum mechanics: CASTEP for material properties from first principles using DFT ➔ GPU acceleration covered by Phil

```
 ______      ________________
/ ___/ |    / /  _/ ___/_  __/
\__ \| | /| / // // /_    / /
___/ /| |/ |/ // // __/   / /
/____/ |__/|__/___/_/    /_/
```

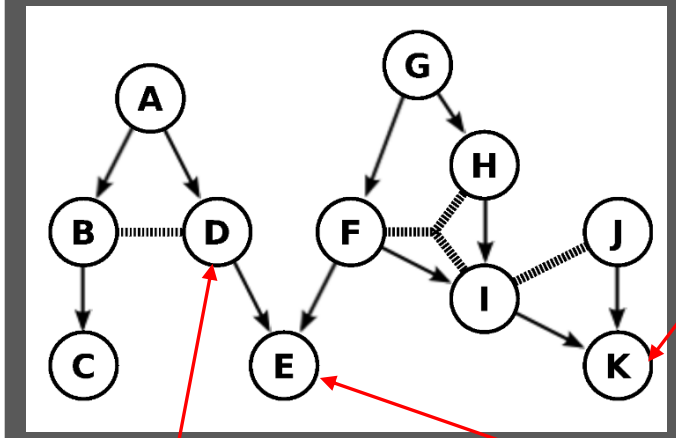
SPH With Inter-dependent Fine-grained Tasking

Jacob Kegerreis 2022 et al. 2022
Part size 14km. 128 million parts

Task-based parallelism in SWIFT & the Challenge of GPU offloading

Task-Based parallelization

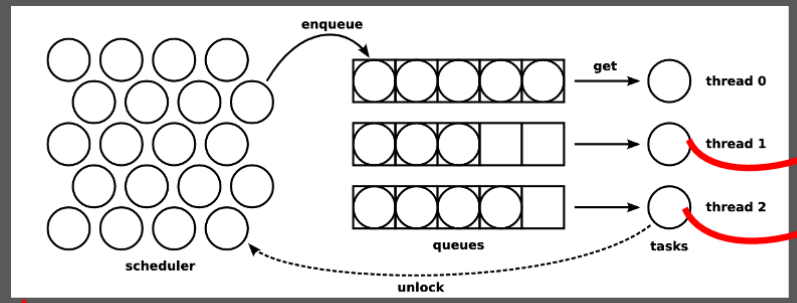
Operations (tasks) can belong to different processes but all execute concurrently (Gonnet et al. 2016)



e.g. Send comms data from cell *j*

e.g. evaluate forces for cell *i*

e.g. Update positions for cell *i*



Thread 0 has packed n_{target} tasks

Each CPU thread packs own data then carries on with next task

GPU offloading strategy

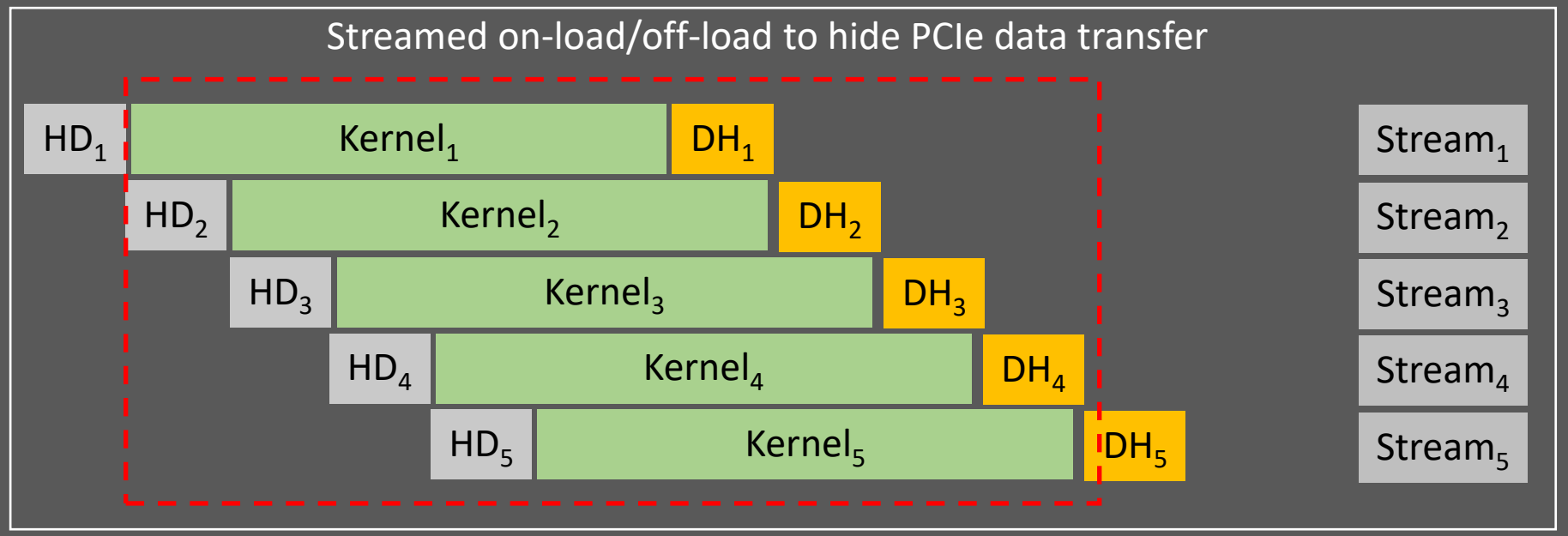
Tasks copied to GPU in bundles

Task results copied to CPU in bundles (as and when ready)

Thread 0 unpacks data and unlocks dep.s as soon as it has got a bundle back

GPU executes bundled tasks as soon as it has the data **and** has enough resources (cores, shared memory)

Leveraging GPU streams/concurrency ➔ More work needed



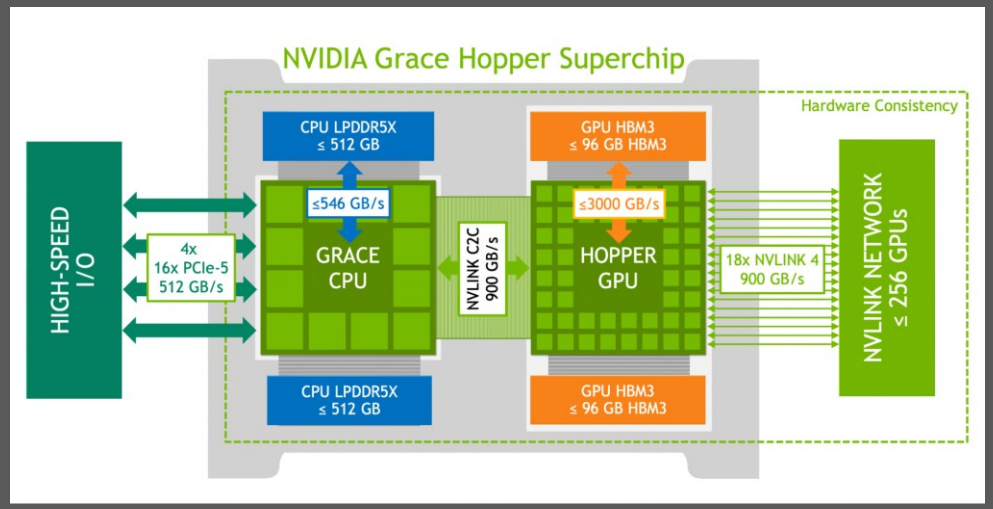
- Current speedup ≈100x (inc. packing/unpacking on host) on RTX3090 for some SWIFT task types and slower for others
- Issues to work on:

- CPU/GPU comms taking longer than kernels over PCIe 3 (achieved throughput ≈ 11GB/s) Theoretical throughput = 16 GB/s
- Will test on A100 testbed with PCIe4 (2x faster) ➔ PCIe5 is 4x faster! **Anyone got H100s we can test on?**

Next steps

- Validate achievable speedups on prev. gen A100s with 16 lane PCIe4 or AMD MI200
- Test on upcoming superchips (e.g. Grace-Hopper) with chip-chip interconnect 7x faster than PCIe5!
- Test on multi-node multi-GPU test beds ➔ likely to be A100 testbed
- Figure out how close we are to Exascale. Aaaand hope we haven't broken SWIFT!

Schematic of Grace-Hopper superchip (Nvidia)



Weak scaling of original SWIFT CPU code (Schaller et al. 2023)