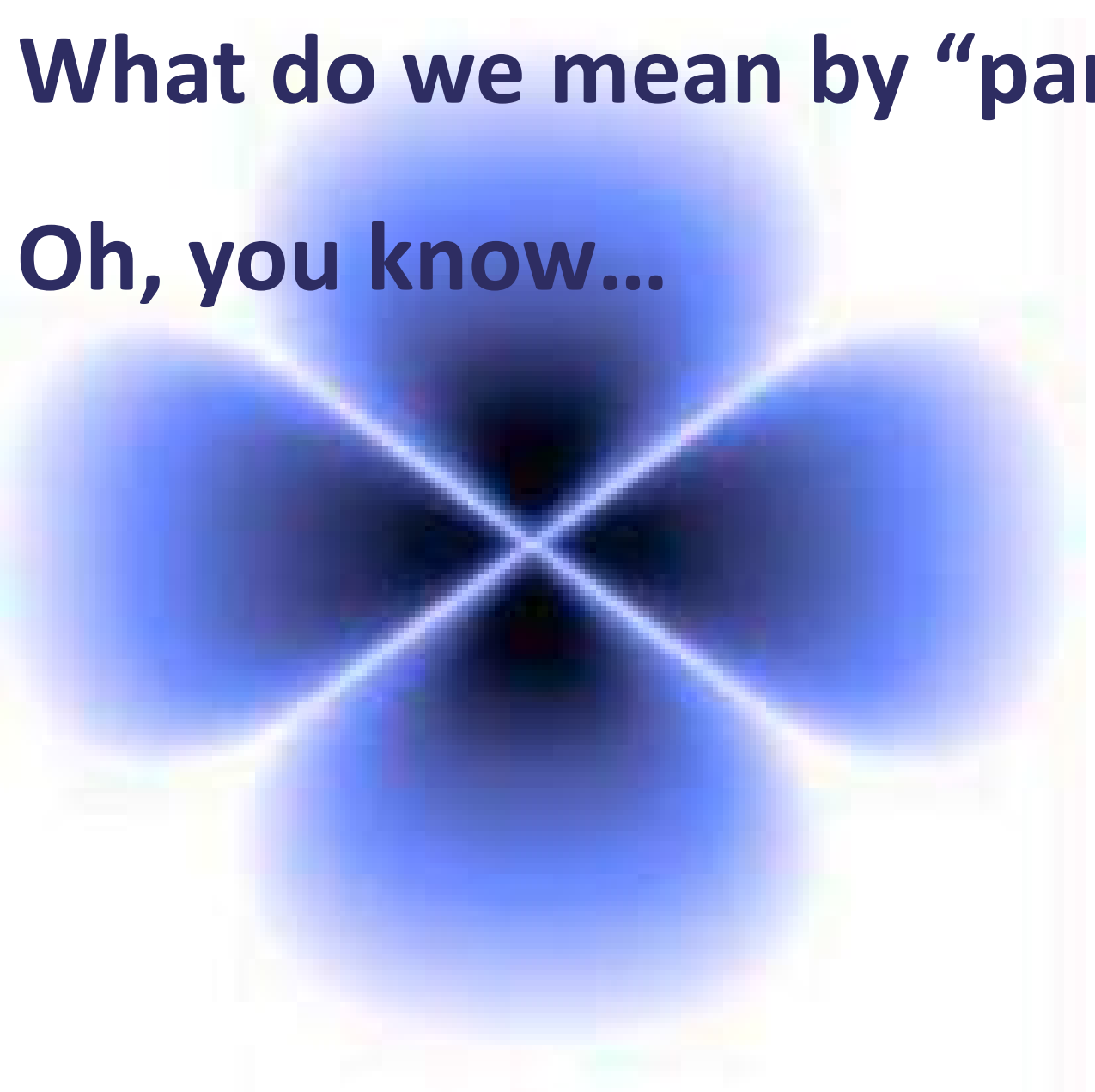
Particles At the eXascale on HPC

Phil Hasnip

University of York

What do we mean by “particles” ?

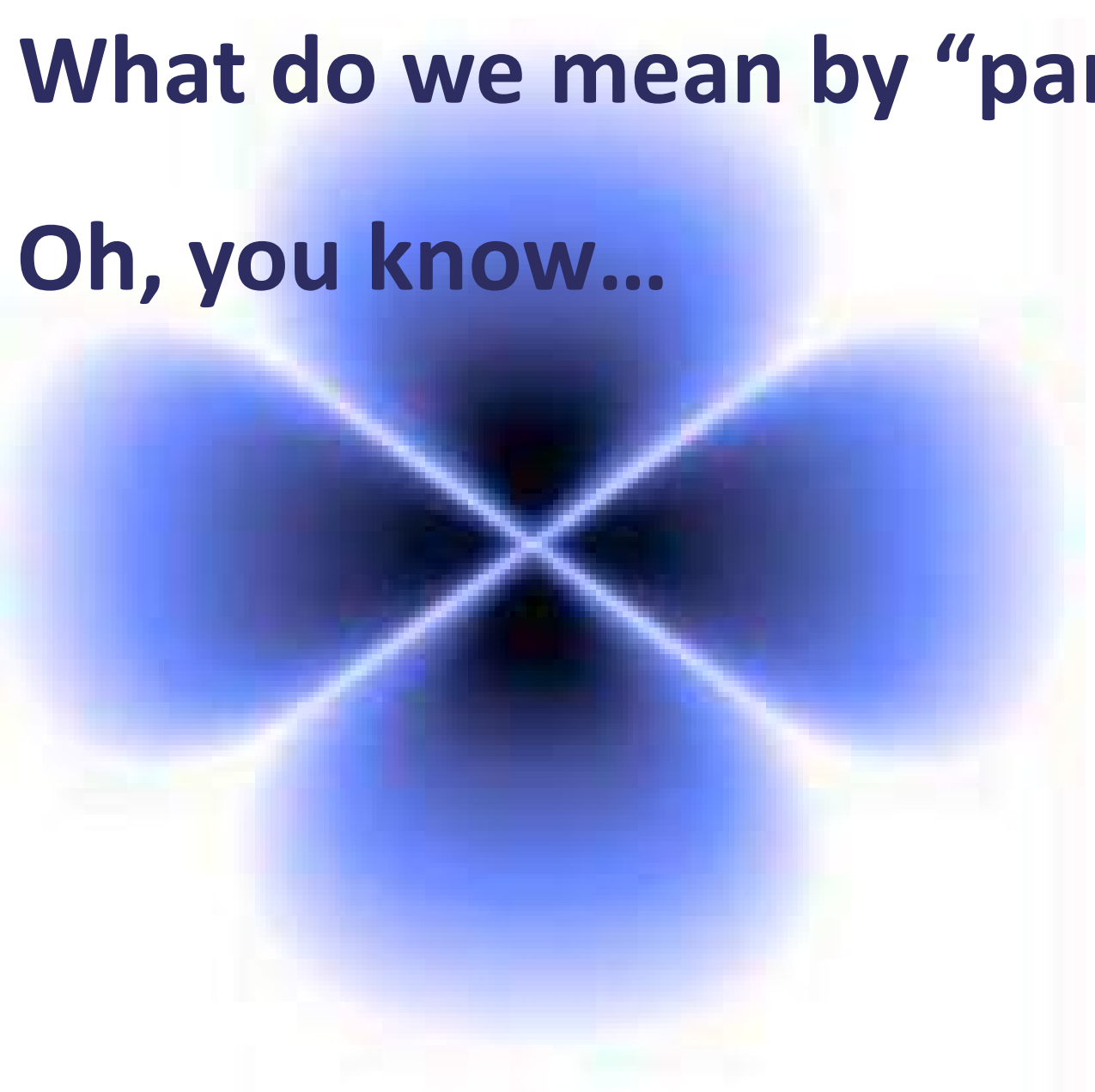
Oh, you know…



electrons…

What do we mean by “particles” ?

Oh, you know…



electrons…

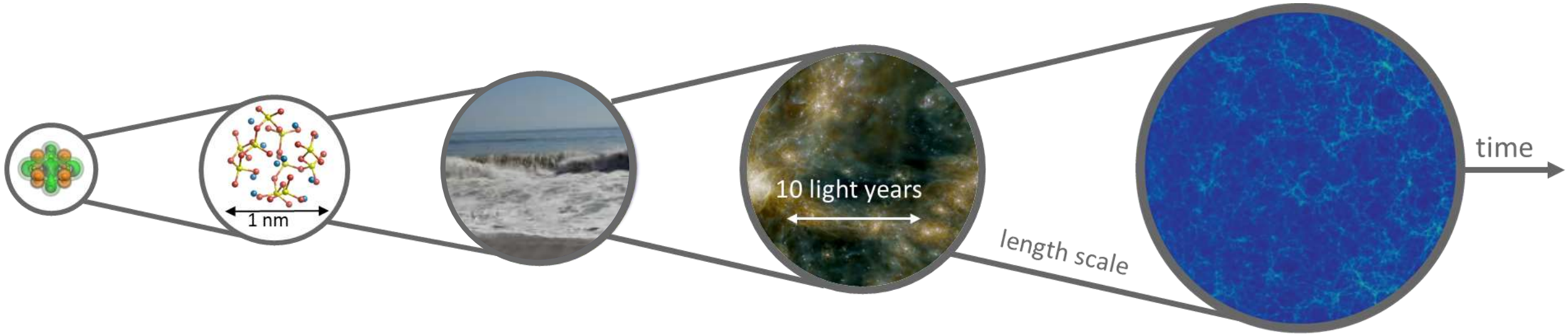
What do we mean by “particles” ?

Oh, you know…

stars…

that sort of thing

From atoms to galaxies



- Over 25 orders of magnitude difference!
- 11 RSEs, 30 academics, 14 project partners
- More than 12 software packages

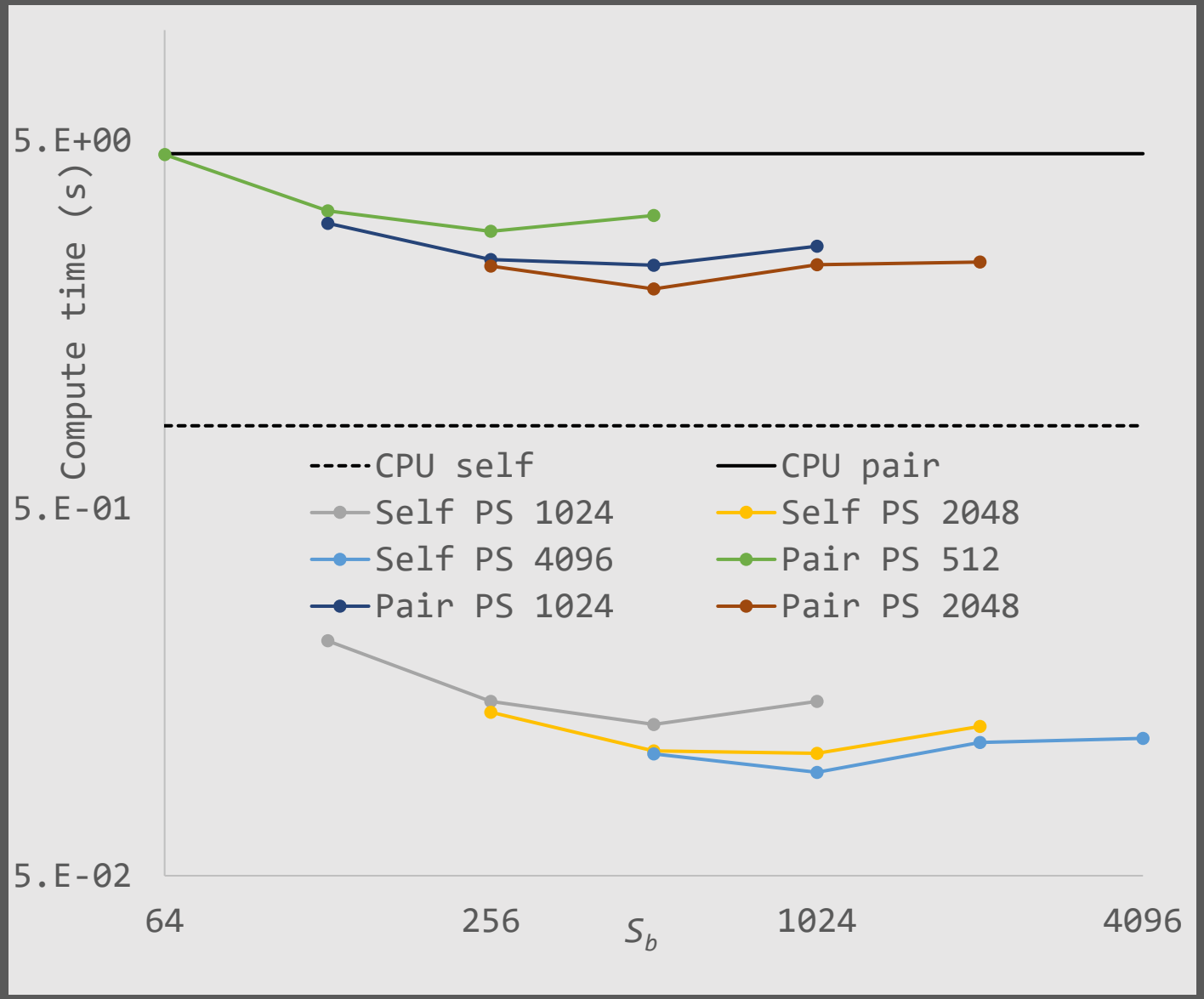
PAX-HPC

ExCALIBUR 10

Hardware Challenges—GPUs

- Optimal memory access different
  - CPU: array of structures
  - GPU: structure of arrays
- Many CPU processes and threads
  - All need access to GPUs concurrently
  - CPU-GPU bandwidth crucial; share GPU RAM?
  - Early Grace-Hopper results good—need more

Hydrodynamics on Grace-Hopper ➔ Compared to 32 Grace cores



- Best speedups:

- 3x for pair tasks
- 10x for self tasks

- Pair tasks slower as ratio of CPU-GPU data transferred/Computations smaller

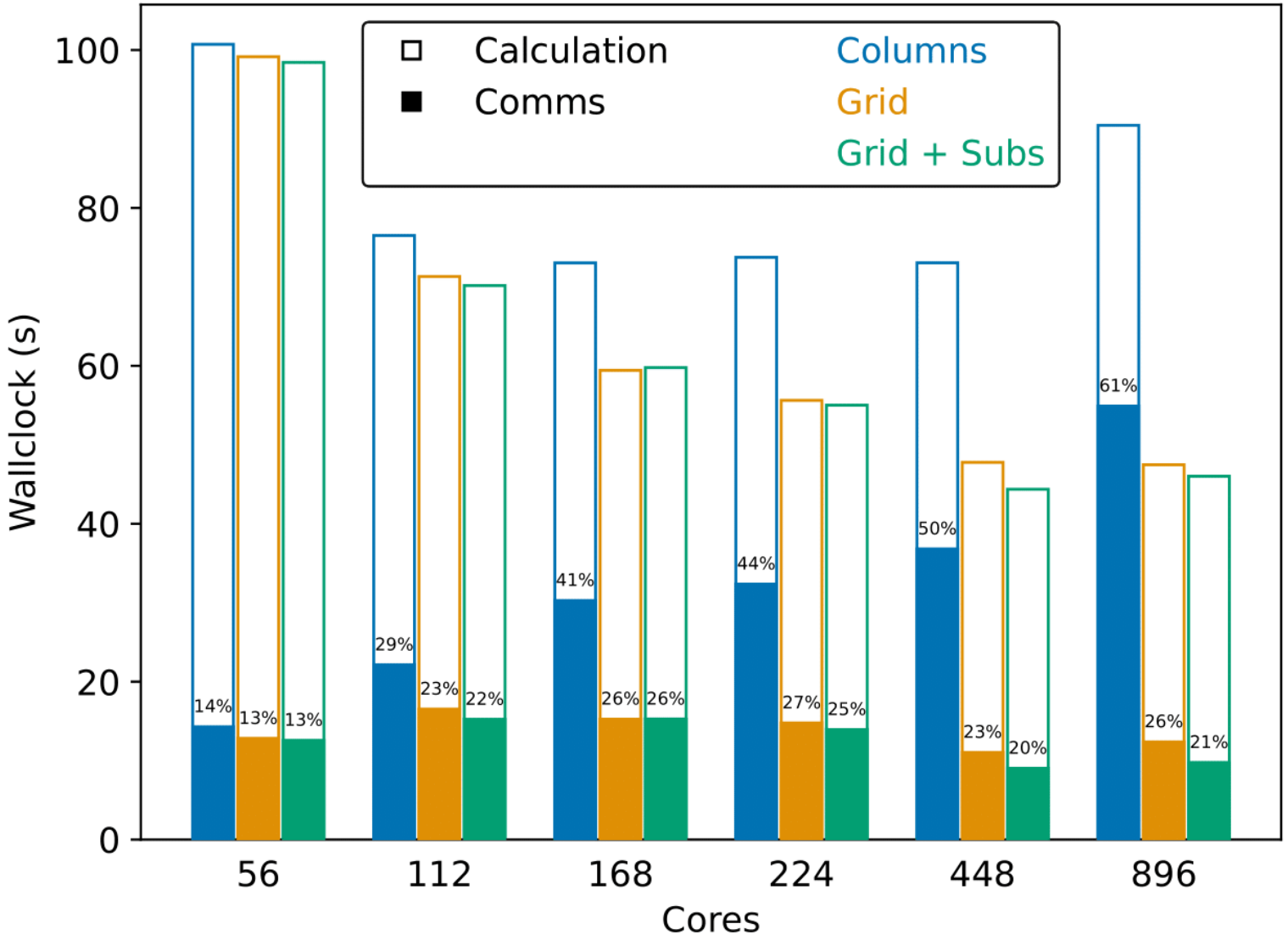
- **CPU-GPU connection speed CRUCIAL to achieving concurrency**

- **Need access to more Grace-Hopper nodes to test scalability!**

Hardware Challenges—Communication

- Scaling limited by latency-bound regime
- Hide with compute/other comms tasks, but:
  - MPI times unreliable (hard to predict)
- Collectives slow, especially all-to-allV
  - DPU acceleration?

CASTEP MPI Performance (Solid benzene on CSD3)

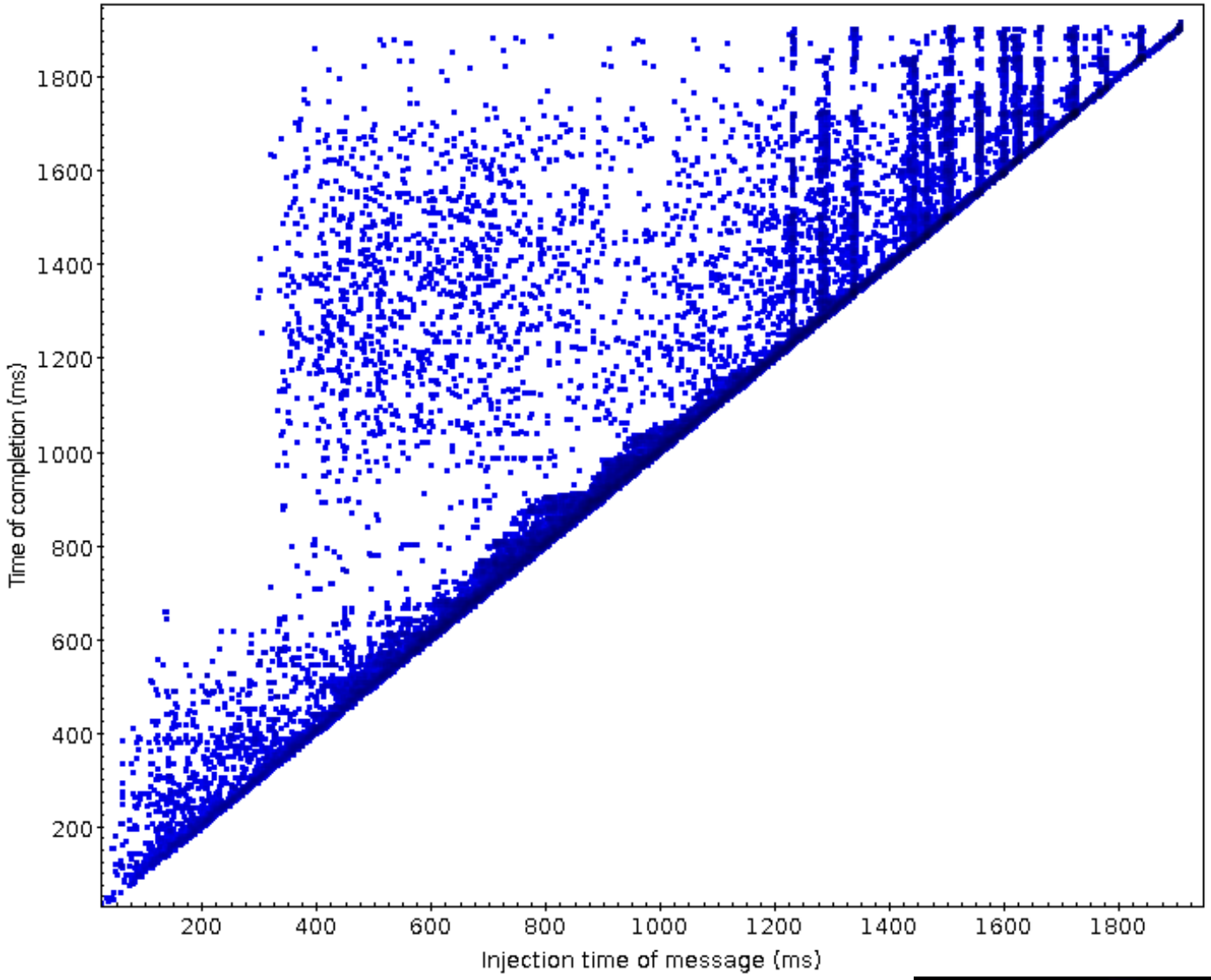


PAX-HPC

ExCALIBUR

SWIFT task-based parallelism

- Try RDMA instead of MPI
  - Faster to exchange data & less jitter
  - Receives take time too
  - Memory shuffling: Add extra threads to move data
- Memory handling is key—MPI hides this detail

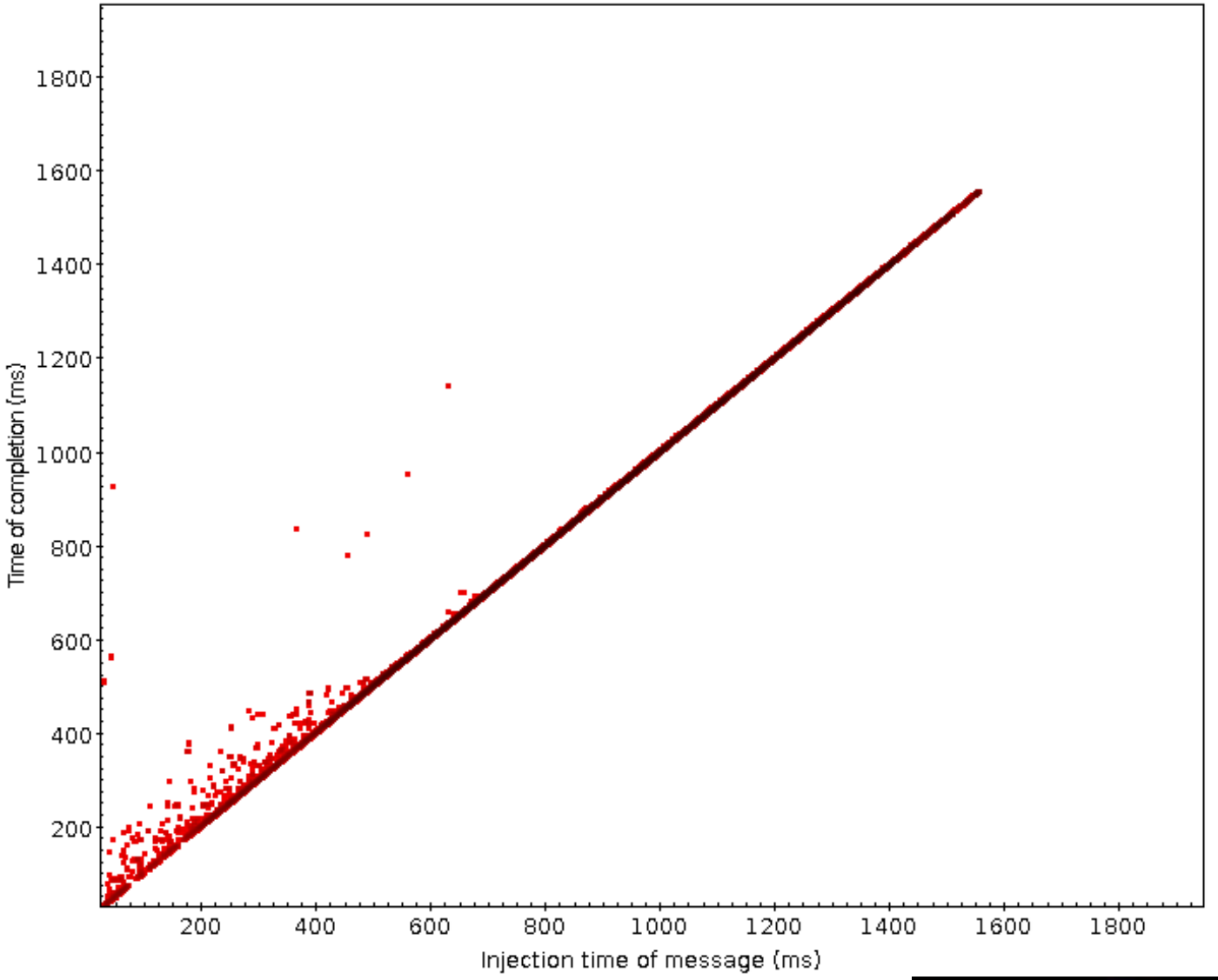


PAX-HPC

ExCALIBUR IO

SWIFT task-based parallelism

- Try RDMA instead of MPI
  - Faster to exchange data & less jitter
  - Receives take time too
  - Memory shuffling: Add extra threads to move data
- Memory handling is key—MPI hides this detail



PAX-HPC

EXCALIBUR IO

Hardware & Enabling Software

- PAX software widely used on global HPC
- Typically 40% of monthly ARCHER2 cycles
- Keen to explore new hardware
- Lots of GPU developments in PAX
- Little engagement with testbeds—why??

Hardware & Enabling Software Testbeds

- RSE time is the critical resource
  - Testbeds not known during project planning
  - No time on project
- Low visibility—what is available?
- Lack of clarity about access
  - When & how to access them?
- All fixable!

Hardware & Enabling Software Testbeds

- Many testbeds are interesting to us
- RSE time is the critical resource
- Hackathons good (e.g. DPU)
  - Focused time, distractions
  - Not long enough!
  - Lose access at end?

EXCALIBUR IO

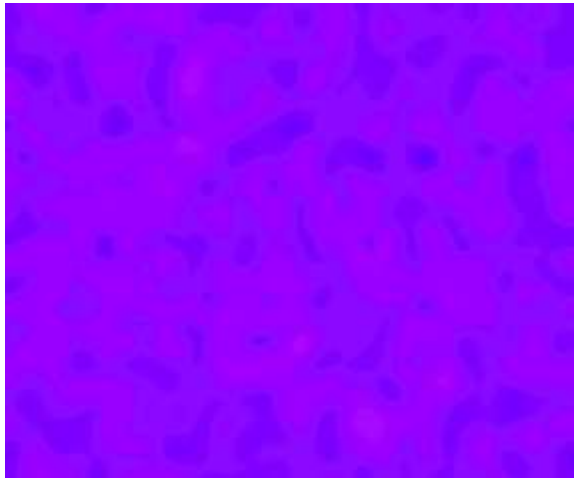
Suggestion

- More testbed hackathons
  - Raises awareness & provides access
  - Gives focused time
- Clear long-term access route
- eCSE-style funding
  - ~3 month (1-6) pilot RSE projects
  - Light touch application, quick turnaround

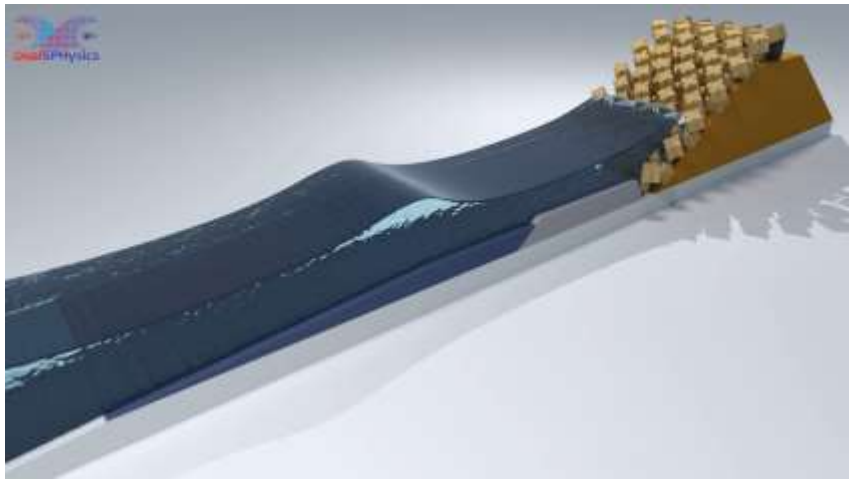
EXCALIBUR 10

PAX-HPC

Engineering and Physical Sciences Research Council



Hasnip, 2020



Altomare et al, 2014

Keggereis et al, 2019