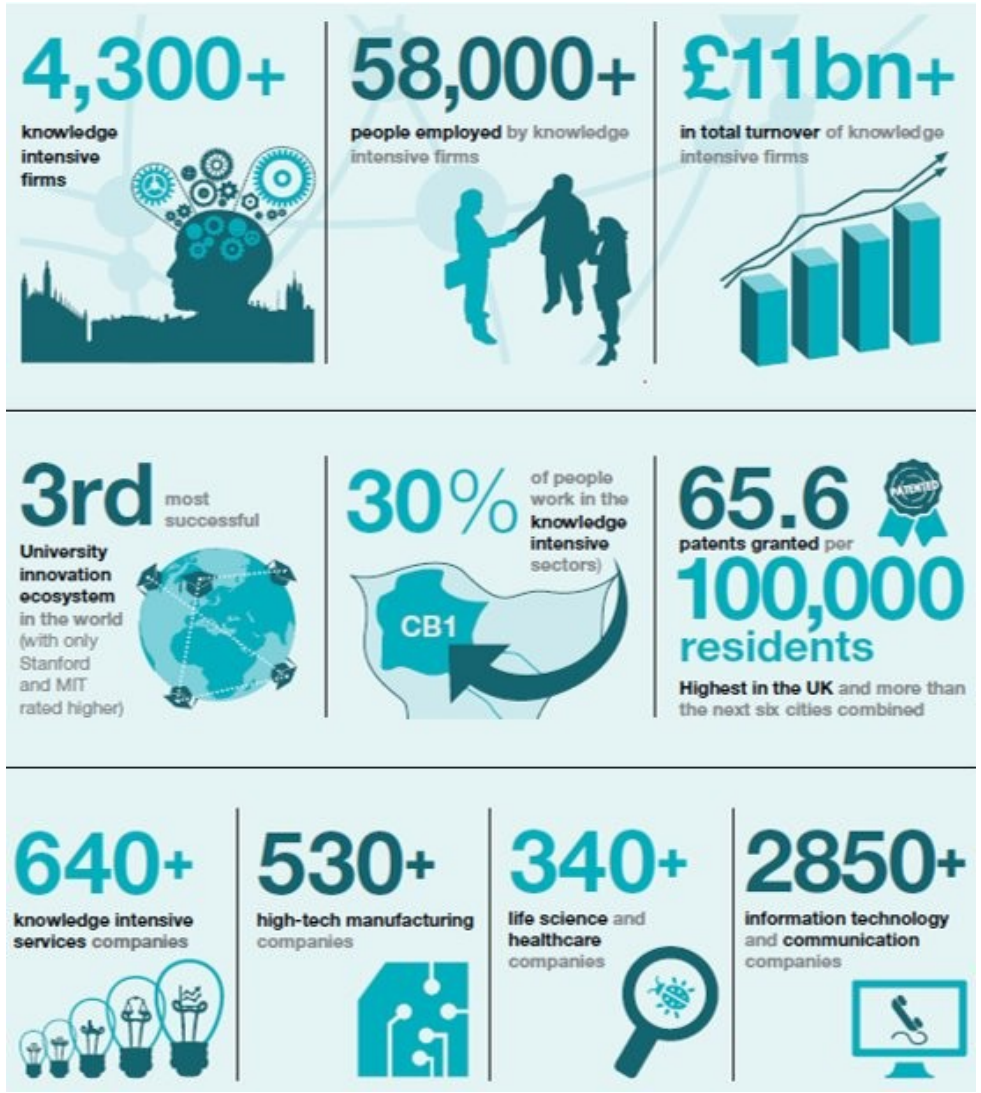
UNIVERSITY OF CAMBRIDGE

MVAPICH at the Cambridge Open Zettascale Lab

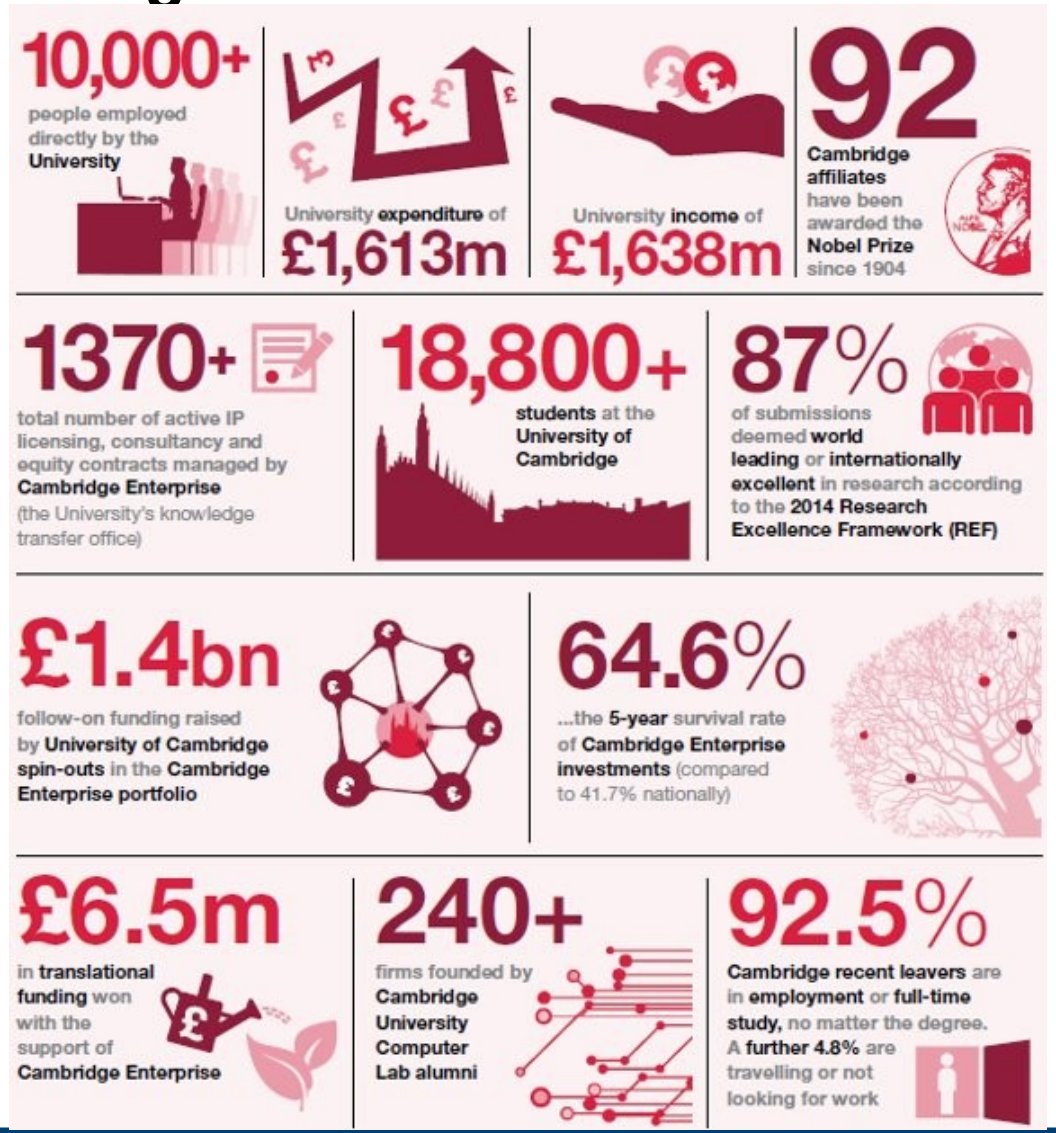
Chris Edsall
MVAPICH User Group August 2022

Research Computing Services

Cambridge Silicon Fen



University of Cambridge



Research Computing Services

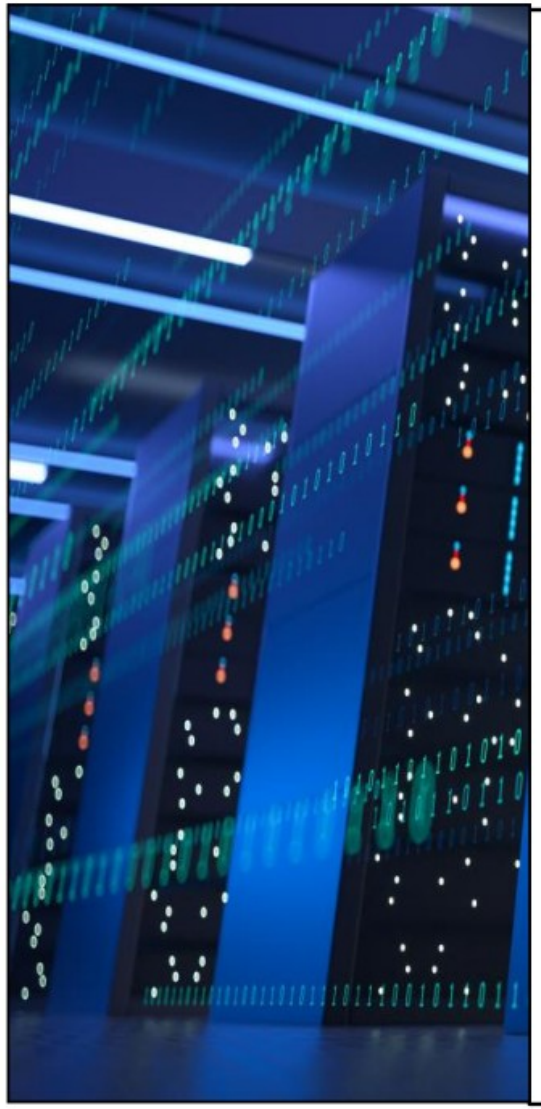
Provide leadership class research computing services - hardware, software and knowledge resources, to:-

- The university of Cambridge research community
- The UK national academic research community
- UK industrial users
- Wider international organisations via consultancy and system integration services via the Dell–Cambridge HPC solution centre & Cambridge spinout “Cambridge Research Computing Ltd”

CSD3

- The Cambridge System for Data Driven Discovery
- Top100 Supercomputer
- Dell Servers
- Partitions
  - Cascade Lake
    - 762 nodes
    - HDR100
  - Ice Lake
    - 544 nodes
    - HDR200
  - "ampere"
    - 80 Nodes
    - 4x A100 80 GiB
    - Dual rail ConnectX6
    - Bluefield2 DPU


CAMBRIDGE OPEN
ZETTASCALE LAB
CAMBRIDGE OPEN
EXASCALE LAB
intel
ISC High Performance



CAMBRIDGE OPEN ZETTASCALE LAB

- $5M public/industry funding - technology scanning Lab, with the goal of making exascale practical and ZettaScale possible
- Academic / industrial partnership for the testing & co-design of leading edge HPC, AI and DA solutions -20 engineers
- Pushing the boundaries of performance but more importantly making large scale HPC system more accessible thereby democratising HPC/AI and DA technologies for everyone at every-scale

UNIVERSITY OF CAMBRIDGE

DELL Technologies

intel

DiRAC

UKRI UK Research and Innovation

UK Atomic Energy Authority

Cambridge Open Zettascale Lab - Themes

- Energy-Efficient Computing
- Extreme-Scale Visualisation
- High-Performance Interconnects for Zettascale
- oneAPI Centre of Excellence
- Scientific OpenStack: Creating a Cloud-Native, Software-Defined Supercomputer
- Zettascale Systems for Urgent Computing
- Zettascale-Class Storage Solutions

Cambridge <-> OSU Collaboration

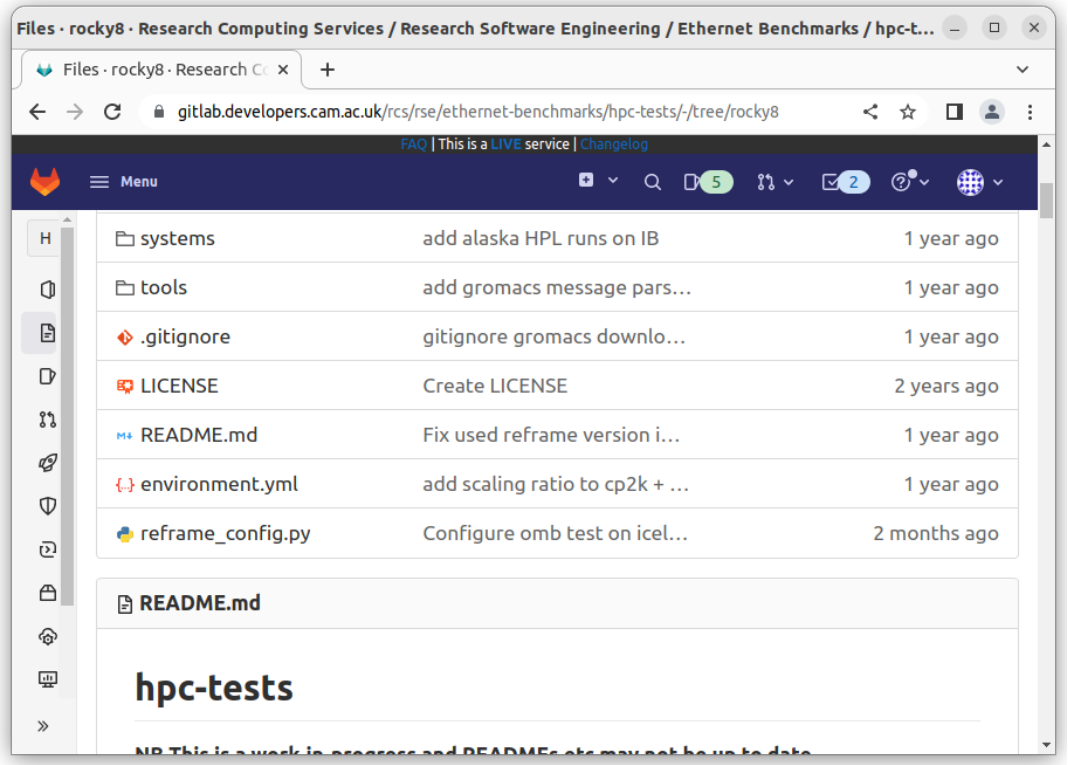
MVAPICH

- Long standing cooperation, all the way back to Wilkes-1
- Increasing collaboration across a range of activities

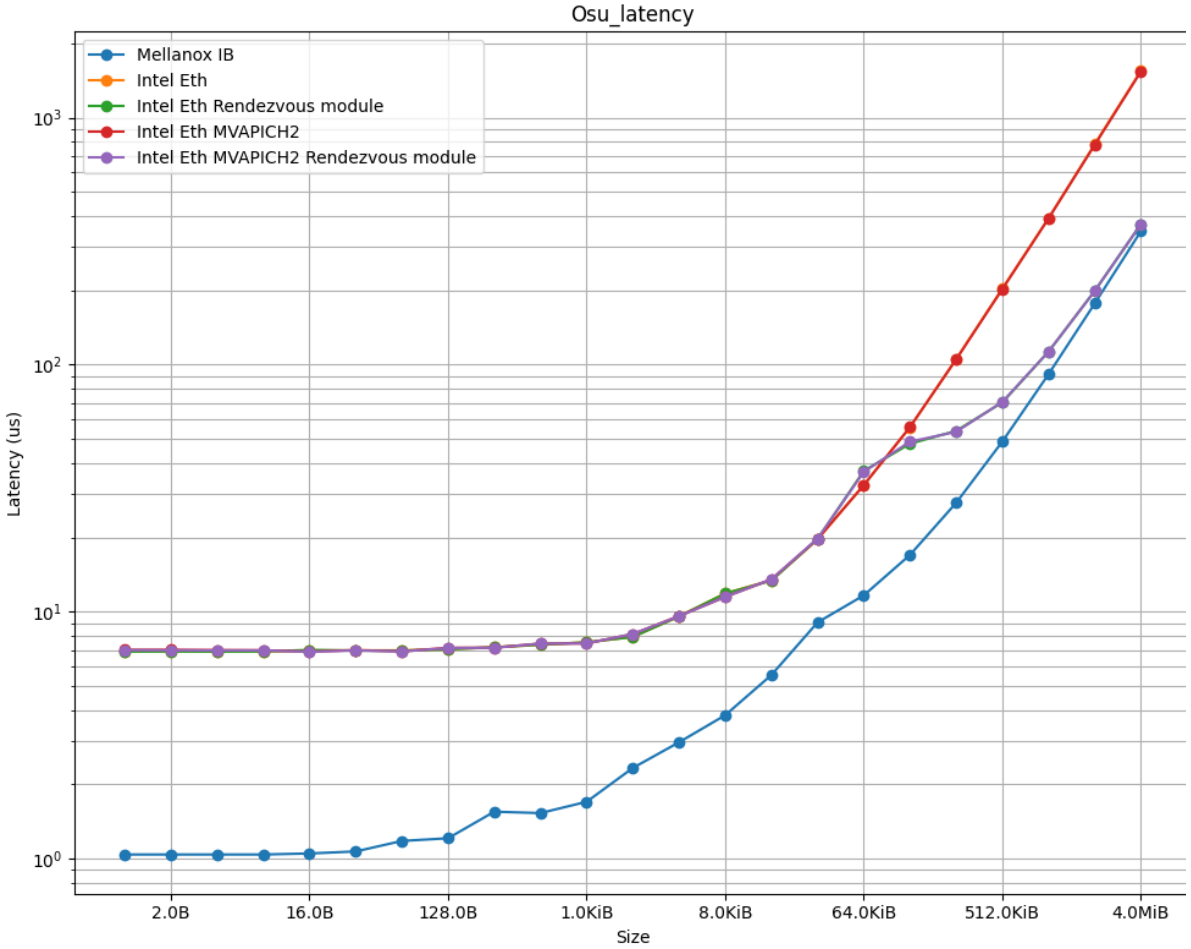
Stack HPC

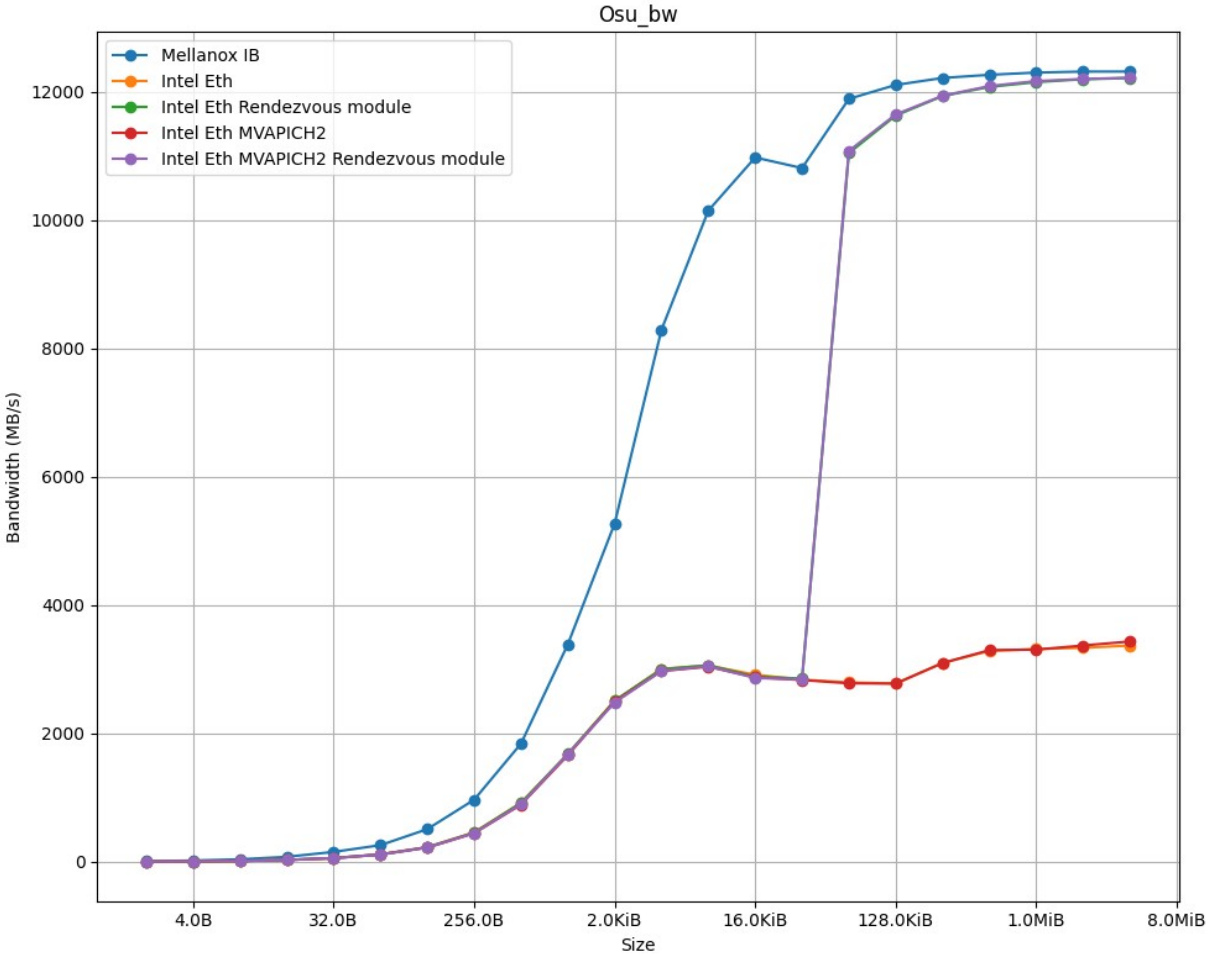
ReFrame

- Open, reproducible
- Spack, ReFrame
- Suite of Benchmarks
  - Synthetic (OSU Micro-Benchmarks)
  - Applications
    - OpenFOAM
    - GROMACS
    - WRF 4.x
    - CP2K
- Using to, e.g.
  - Compare ethernet NICs
  - MPI libraries

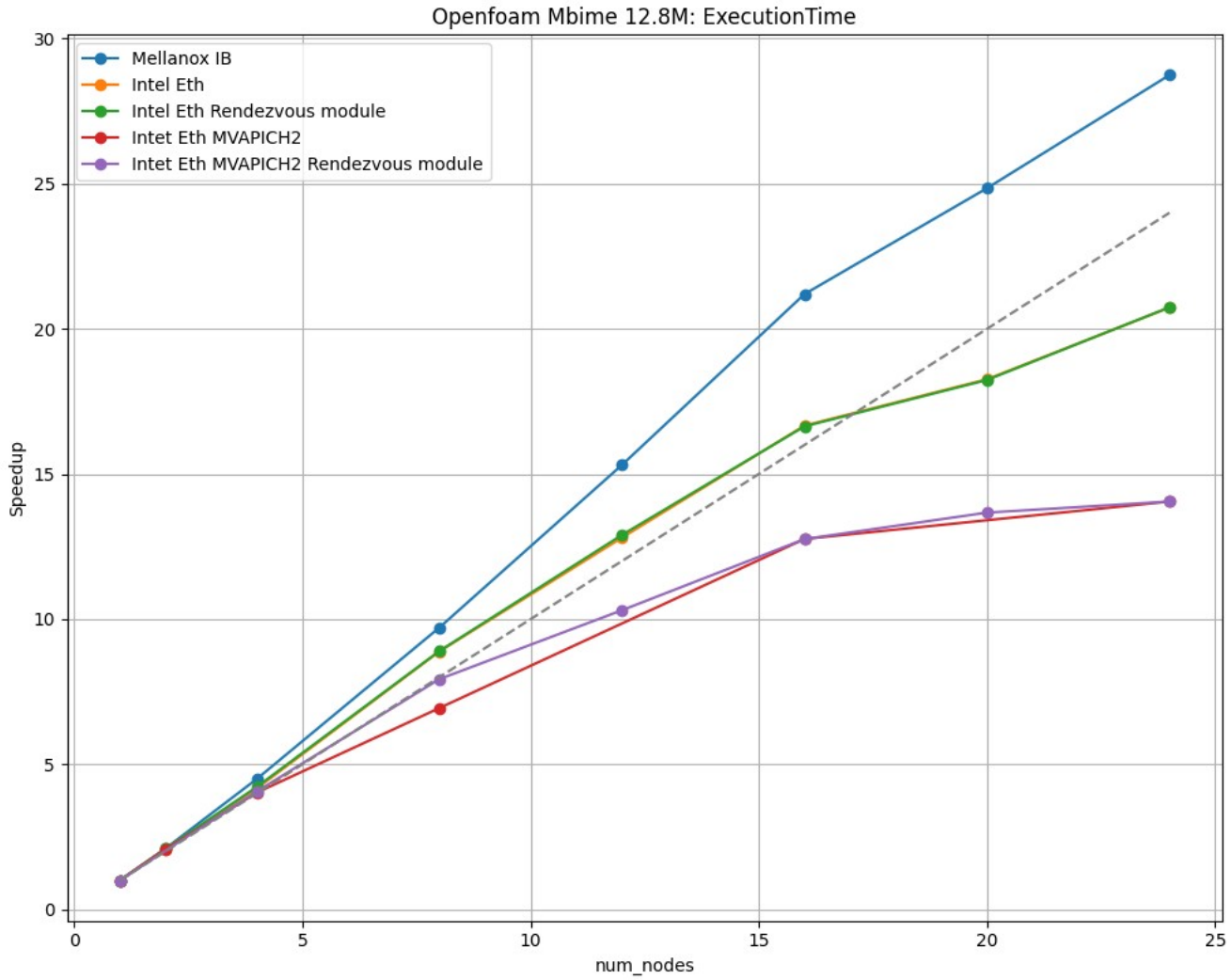


- Work by **Kacper Kornet** (Zettascale) and **Moustafa Abduljabbar** (OSU)
- Experimental partition on CSD3
  - Intel® Ethernet 800 Series (code named Columbiaville) (100Gb/s)
  - CascadeLake 56 nodes available, currently 28 each configured Intel Ethernet Fabric Suite and MOFED (because of conflict)
  - Dell ethernet switch (model Z9332, 400 Gb/s ports with 4–way split cables, 4X100 Gb/s)
- Morocco Cluster (same nodes and software as CSD3)
  - Mellanox ConnectX-6 (limited in our case 50 Gb/s because of split cables)
- Requires MVAPICH >=3 for PSM3 libfabric provider
  - Two choices for PSM3:
    - Intel Ethernet Fabric Suite <- this is the one we used in this work
    - Latest Intel MPI (2021.6) also ships PSM3

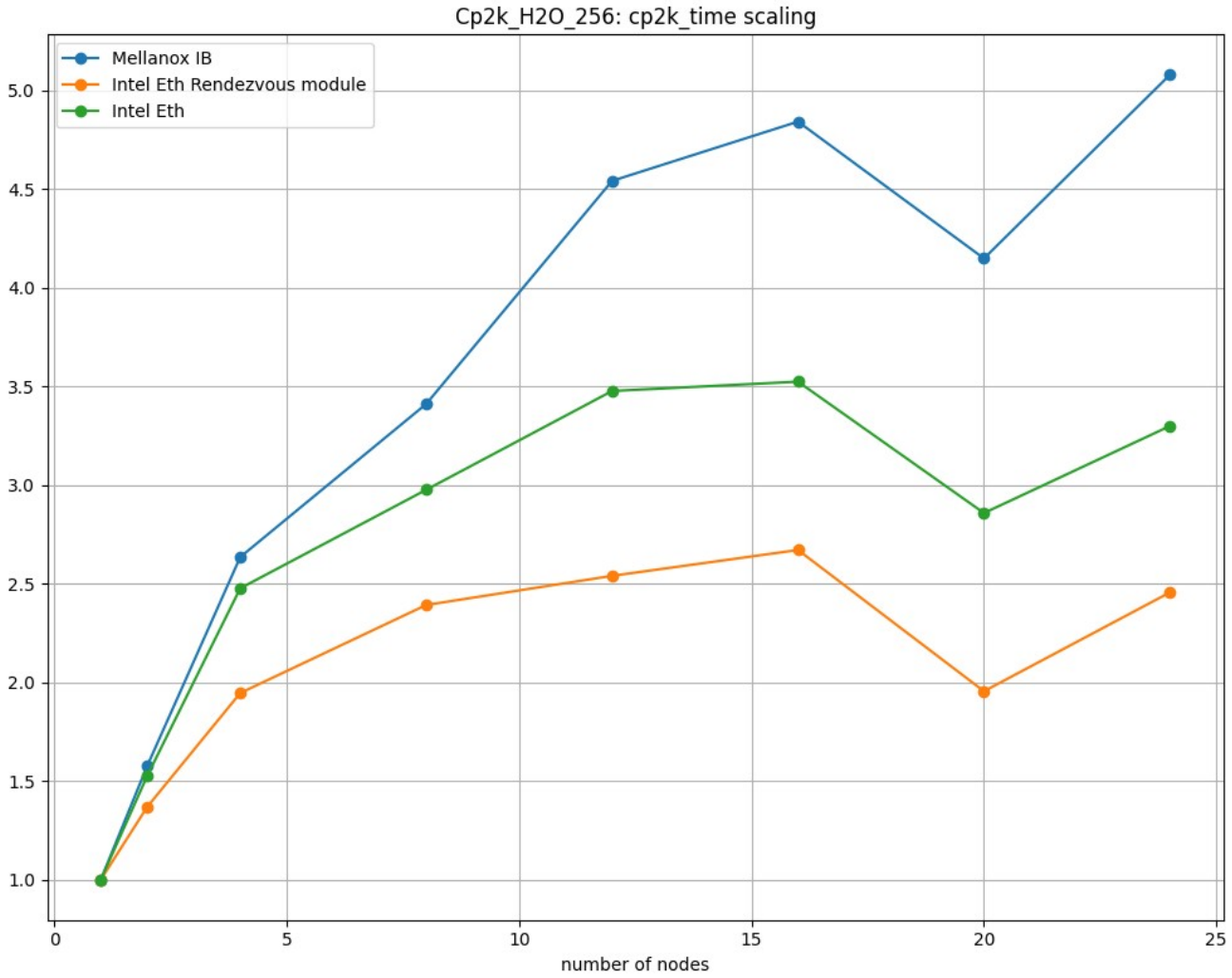

Osu_latency
Mellanox IB
Intel Eth
Intel Eth Rendezvous module
Intel Eth MVAPICH2
Intel Eth MVAPICH2 Rendezvous module
Latency (us)
10^3
10^2
10^1
10^0
2.0B
16.0B
128.0B
1.0KiB
8.0KiB
64.0KiB
512.0KiB
4.0MiB
Size


Osu_bw
Mellanox IB
Intel Eth
Intel Eth Rendezvous module
Intel Eth MVAPICH2
Intel Eth MVAPICH2 Rendezvous module
Bandwidth (MB/s)
12000
10000
8000
6000
4000
2000
0
4.0B
32.0B
256.0B
2.0KiB
16.0KiB
128.0KiB
1.0MiB
8.0MiB
Size

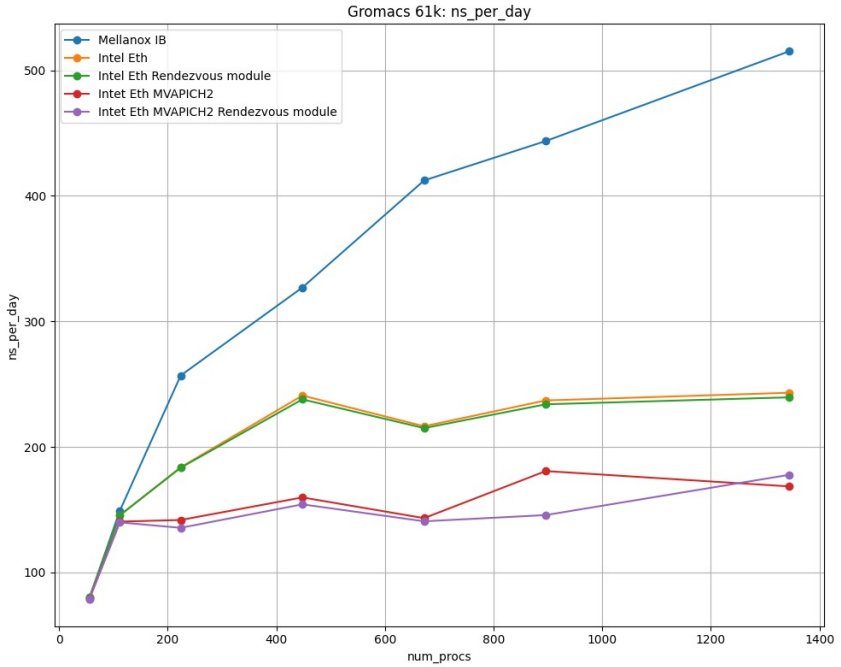
CFD - OpenFOAM

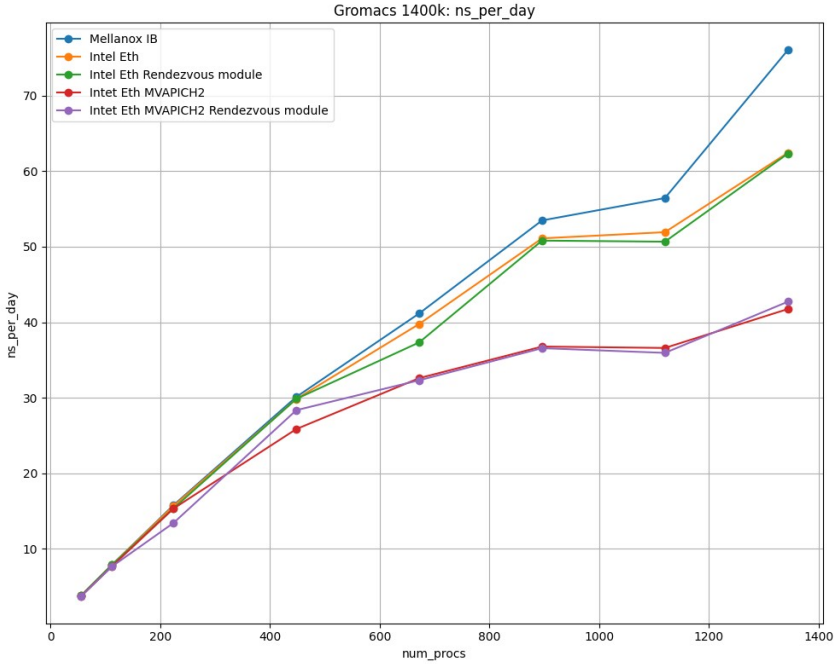


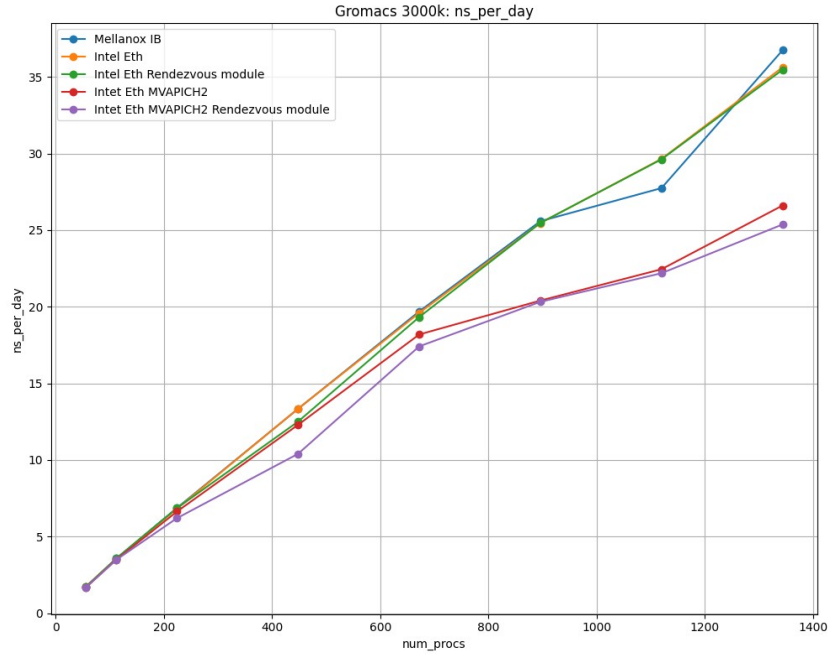
Materials CP2K



Molecular Dynamics - GROMACS







DPU

- MVAPICH2-DPU
- Accelerate MPI collective operations
- Applications?
  - CASTEP – UK based materials science code
    - MPI/OpenMP + accelerator offload
    - Scales to O(10k) processors
    - Top5 code on ARCHER2
    - Heavily FFT bound (30% wall time waiting on alltoallv comms)
    - Aim: communication hiding by overlapping computation and communication
    - Preliminary work (**Arjen Tamerus**, Zettascale Lab): would require significant re-architecting of the code.

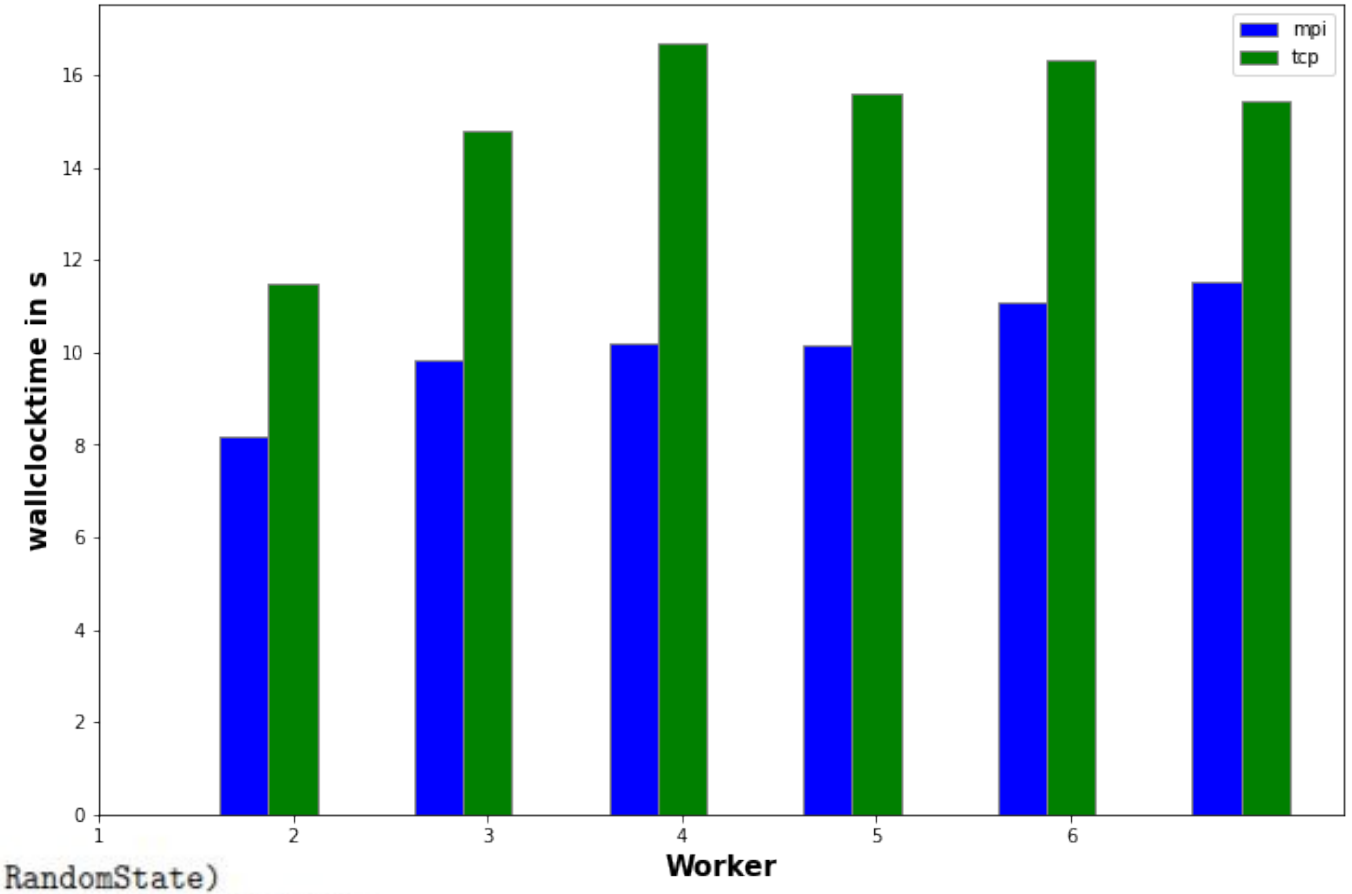
MVAPICH

dask

Data Analytics - Methods

- Work by **Stefanie Reuter** (Zettascale Lab) with assistance from **Amir Shafi** (OSU)
- Installed MVAPICH2 2.3.7
  - (Next try MVAPICH2-GDR, MVAPICH2-X)
- Installed the coda environment (based on python 3.8, dask 2021.1.1, distribute 2021.1.1)
- mpi4dask-0.2
- Ran test cases
  - numpy_sum_mpi.py on CSD3 Icelake
  - cupy_sum.py on CSD3 ampere partition (Wilkes-3)

Dask – CPU Sum

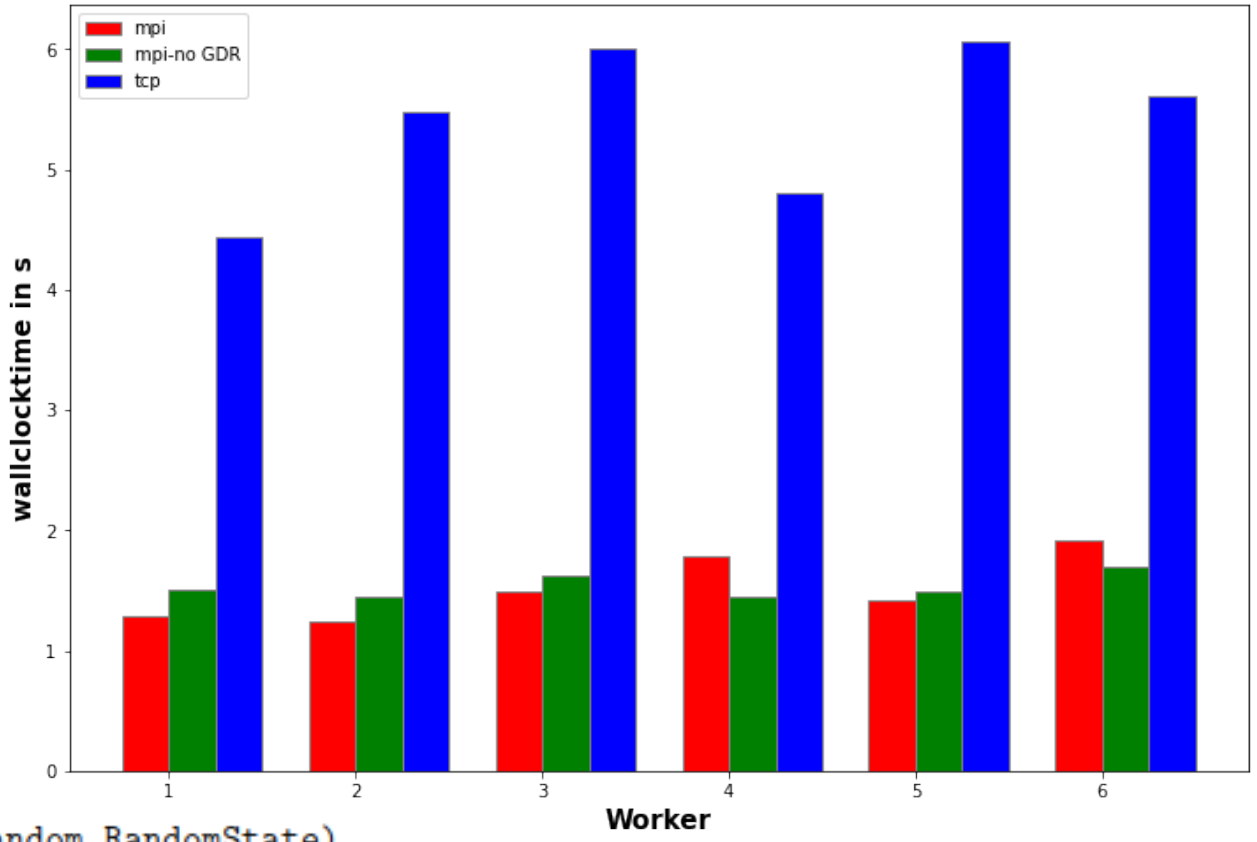



```
for i in range(RUNS):
    start = time.time()
    rs = da.random.RandomState(RandomState=numpy.random.RandomState)
    a = rs.normal(100, 1, (int(3e4), int(3e4)), chunks=(int(5e3), int(5e3)))

    x = a + a.T
    xx = await client.compute(x)

    duration = time.time() - start
```

Dask – GPU Cupy Sum




```
for i in range(RUNS):
    start = time.time()

    rs = da.random.RandomState(RandomState=cupy.random.RandomState)
    a = rs.normal(100, 1, (int(20000), int(20000)), \
                  chunks=(int(4000), int(4000)))
    x = a + a.T

    xx = await client.compute(x)

    duration = time.time() - start
```

DASK - Applications

Once we have got the software stack running and validated the next step is to try it on real world applications:

- Climate
- Square Kilometer Array

Summary

- We've set up the Zettascale Lab to look at challenges around the next generation of supercomputers
- Cambridge and OSU have begun a deeper collaboration

Contacts

- https://www.zettascale.hpc.cam.ac.uk/
- info-cozl@hpc.cam.ac.uk
- https://twitter.com/ZettascaleLab

- cje57@cam.ac.uk
- https://twitter.com/hpcchris