Science and Technology Facilities Council

Conference Proceedings
DL-CONF-2022-001

Computing Insight UK 2021

Manchester Central Convention Centre, UK

9th-10th December, 2021

G Lomas, D Jones (editors)

February 2022

Enquiries concerning this report should be addressed to:

Chadwick Library
STFC Daresbury Laboratory
Sci-Tech Daresbury
Keckwick Lane
Warrington
WA4 4AD

Tel: +44(0)1925 603397
Fax: +44(0)1925 603779
email: librarydl@stfc.ac.uk

Science and Technology Facilities Council reports are available online at:
https://epubs.stfc.ac.uk

ISSN 1362-0223

Conference Proceedings

Computing Insight UK (CIUK) 2021 took place on 9th and 10th December 2021 at the Manchester Central Convention Centre. These proceedings are a record of the presentations and posters from the Conference.

The CIUK Organising Committee would like to thank the exhibitors, sponsors, presenters and attendees who help to make the Conference a continued success.

GOLD Sponsors	SILVER Sponsors	BRONZE Sponsors
VAST DATACORE	ATEMPO Atos BOSTON Servers | Storage | Solutions BIOS IT Orchestrating a brighter world NEC	ALTAIR dug OCF GIGAIO KAO DATA SCAN ddn YELLOWDOG CORNELIS NETWORKS nexstor

Computing Insight UK 2021 Introduction

Computing Insight UK (CIUK) 2021 was the 32nd edition of an annual conference organised by the Science and Technology Facilities Council's (STFC) Scientific Computing Department (SCD).

Following the switch to an online event in 2020 due to the covid pandemic, we were delighted to be able to return to a physical, face-to-face conference in 2021. The event was held on the 9-10 December at Manchester Central.

The theme for CIUK 2021 was "Heterogeneous Computing" with sub-themes including subjects such as "Why Heterogeneous Computing?", "Hardware Sub-Systems", "Co-Design and Environment" and "Emerging Technologies".

As with previous years CIUK 2021 also included a student poster competition, with the selected finalists displaying their posters online and also physically during the conference, and the presentation of the CIUK Jacky Pallas Memorial Award, which this year was awarded to Dr Niall Jeffrey (Ecole Normale Supérieure, France and University College London) for his work on "Mapping dark matter with the Dark Energy Survey and AI". Niall presented his work as part of the main programme during the conference.

CIUK 2021 also saw the second edition of the CIUK Cluster Challenge competition. The defending champions from Durham University took on a combined team of students from Bristol University and Bath University, with a series of online challenges in the weeks leading up to the conference and then four more challenges during the conference in Manchester. Team Bristol / Bath took the title after a closely fought competition and earned their place at the ISC'22 Cluster Challenge competition where they will represent CIUK against the best student teams from around the word.

CIUK 2021 Programme

Day 1 - Friday 10 December

	Session 1: Why Heterogeneous Computing?
	Session 2: Hardware Sub-Systems
	Session 3: The ExCALIBUR Programme

TIME	MAIN PROGRAMME	BREAKOUT SESSIONS
From 08:30	**Registration Open (Main Foyer) Exhibition Open (Gallery)**	
09:15 - 09:30	**Tom Griffin** (Director, Scientific Computing, STFC) *Welcome and Introduction*	
09:30 - 10:00	**Steve Hindmarsh** (Head of Scientific Computing, The Francis Crick Institute) *Heterogeneous Computing at the Crick*	
10:00 - 10:30	**Phil Hasnip** (University of York) *Portable acceleration of materials modelling software: CASTEP, GPUs and OpenACC*	CoSeC Computational Science Centre for Research Communities **Annual Conference 2021** Thursday 9 December @ CIUK 2021
10:30 - 11:00	**Dr Igor Baratta** (Department of Engineering, University of Cambridge) *Heterogeneous Programming for Finite Element: insights from benchmarks*	
11:00 - 11:30	**REFRESHMENTS**	
11:30 - 12:00	**Paul Calleja** (University of Cambridge) *The Modern Cloud Native Heterogeneous Super Computer – A Converged Platform for Simulation, AI & Data Analytics*	
12:00 - 12:30	**Alastair Basden** (Durham University) *The Durham Intelligent NIC Environment (DINE)*	
12:30 - 13:00	**Andrew Edmondson** (Research Software Group Leader, Advanced Research Computing, University of Birmingham) *Installing and curating software for heterogeneous compute environments*	
13:00 - 14:00	**LUNCH**	

14:00 - 14:30	**The ExCALIBUR Programme** **Dr Elizabeth Bent** (Senior Portfolio Manager, UKRI-EPSRC) *Harnessing Exascale Computing – an ExCALIBUR overview* **Rob Akers** (UKAEA) *ExCALIBUR – exploiting the exascale to bottle a star* **Nigel Wood** (Met Office) *ExCALIBUR and the quest for the holy grail of weather & climate prediction* **ExCALIBUR Panel Discussion and Q&A Session**	
14:30 - 15:00		
15:00 - 15:30		
15:30 - 16:15	**REFRESHMENTS**	
16:15 - 17:00	**Martyn Guest** (ARCCA, Cardiff University) *Performance of Computational Chemistry Codes. An Analysis of Molecular Dynamics and Electronic Structure Applications on Multi-core Processors*	
17:00 - 18:00	**Keynote Presentation - Simon McIntosh-Smith (Bristol University)** *Heterogeneous Computing: past, present and future*	
18:00 - 20:00	**CIUK 2021 Networking Event** Join us at Manchester Central for wine, beer and nibbles. All registered attendees are invited!	

Day 2 - Friday 10 December

	Session 4: Co-Design and Environment
	Session 5: The Skills Gap
	Session 6: Emerging Technologies

TIME	MAIN PROGRAMME	BREAKOUT SESSIONS
From 08:30	**Registration Open (Main Foyer)** **Exhibition Open (Gallery)**	
09:30 - 10:00	**Professor Mark Parsons** (EPCC Director at The University of Edinburgh / EPSRC Director of Research Computing) *The UK Exascale Project*	
10:00 - 10:30	**Jeff Hammond and Filippo Spiga (NVIDIA)** *Shifting through the Gears of GPU Programming: Understanding Performance and Portability Trade-offs*	

10:30 - 11:00	**The Jacky Pallas Memorial Presentation** **Dr Niall Jeffrey (Ecole Normale Supérieure & University College London)** *Mapping dark matter with the Dark Energy Survey and AI*	 Spectrum Scale User Group
11:00 - 11:30	**REFRESHMENTS**	
11:30 - 12:00	**The UK Skills Gap** **Richard Gunn** (UKRI) Skills gaps in the context of the wider approach to DRI within UKRI **Michael Ball** (BBSRC) UKRI's approach to supporting DRI skills **Prof Mark Wilkinson** (DiRAC) The DiRAC Facility HPC skills training programme **Christine Kitchen** (Cardiff University) Ideas to address skills gaps **Andrew Medhurst** (Inspire People) Title TBC **Skills Gap Panel Discussion and Q&A Session**	
12:00 - 12:30		
12:30 - 13:00		
13:00 - 14:15	**LUNCH**	
14:15 - 14:30	**Award Presentation** The CIUK 2021 Student Cluster Challenge and Poster Competition	
14:30 - 15:00	**Prof Viv Kendon** (University of Strathclyde) *QEVEC: integrating quantum computing with HPC*	
15:00 - 15:30	**Nick Brown** (EPCC, University of Edinburgh) *FPGAs for scientific workloads: The why and the how*	
15:30 - 16:00	**Gihan Mudalige (Reader (Associate Professor), University of Warwick, Department of Computer Science)** *Multi-Layered Abstractions for Performance Portability - Lessons Learnt and Challenges*	
16:00	**CIUK 2021 CLOSES**	

Welcome to CIUK 2021

"Heterogeneous Computing“

Science and Technology Facilities Council

Scientific Computing

- Two Days of Presentations
- Exhibition of the Latest Technology
- Parallel Breakout Sessions
- Student Poster Competition
- CIUK Cluster Challenge
- Research Zone
- CIUK 2021 Networking Event

CIUK 2021

Welcome to CIUK 2021

Science and Technology Facilities Council
Scientific Computing

COVID Safety Measures

Mandatory

Temperature checks on entrance to venue

Adequate ventilation (to current regulation levels) in the exhibition and presentation areas

Spacing of seating within the presentation areas and seating in communal areas

Use of face coverings in communal areas

Advisable

All attendees should complete a lateral flow test before arriving at Manchester Central. If the result is positive then you should not attend.

Attendees should scan the QR code for track & trace purposes

Regular hand sanitising whilst on site

Maintain social distancing where possible and respect other attendees space

Please respect our COVID safety measures…
Wear a face covering ✓
Respect social distancing ✓

Thank you!

CIUK 2021

CIUK Code of Conduct

Science and Technology Facilities Council
Scientific Computing

The organisers are committed to making this conference productive and enjoyable for everyone, regardless of sex, gender identity, sexual orientation, disability, age, physical appearance, body size, ethnicity, nationality or religion/belief. We will not tolerate harassment of participants in any form.

We are committed to achieving a balanced and diverse panel of speakers at all our events by inviting speakers of all races, ethnicities, genders, ages, abilities, religions, and sexual orientation without compromising the quality, and remaining within the topic, of the programme.

CIUK Code of Conduct

Science and Technology Facilities Council
Scientific Computing

Behave professionally.

Harassment and sexist, racist, or exclusionary comments or jokes are not appropriate.

Harassment includes sustained disruption of talks or other events, inappropriate physical contact, sexual attention or innuendo, deliberate intimidation, stalking, and photography or recording of an individual without consent.

It also includes offensive or belittling comments related to sex, gender identity, sexual orientation, disability, age, physical appearance, body size, ethnicity, nationality or religion/belief.

All communication should be appropriate for a professional audience including people of many different backgrounds. Sexual language and imagery are not appropriate.

Be kind to others. Do not insult or put down other attendees.

Delegates are reminded that STFC holds the right to remove any person who does not adhere to the code of conduct.

CIUK 2021

GOLD Sponsors	SILVER Sponsors	BRONZE Sponsors
VAST DATACORE	Atempo Atos BOSTON Servers | Storage | Solutions BIOS IT Orchestrating a brighter world NEC	ALTAIR dug OCF GIGAIO KAO DATA SCAN ddn YELLOWDOG CORNELIS NETWORKS nexstor

CIUK 2021 Research Zone

N8 CIR BEDE · JASMIN · cirrus Powered by epcc · HPC Midlands+ · MMM Hub · NI-HPC Northern Ireland · B

CIUK 2021 MEDIA PARTNER

SCIENTIFIC COMPUTING WORLD

CIUK 2021 EXHIBITORS

X	CIUK Information Point
R	Research Zone - BEDE
R	Research Zone - SULIS
R	Research Zone - CIRRUS
R	Research Zone - MMM Hub
R	Research Zone - NI-HPC
R	Research Zone – JASMIN
R	Research Zone – Baskerville
P	Poster – Thomas Johnson
P	Poster – Megan Ratcliffe
P	Poster – Adam Tuft
P	Poster – Kieran Woodward
P	Poster – Jemma Bennett
P	Poster – Shrey Bhardwaj
1	
2	Cornelis Networks
3	Atos
4	Atempo
5	Boston Limited
6	BIOS IT
7	NEC Deutschland GmbH
8	DUG Technology
9	DDN
10	SambaNova Systems Inc
11	
12	Alces Flight
13	Kao Data
14	
15	SCAN
16	StrongBox Data Solutions
17	Scientific Computing World
18	GigaIO
19	NVIDIA
20	YellowDog
21	SoftIron
22	Panasas
23	Nexstor
24	Altair
25	VAST Data
26	Intel
27	Lenovo
28	IBM
29	OCF
30	DataCore Software / NAS

Welcome to CIUK 2021

Keynote Presentation

Thursday 9 December
17:00 – 18:00

Simon McIntosh-Smith
(Bristol University)

Heterogeneous Computing: past, present and future

UKRI Science and Technology Facilities Council
Scientific Computing

CIUK 2021

Welcome to CIUK 2021

UKRI Science and Technology Facilities Council
Scientific Computing

Jacky Pallas Memorial Award

Friday 10 December
10:30 – 11:00

Dr Niall Jeffrey

(Ecole Normale Supérieure & University College London)

Mapping dark matter with the Dark Energy Survey and AI

CIUK 2021

Welcome to CIUK 2021

Science and Technology Facilities Council
Scientific Computing

Want to ask a question at the end of a presentation? We are using slido…

slido

Join at
slido.com
#CIUK2021

CIUK 2021

Welcome to CIUK 2021

Science and Technology Facilities Council
Scientific Computing

CoSeC Annual Conference

CoSeC
Computational Science Centre for Research Communities
Annual Conference 2021
Thursday 9 December @ CIUK 2021

Thursday 9 December
10:00 – 16:00

Spectrum Scale User Group

Spectrum Scale User Group

Friday 10 December
10:00 – 12:00

CIUK 2021

Welcome to CIUK 2021

UKRI Science and Technology Facilities Council

Scientific Computing

N8 CIR Computationally Intensive Research

BEDE

hpc MIDLANDS+

JASMIN

cirrus Powered by epcc

MATERIALS AND MOLECULAR MODELLING HUB

NI-HPC NORTHERN IRELAND HIGH PERFORMANCE COMPUTING

CIUK 2021

Welcome to CIUK 2021

Science and Technology Facilities Council
Scientific Computing

Join the conversation on Twitter…

@CompInsightUK
#CIUK2021

CIUK 2021

Welcome to CIUK 2021

Science and Technology Facilities Council
Scientific Computing

COVID Safety Measures

Mandatory

Temperature checks on entrance to venue

Adequate ventilation (to current regulation levels) in the exhibition and presentation areas

Spacing of seating within the presentation areas and seating in communal areas

Use of face coverings in communal areas

Advisable

All attendees should complete a lateral flow test before arriving at Manchester Central. If the result is positive then you should not attend.

Attendees should scan the QR code for track & trace purposes

Regular hand sanitising whilst on site

Maintain social distancing where possible and respect other attendees space

Please respect our COVID safety measures…
Wear a face covering ✓
Respect social distancing ✓

Thank you!

CIUK 2021

CIUK 2021 Presentations

Steve Hindmarsh (Head of Scientific Computing, The Francis Crick Institute)

Heterogeneous Computing at the Crick

Abstract: The presentation will provide an overview of the current and future Scientific Computing landscape at the Crick, which features a heterogeneous range of compute resources and (crucially) the skills and support for researchers to apply them effectively. It will also show examples of Crick research to understand life and benefit human health, made possible by Scientific Computing.

The overall premise of the talk is that our researchers need access to a wide range of computing resources to enable their research, and help to make the best use of them.

Bio: I joined the Crick in 2017 and lead the Scientific Computing Science Technology Platform (core facility) with a team of 20+ staff providing specialist scientific computing services including software development/engineering, machine learning/AI, research data services and databases to over 1500 researchers. We provide specialist support to 100+ research groups and 15 other Science Technology Platforms at the Crick as well as participating in collaborations across the wider biomedical community.

I am a biology graduate that 'defected' to IT, with over 20 years of experience working with scientists to provide the scientific computing tools to enable their research. I was previously Head of IT at the NERC Centre for Ecology & Hydrology, where I was responsible for provision of scientific computing and core IT services and support.

UKRI Medical Research Council

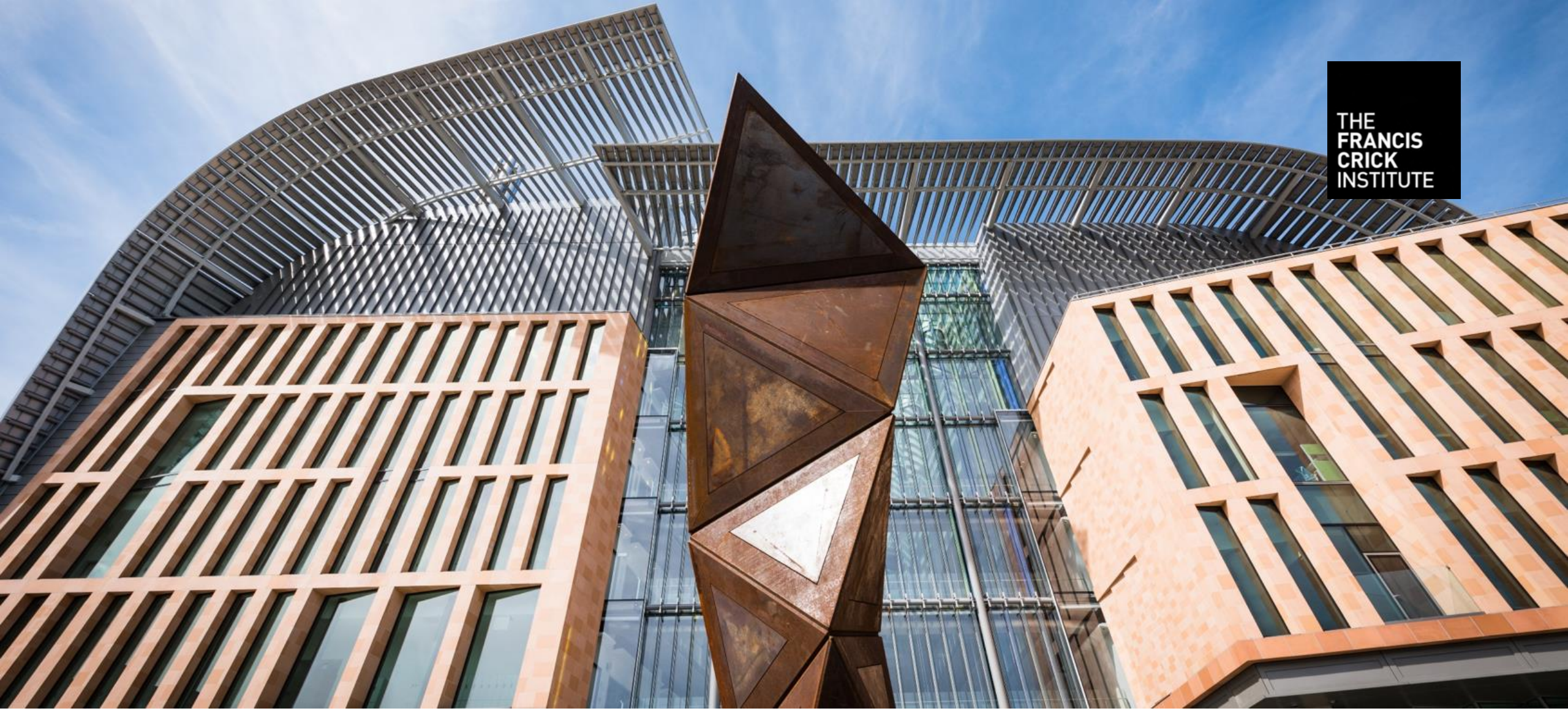
UCL

Imperial College London

King's College London

Wellcome

Cancer Research UK



UKRI Medical Research Council

UCL

Imperial College London

King's College London

Wellcome

Cancer Research UK

Heterogeneous Computing at the Crick

Steve Hindmarsh
Head of Scientific Computing, The Francis Crick Institute
CIUK 9th December 2021

Steve.Hindmarsh@crick.ac.uk

THE FRANCIS CRICK INSTITUTE

The Francis Crick Institute

- A biomedical discovery institute dedicated to ***understanding the fundamental biology underlying health and disease***.
- Founded in 2015 with the merger of two London research institutes from the Medical Research Council and Cancer Research UK into the Crick
- Supported by our founding partners:

MRC Medical Research Council · UCL · Imperial College London · King's College London · Wellcome · Cancer Research UK

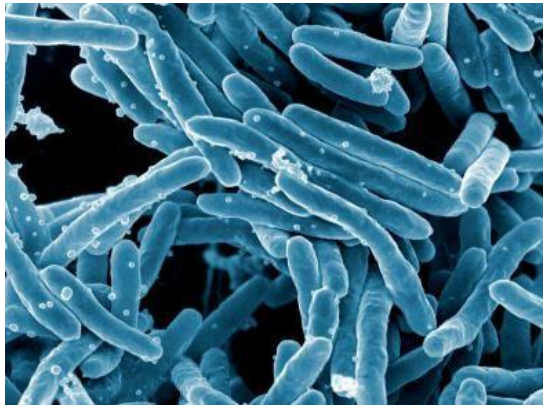
THE FRANCIS CRICK INSTITUTE

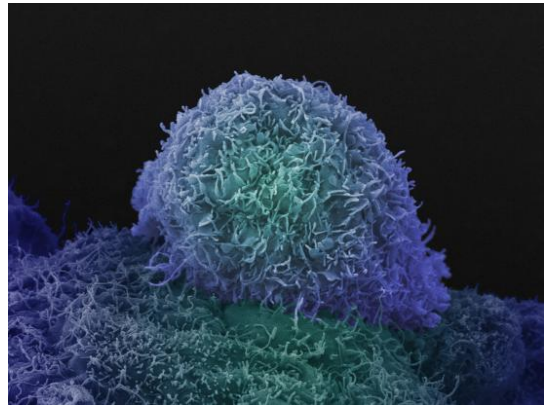
“Discovery without boundaries”

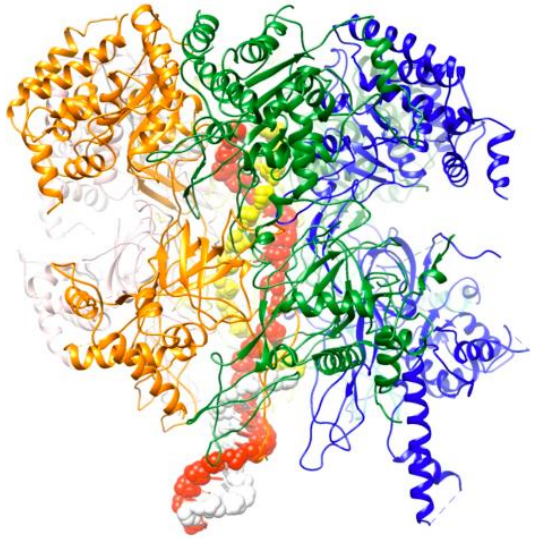
- Core research areas:

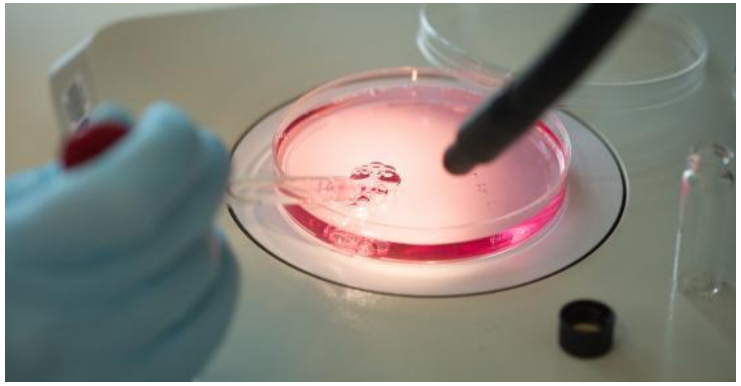
- Growth and development
- Health and ageing
- Human biology
- Cancer
- Immune system
- Infectious disease (including COVID-19!)
- Neuroscience

Multi-disciplinary approach: biology, physics, chemistry, bioinformatics, maths, engineering…









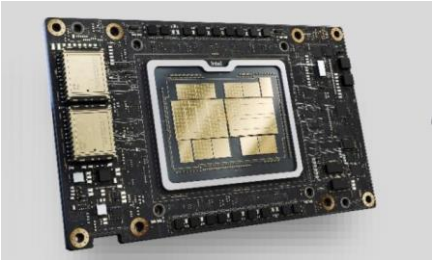
Spoiler alert: Not just HPC!

THE FRANCIS CRICK INSTITUTE

HPC

+ ? + ?

Compute - mainstream research trends:



Past	Present	Future
Gigaflops	Petaflops	Exaflops
CPU	CPU + GPU	CPU + GPU + FPGA + custom ASICs + xPU + Quantum?

↑ Processor heterogeneity – **workload** specific
↑ Parallelism (# processor cores and # nodes)
↑ Memory capacity & bandwidth (but ↓ per core)
↑ Power and cooling requirements

Cloud:

Past	Present	Future
On-site cluster	On-site clusters + Private Cloud/ VMs + **niche** Public Cloud	On-site/ Public Cloud **hybrid**

- **Workload** dependent – ‘data gravity’, network, new processor types
- Cost? (Storage vs Compute), CAPEX → OPEX Funding
- Not everything will be cloud!

Ease of access:

Past	Present	Future
UNIX/Linux Command line Interface (CLI)	CLI + basic GUI	CLI + GUIs + embedded in applications

↑ Access by researchers, clinicians, industry partners etc.

↓ Barrier to HPC benefits

CAMP: Crick Data Analysis and Management Platform

THE FRANCIS CRICK INSTITUTE

OCF

DDN STORAGE

Mellanox TECHNOLOGIES

CAMP

Lenovo

DELL Technologies

intel

NVIDIA

IBM Spectrum Scale

Crick CPU cluster (2016-)

THE FRANCIS CRICK INSTITUTE

~3000 physical cores, (6000 virtual cores with hyperthreading)

- 194 Regular CPU nodes: 2 x 8-core Intel Haswell, 256 GB RAM
- 4 High RAM CPU nodes: 4 x 12-core Intel Haswell, 1.5 TB RAM
- 8 interactive CPU nodes: 2 x 12-core Intel Skylake, 384 GB RAM
- InfiniBand FDR + 40G Ethernet

OCF

Lenovo

intel

Mellanox
TECHNOLOGIES

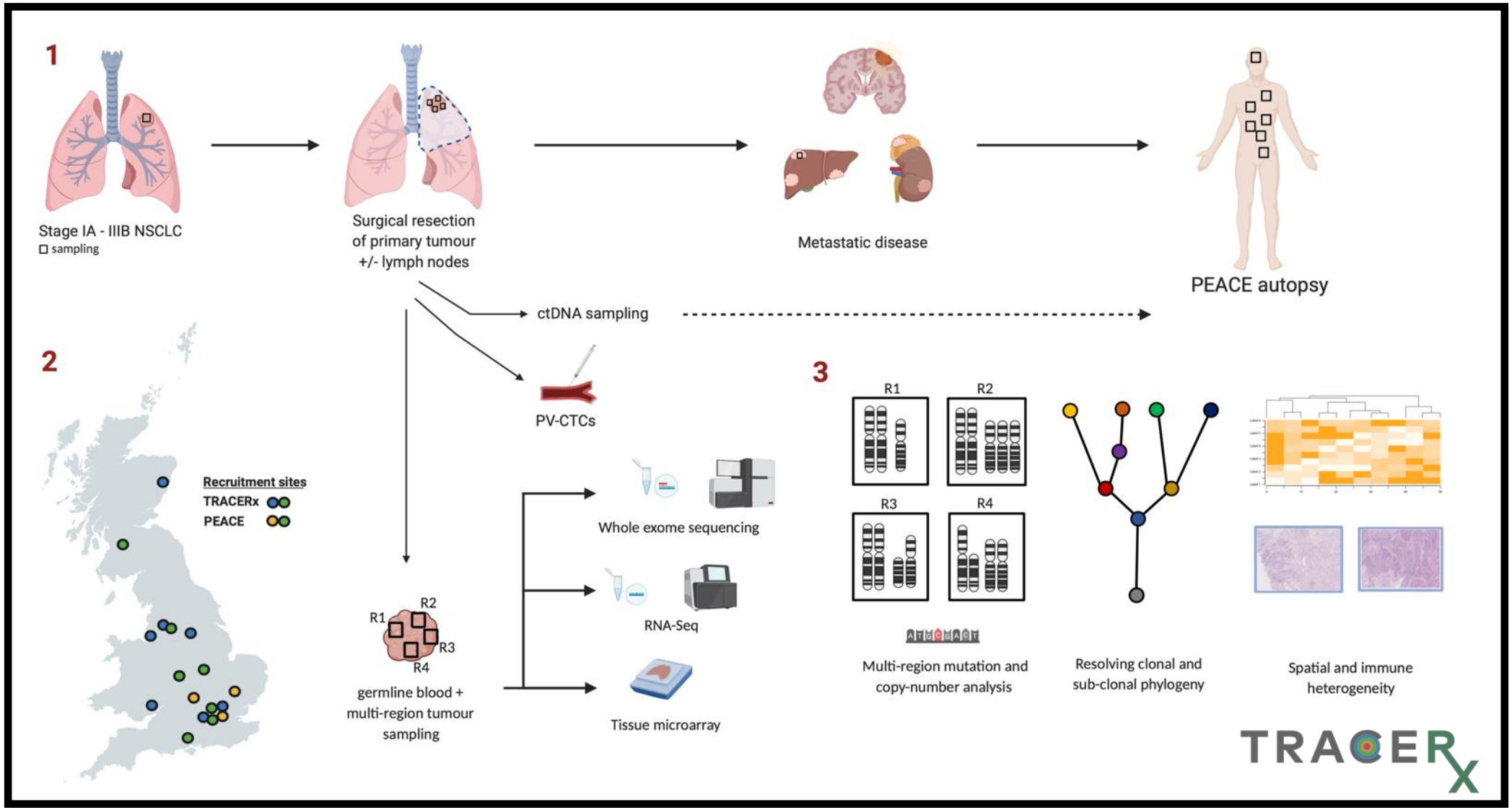
- Workloads: Genomics, data analysis, molecular modelling

CPU Research Applications

THE FRANCIS CRICK INSTITUTE

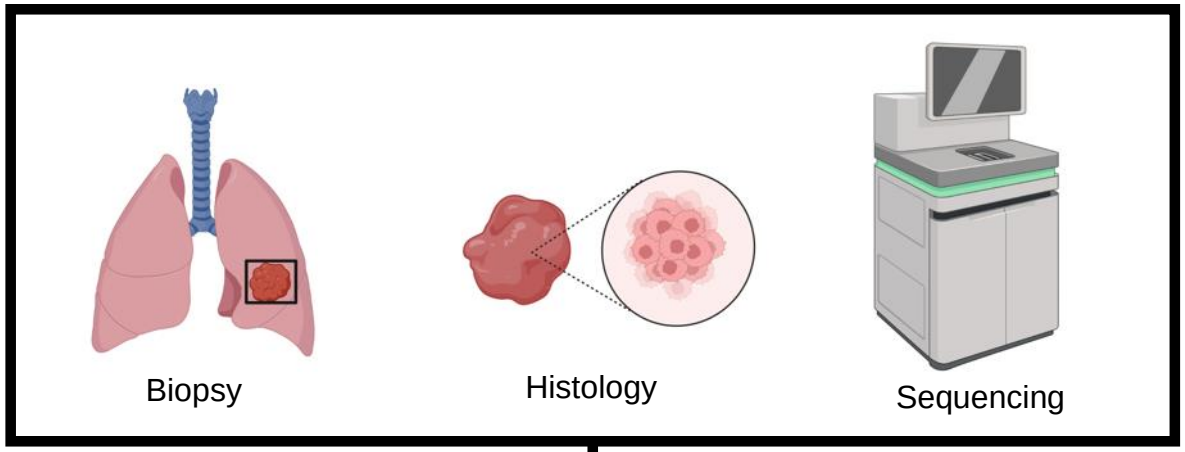
- Cancer evolution genomics
- COVID-19 variant tracking

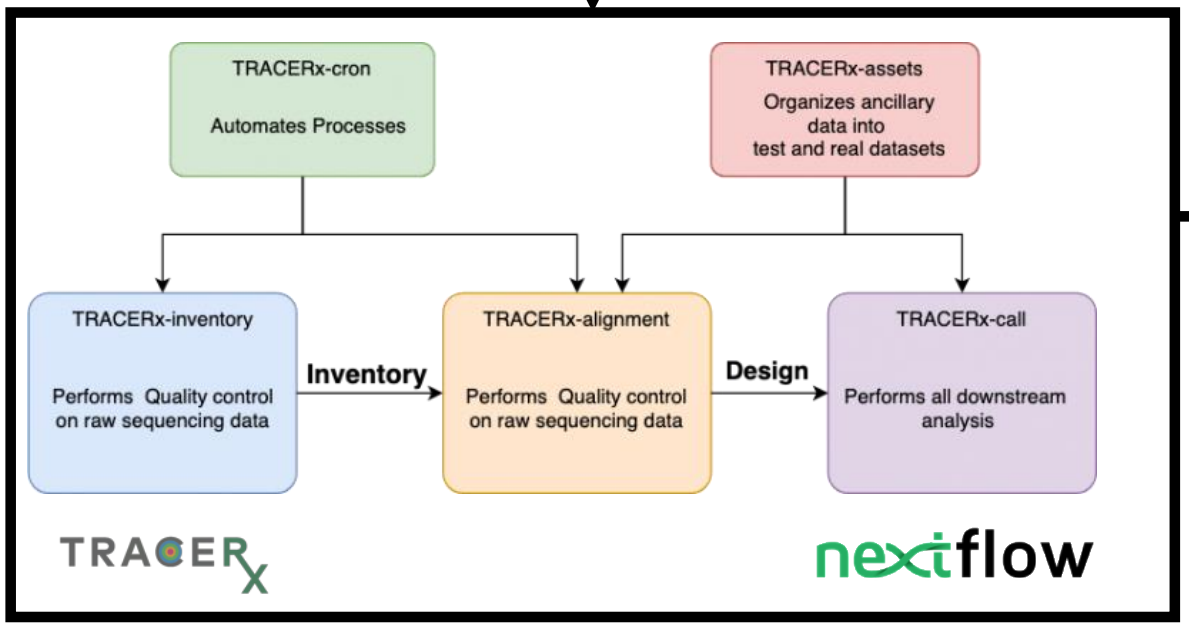
The TRACERx study aims to profile the evolutionary history of 842 patients with non-small cell lung cancer



Bailey et al 2021

The Crick's HPC facility has ensured the processing of whole exome sequencing data





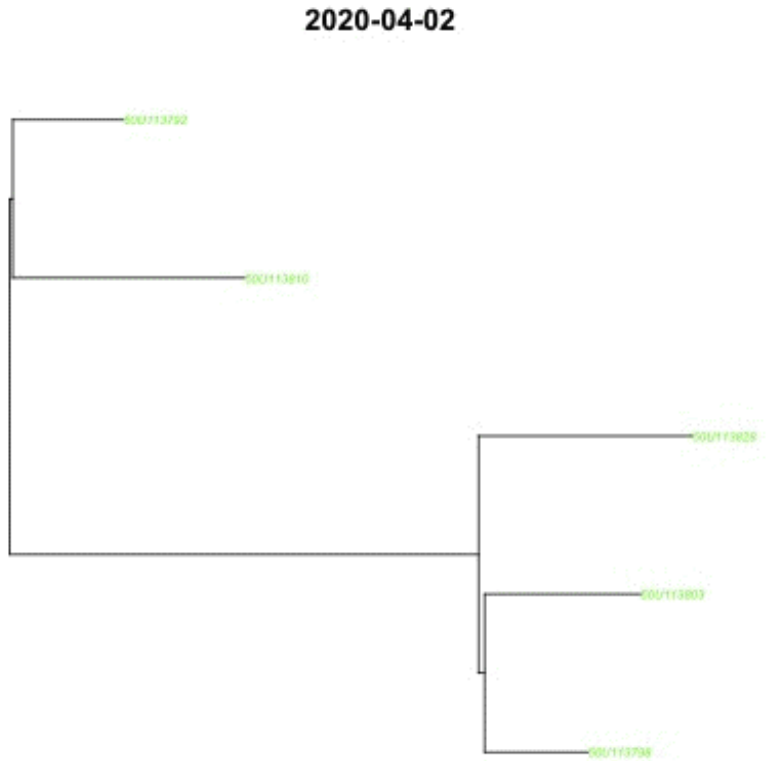
The exome makes up 1-2% of the genome
The TRACERx study has now expanded to whole genome sequencing with 357 samples processing

Jamal-Hanjani et. al 2017
Rosenthal et al 2019
Biswas et al 2019
Lopez et al 2020

COVID-19 sequencing (part of COG-UK variant tracking)

THE FRANCIS CRICK INSTITUTE

- All PCR+ tests processed by Crick were sequenced
- Sequence data processed on CPU cluster
- Animation shows new variants over time
- Data fed into COG-UK



COG-UK COVID-19 GENOMICS UK CONSORTIUM

Crick GPU clusters (2019-)

Main GPU cluster:

- 40 nodes: 4 x Nvidia V100 32 GB NVLink, 2 x 20-core Intel Skylake, 768 GB RAM

Structural Biology (cryo-EM) GPU cluster:

- 11 nodes: 4 x Nvidia RTX5000 16 GB, 2 x 20-core Intel Skylake, 348 GB RAM
- (Replaced local GPU workstations for cryo-EM)

Interactive GPU cluster:

- 5 nodes: 4 x Nvidia RTX5000 16 GB, 2 x 20-core Intel Skylake, 348 GB RAM

InfiniBand FDR + 40G Ethernet

- Workloads: cryo-EM, image processing, AI/ML

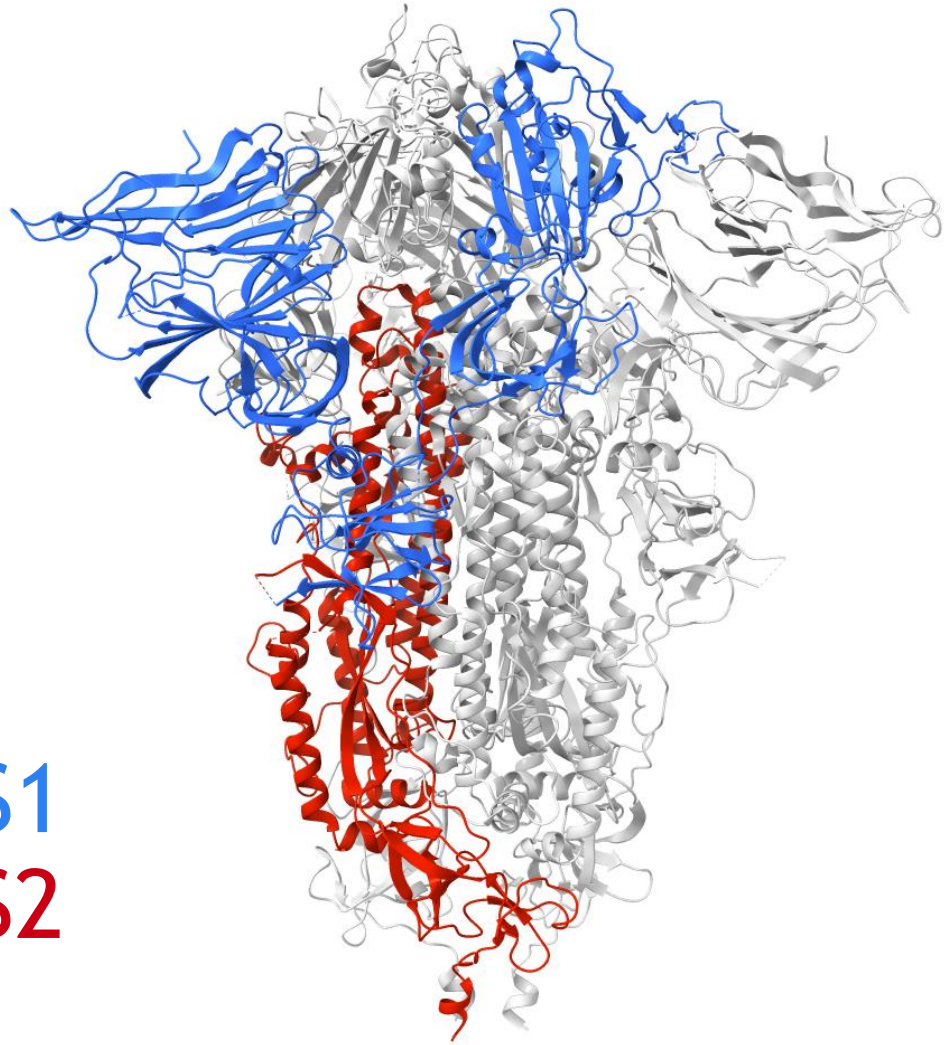
GPU Research Applications

THE FRANCIS CRICK INSTITUTE

- SARS-CoV-2 structure
- Dynamic protein structures
- Cell organelle image segmentation

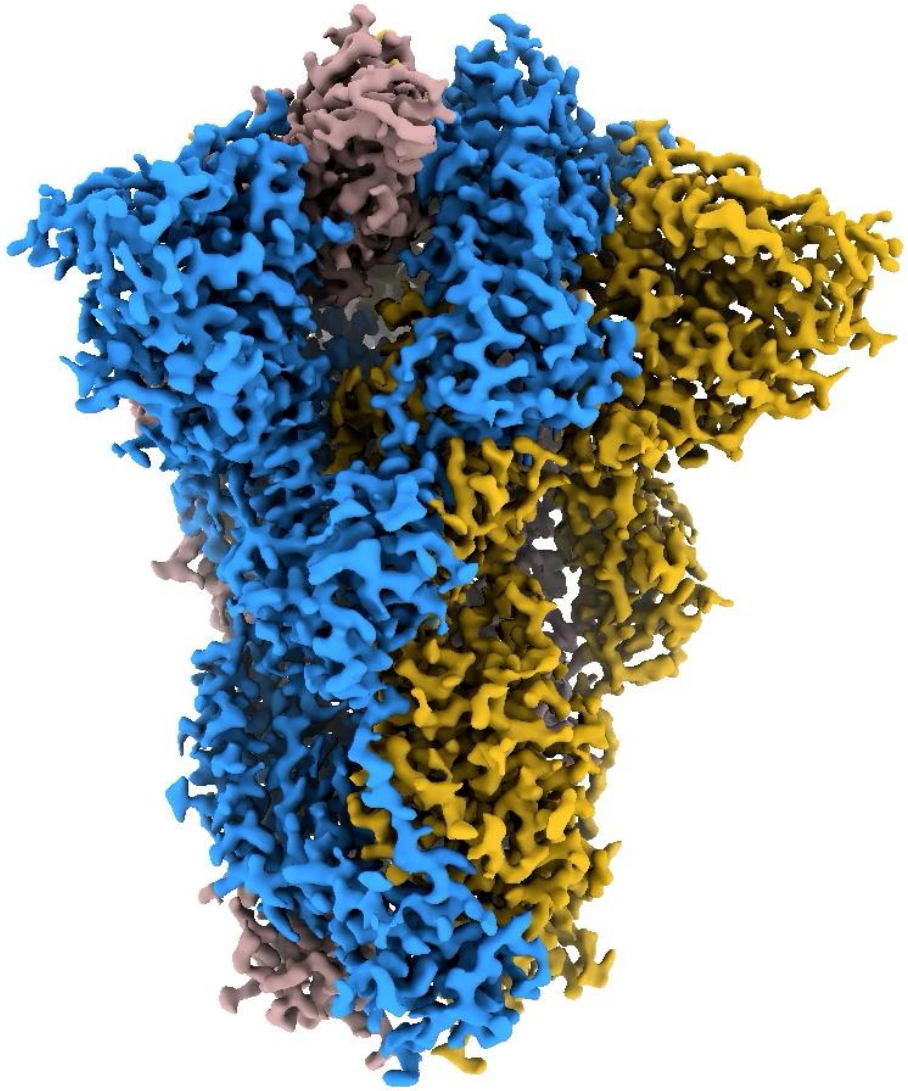
SARS-CoV-2 Spike Structure at 2.6Å

THE FRANCIS CRICK INSTITUTE



S1

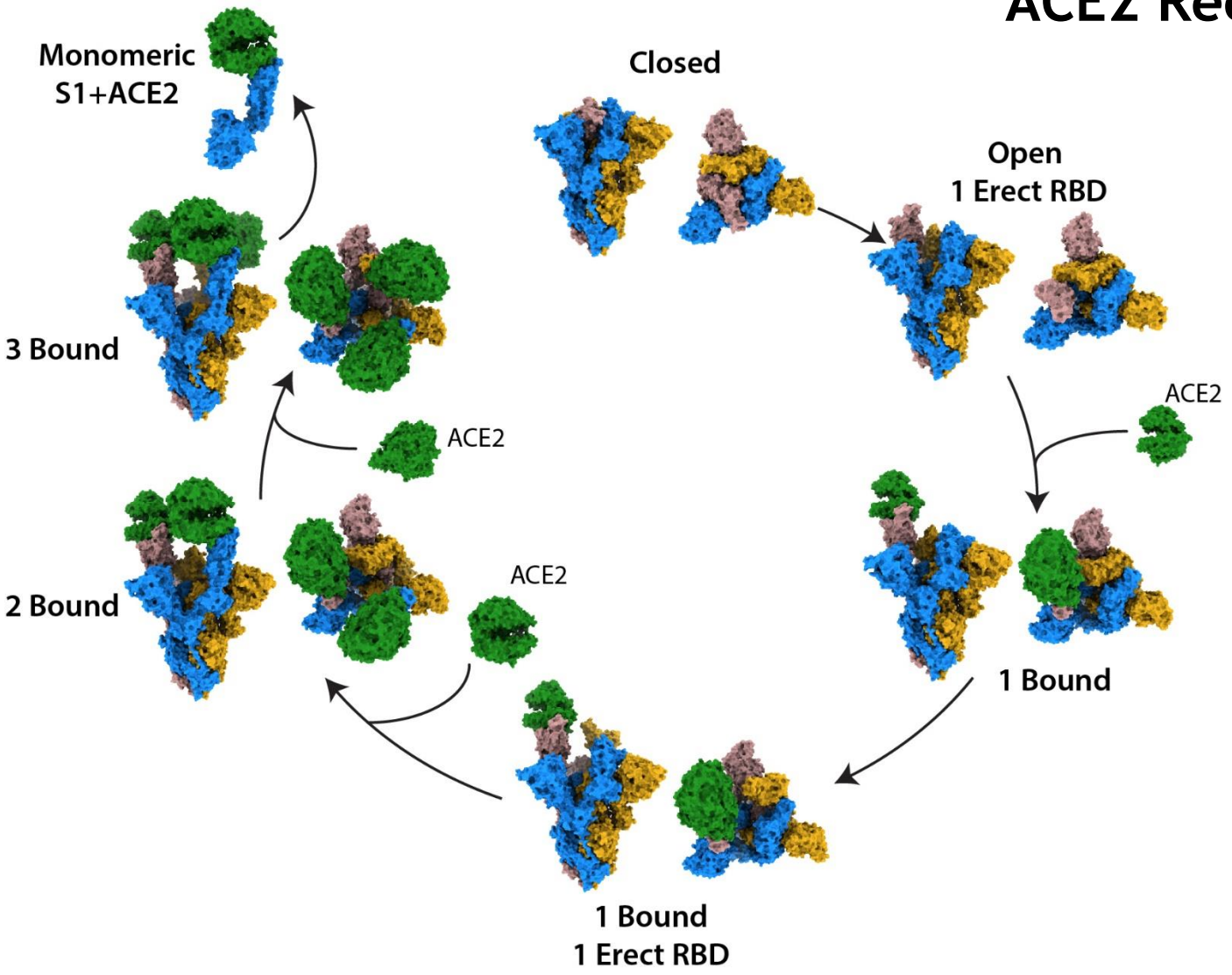
S2

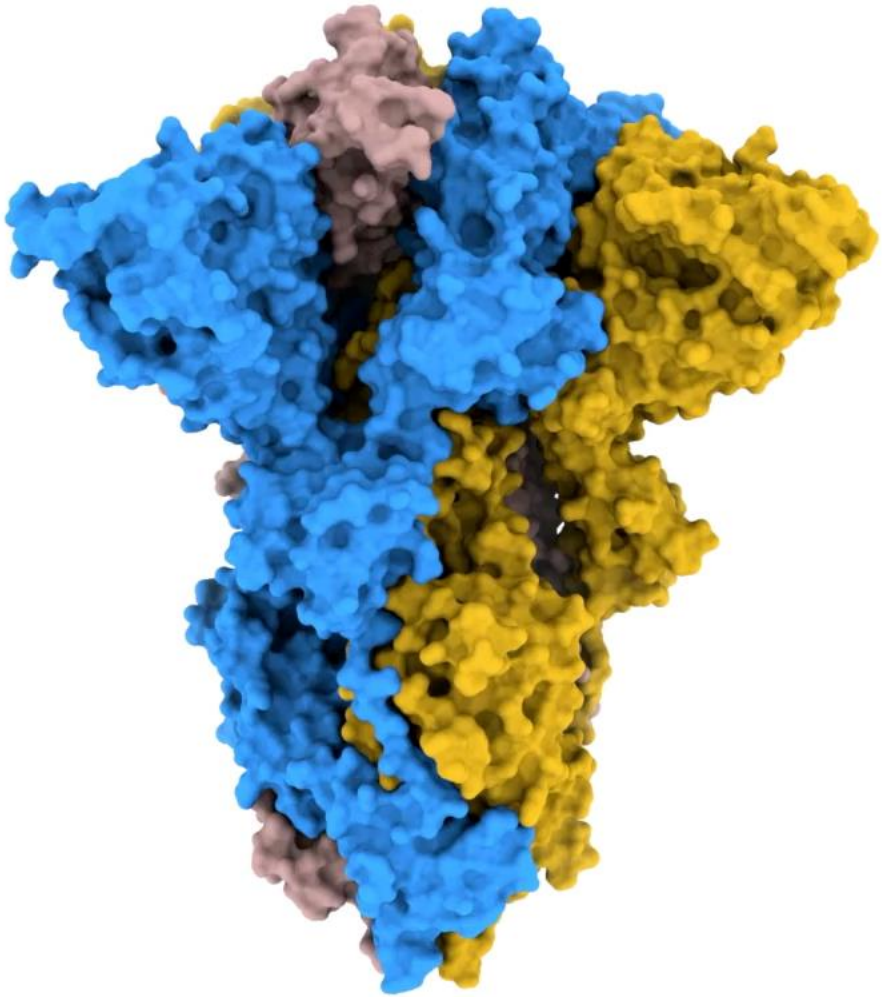


EMPIAR-10469

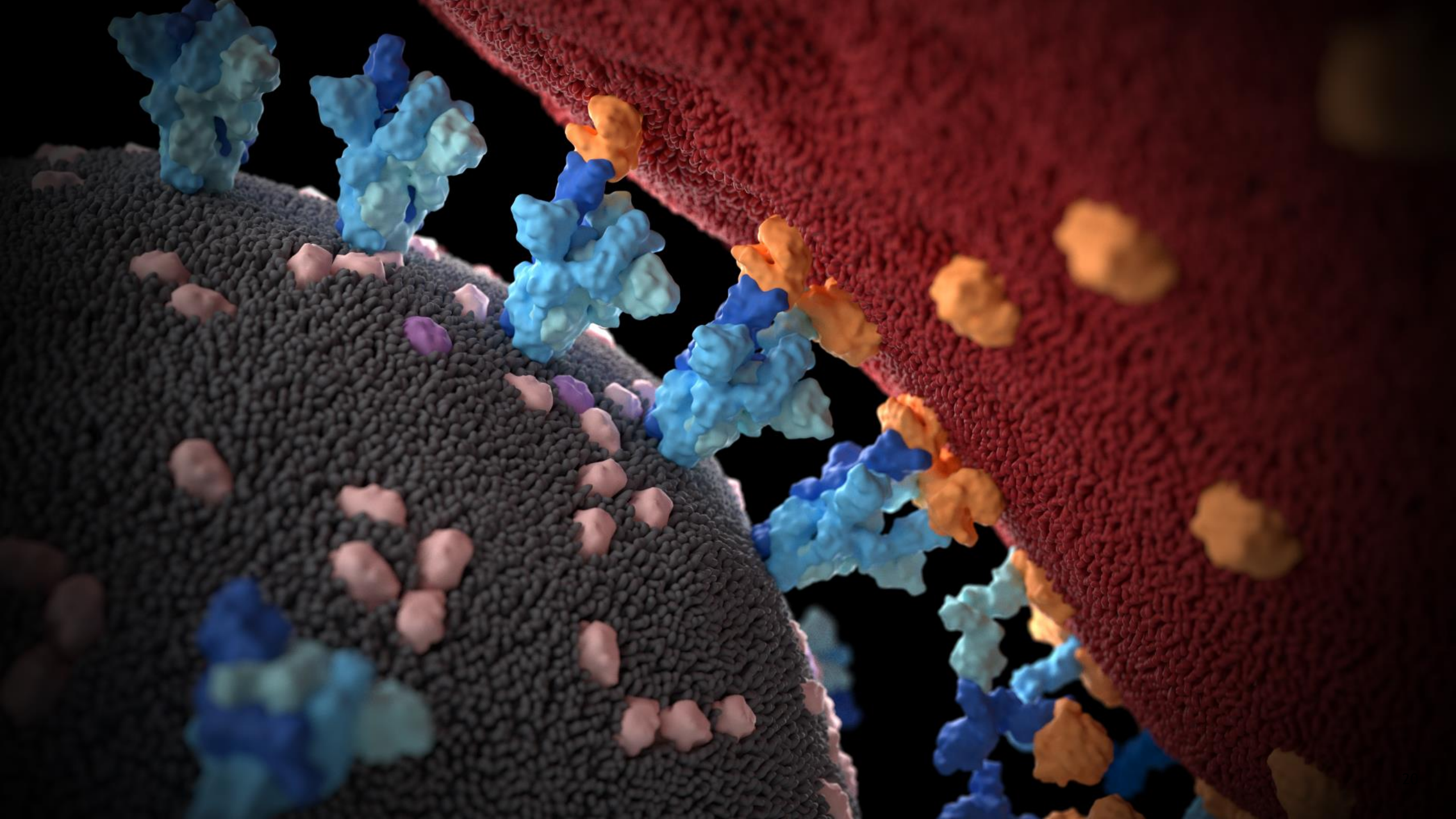
Wrobel*, Benton* et al., NSMB, 2020

ACE2 Receptor Binding to SARS-CoV-2 Spike





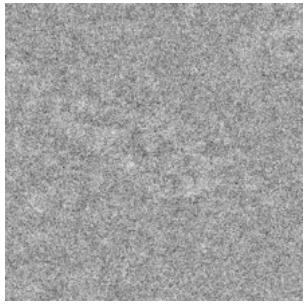
Benton et al., Nature, 2020



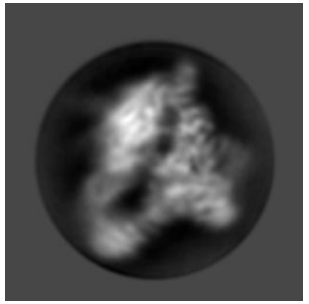
Dynamic protein structures from cryo-EM

Software solutions
cryoDRGN and others

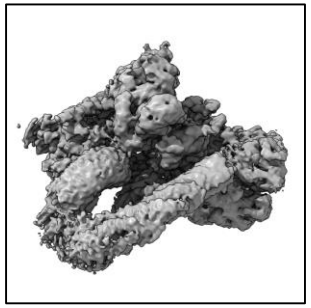
Cryo-EM data



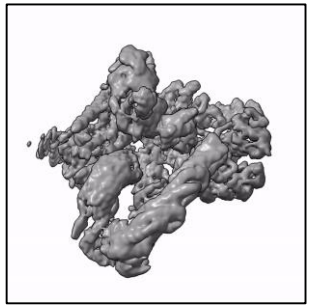
2D alignment



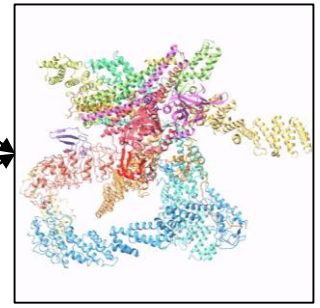
3D alignment



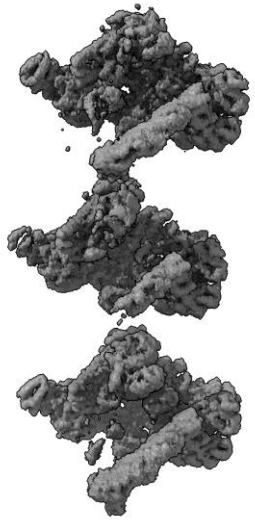
Dynamics



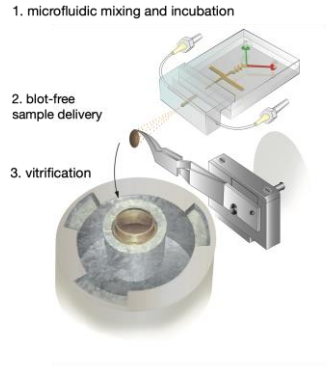
Understanding
function

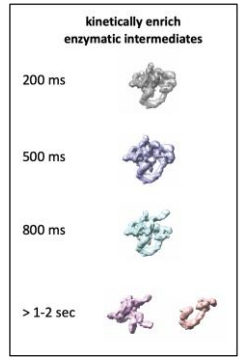


Different classes



Hardware solutions
time-resolved sample preparation





Cryo-em reconstruction
Märt-Erik Mäeots

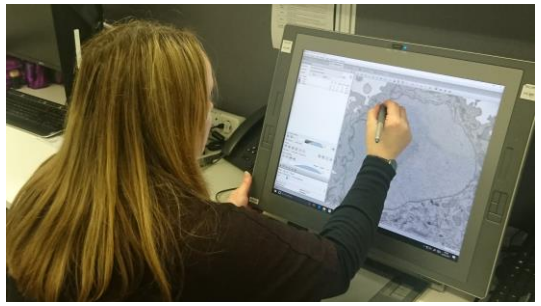
Cell image segmentation using Deep Learning

THE FRANCIS CRICK INSTITUTE

Collaboration with the Electron Microscopy core facility and the Zooniverse citizen science team – called Etch-a-cell.

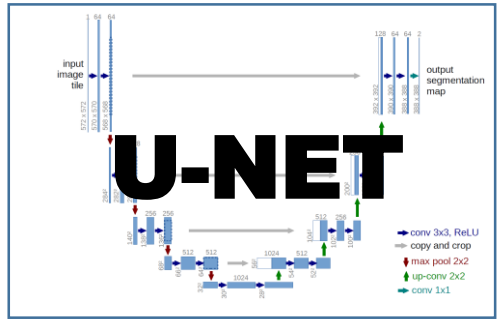
- Organelle segmentation starting with Nuclear Envelope and moving on to Mitochondria and Endoplasmic Reticulum
- Using crowd sourced annotations to train deep learning models

Expert data



Crowd-sourcing

Machine Learning

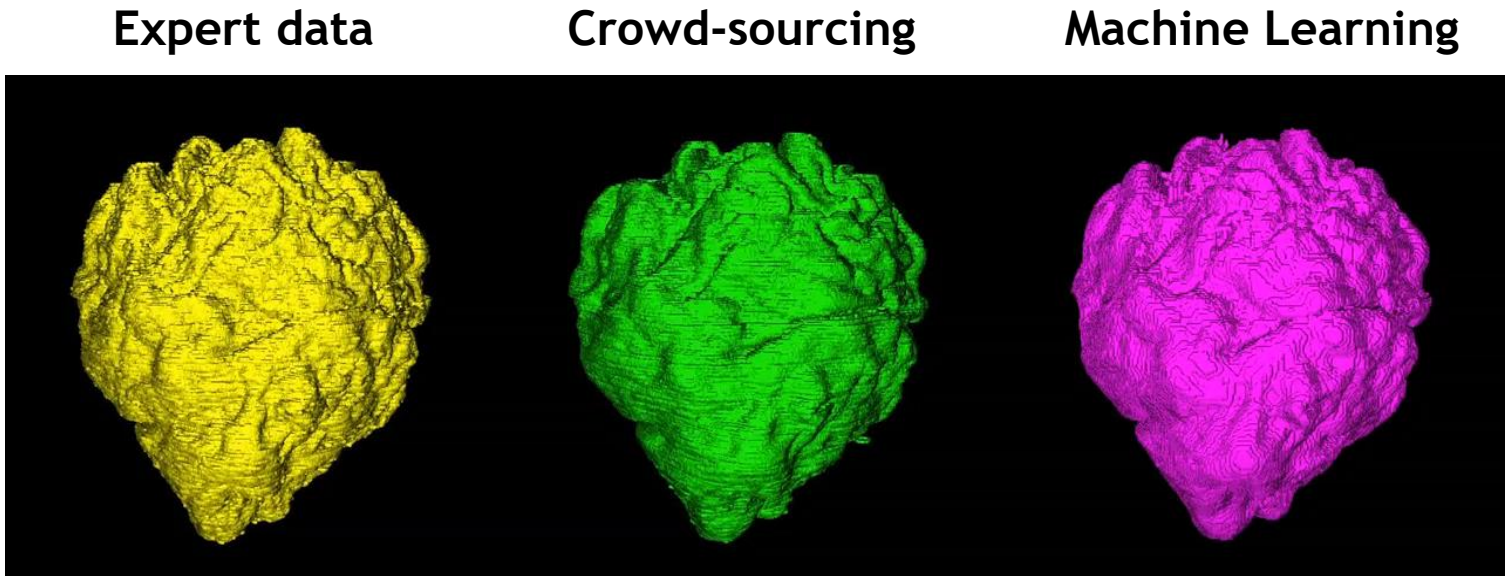


https://github.com/FrancisCrickInstitute/Etch-a-Cell-Nuclear-Envelope | https://www.zooniverse.org/

Segmentation results and next steps

THE FRANCIS CRICK INSTITUTE

- Dice score >0.95 on unseen images



Next Steps

- Future goal to improve generalization across cell and microscope types
- Strategies include training on a greater variety of training data, normalization schemes and adopting alternate model formulations

Crick research data storage 2016-

THE FRANCIS CRICK INSTITUTE

- 11 PiB DDN GRIDScaler: 2 x SFA12K + 2 x GS7K
- MEDIAScaler – NFS/SMB presentation to Mac/Win/Linux clients
- IBM GPFS/Spectrum Scale v4 > v5

OCF

DDN STORAGE

IBM Spectrum Scale

- Backbone of our research capability
- Special thanks to DDN for their excellent support and IBM for v5 licence transition

(Join the Spectrum Scale User Group tomorrow for more details!)

New Crick research data storage 2022-

THE FRANCIS CRICK INSTITUTE

- Lenovo DSS-G
- 15 PiB HDD + 1.2 PiB NVMe
- CES protocol nodes
- IBM Spectrum Scale v5
- Expandable to 30 PiB just by adding 1 PiB disk shelves
- ~9 PiB data migration using Atempo Miria

OCF

Lenovo

IBM Spectrum Scale

Other compute resources

THE FRANCIS CRICK INSTITUTE

- Cloud – AWS and GCP pilots
- Virtual GPU desktops (VMware) – e.g. iterative ML model development

The future of Crick compute is heterogeneous!

THE FRANCIS CRICK INSTITUTE

- Procurement planned in 2022 to replace CPU cluster (EOS June 2023)
- CPU + GPU? + custom ASICs?
- Driven by workloads
- Easy access to cloud (software and networking)
- ‘Spectrum of compute’ for researchers
- Containerisation/virtualisation support
- Software developers/engineers needed!
- GPU cluster EOS 2024
- Quantum?

One size does not fit all… workload specific

Not just HPC - the whole picture

THE FRANCIS CRICK INSTITUTE

HPC

People, Support & Expertise

Relationships

Scientific Computing core facility

Providing Crick researchers with advanced scientific computing platforms, support and skills to deliver discoveries to change lives

We provide a broad range of support for scientific computing across 3 teams, total 20 staff:

- Research Data Services / Database Team
- Software Development & Machine Learning Team
- Research Computing Platforms/HPC

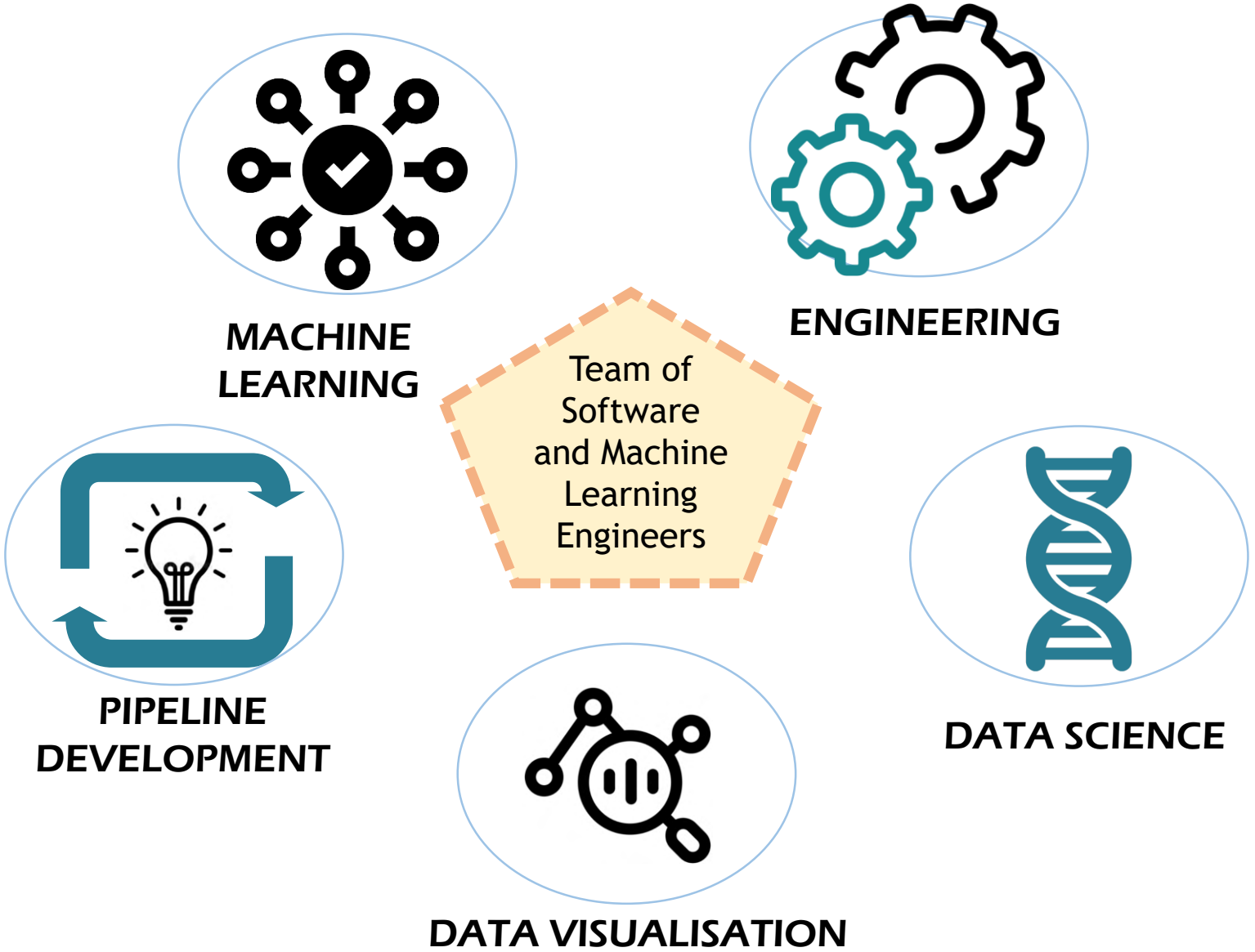
Karen Ambrose
Research Data Services

Amy Strange
Software Development & Machine Learning

Wei Xing
Research Computing Platforms/HPC

Software Development and Machine Learning

THE FRANCIS CRICK INSTITUTE



My own next move…

NBI Partnership

Healthy Plants, Healthy People, Healthy Planet

Hiring in London and Norwich…

Crick HPC & Research Data Systems Engineer

https://www.crick.ac.uk/careers-study/vacancies/2021-10-27-hpc-research-data-systems-engineer

NBI Research Computing Junior Sys Admin

https://jobs.nbi.ac.uk/Details.asp?vacancyID=16662

Talk to me/get in touch:

Steve.Hindmarsh@crick.ac.uk

crick.ac.uk

Phil Hasnip (University of York)

Portable acceleration of materials modelling software: CASTEP, GPUs and OpenACC

Abstract: CASTEP is a leading first-principles materials modelling program, which uses quantum mechanics to predict the chemical, electronic and physical properties of materials. CASTEP is parallelised using OpenMP and MPI, and is widely used on HPC facilities, including ARCHER2 where it typically consumes 5% of monthly CPU cycles. In this presentation we will present work to extend CASTEP to exploit heterogeneous architectures, in particular GPGPUs, using OpenACC. We discuss the challenges and opportunities presented by accelerator-based architectures, and the approach taken in the CASTEP OpenACC port. Whilst the port is still under active development, early performance results show that significant speed-ups may be gained, particularly for materials simulations using so-called “non-local functionals,” where speed-ups can exceed a factor of ten.

Bio: Phil Hasnip is a physicist and computer programmer in the Department of Physics at the University of York. He grew up in the 1980s, where he learned physics at school and computer programming on his Sinclair ZX Spectrum. He is an EPSRC Research Software Engineering Fellow, and writes computer software to tackle problems in physics research, with a particular interest in making scientific software user-friendly, scalable, efficient and reliable. Phil is a lead developer of the quantum mechanical materials modelling program CASTEP, and is also on the "Materials and Molecular Modelling" working group for the UK's ExCALIBUR exascale HPC programme, working to ensure the key UK materials modelling methods are ready for the next generation of HPC machines.

Dr Igor Baratta (Department of Engineering, University of Cambridge)

Heterogeneous Programming for Finite Element: insights from benchmarks

Abstract: Finite Element software packages have struggled to get good performance from GPUs and accelerators in the past. Partly, this was due to the limited memory capacity of earlier devices, but it is also essential to consider the memory layout, the details of the kernel code, and where it runs.

This presentation will consider the performance of the finite element method on different architectures and accelerators using the FEniCS finite element libraries and the SYCL programming model. SYCL is a modern kernel-based parallel programming model that allows one code to be written, which can run on multiple types of computational devices.

This kernel-based model matches nicely with the latest FEniCS data-centric design: the top-level C++ library creates data to operate in parallel (geometry, topology, and dof layout information) and an automatic code generator emits efficient code that can be used as part of the kernel.

Finite element codes have several parts, often subdivided into: mesh and domain handling, formulation of the equations of interest, assembly and linear solve. The linear solve is the most computationally intensive, and for all problems of size worth consideration on HPC systems, a direct solve is not scalable. Iterative solvers are a necessity, and the cost per iteration is strongly affected by the cost of memory access. Various strategies can be used to minimise data movement, and this often means considering some merging of the assemble and solve phases. We will demonstrate some of the approaches we have taken with FEniCS and SYCL to obtain the best performance from the currently available hardware.

We will also discuss how different ways of expressing parallelism can affect the performance of finite element code on heterogeneous architectures. We will consider how arranging memory transfer and allocations can reduce latency and increase throughput in different accelerators. Finally, we will show some performance results from our Excalibur Project (Excalibur-SLE) using our code in the well-known CEED benchmarks from the ECP programme. We will present results for several architectures, including Intel Ice Lake CPU and NVIDIA A100 GPU.

Bio: Igor Baratta is a Research Associate in Scientific Computing at the University of Cambridge. He completed his PhD and undergraduate degrees in Electrical and Computer Engineering at UFMG in Brazil. Prior to completing his PhD, he was an R&D Engineer working on computational electromagnetics at a multinational aerospace company.

Heterogeneous Programming for Finite Element Methods:

Insights from benchmarks

Igor Baratta, Chris Richardson and Garth N. Wells
Department of Engineering

Outline

Why heterogeneous computing?

Finite Element Overview

Benchmarks

Results

Where to go from here?

Compute variety at scale

TOP 500
The List.

Supercomputer Fugaku	Summit	Sierra	Sunway TaihuLight	Perlmutter
				
A64FX 48C 2.2GHz	IBM POWER9 22C 3.07GHz NVIDIA Volta GV100	IBM POWER9 22C 3.1GHz NVIDIA Volta GV100	Sunway SW26010 260C 1.45GHz, Sunway	AMD EPYC 7763 64C 2.45GHz NVIDIA A100 SXM4 40 GB
Rmax (TFlop/s) 442,010.0	Rmax (TFlop/s) 148,600.0	Rmax (TFlop/s) 94,640.0	Rmax (TFlop/s) 93,014.6	Rmax (TFlop/s) 70,870.0

Future exascale systems

System	Delivery	CPU	Accelerator
Tianhe-3	2021	Phytium	Phytium
Frontier	2021?	AMD	NVIDIA
Aurora	2022	Intel	Intel
El Capitan	2023	AMD	AMD

- CPU + Accelerator
- Different vendors
- Complex software stack

TIER 2 HPC

CSD3 is one Tier 2 National HPC Services hosted by the Research Computing Services at the University of Cambridge.

CPU Nodes

Cascade Lake - 2x Intel(R) Xeon(R) Platinum 8276

Ice Lake - 2x Intel(R) Xeon(R) Platinum 8368Q

GPU Nodes

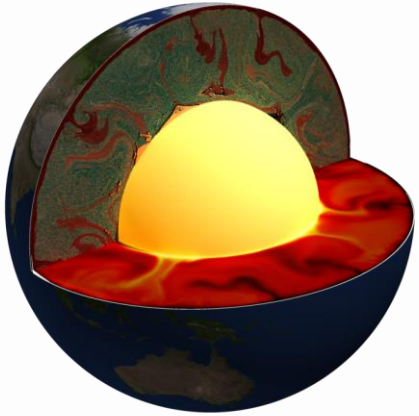
1x Intel Xeon E5-2650 12-Core + 4x NVIDIA P100 16GB

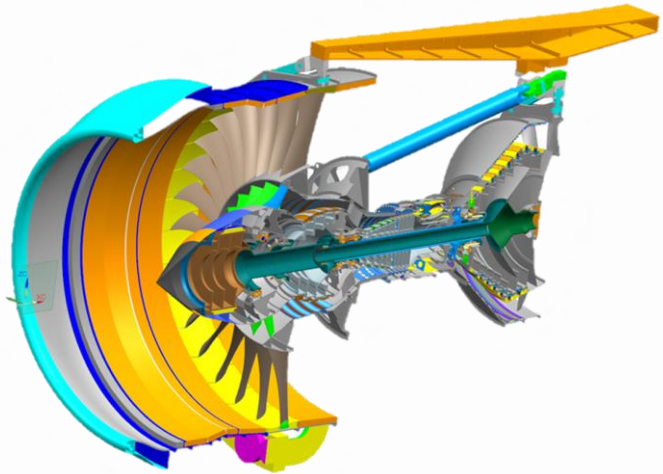
2x AMD EPYC 7763 64-Core + 4x NVIDIA A100-SXM-80GB GPUs

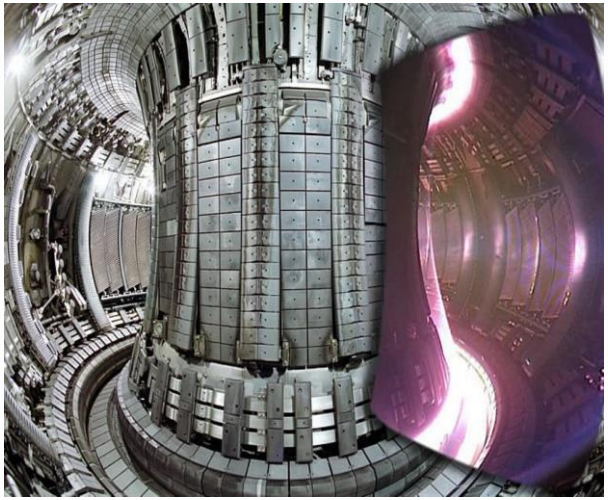
\+ Fujitsu A64FX, AMD MI100, Intel GPUs

Finite element overview

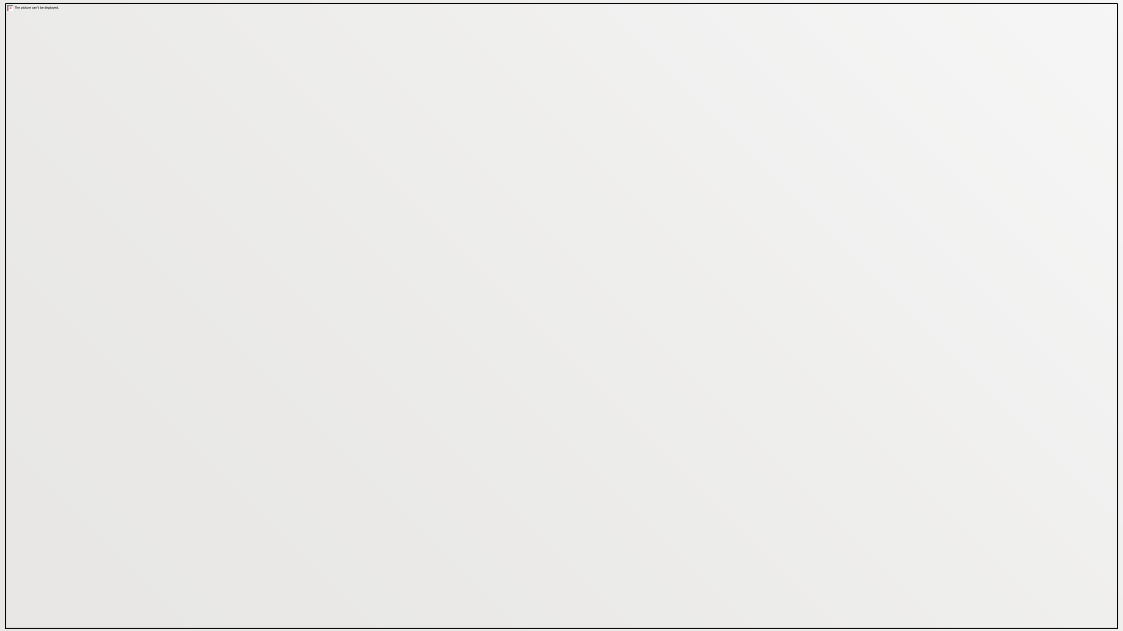
FENICS PROJECT







Finite element overview



Linear Algebra Backend

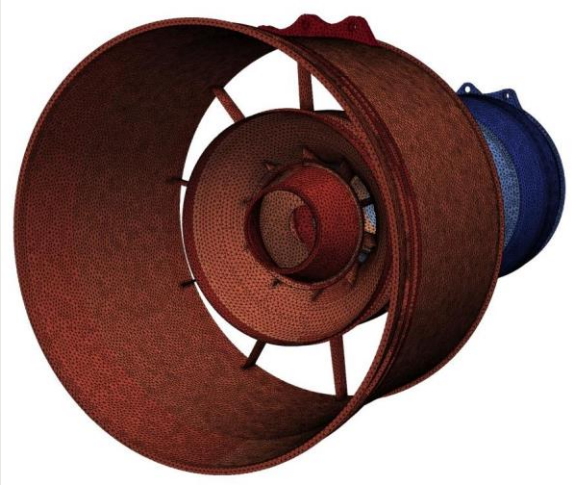
PETSc

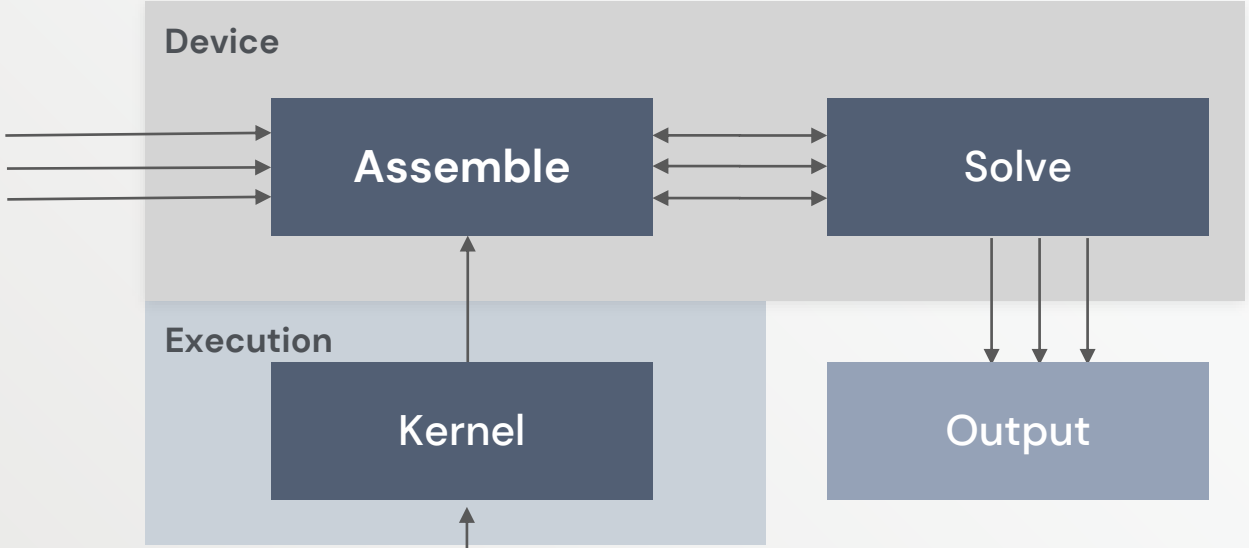
TRILINOS

Ginkgo

AmgX

Modular finite element method





$$-\nabla \cdot (\mu \nabla u) + \gamma u = f$$

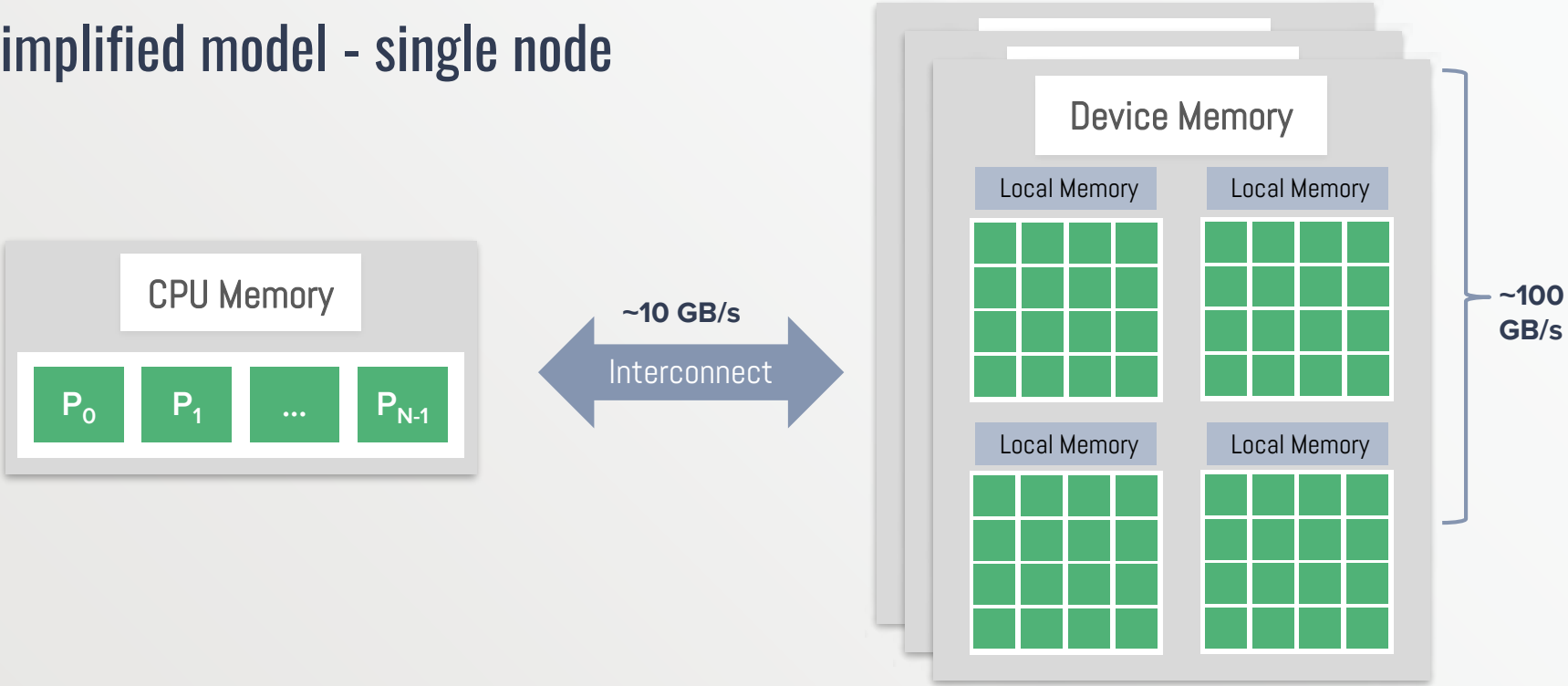
```
element = FiniteElement("Lagrange", tetrahedron, 3)
...
a = mu*inner(grad(u), grad(v)) * dx + gamma*inner(u, v) * dx
L = inner(f, v) * dx
```

Programming model

SYCL is a high-level single source parallel programming model, that can target a range of heterogeneous platforms:

- uses completely standard C++;
- both host CPU and device code can be written in the same C++ source file;
- open standard coordinated by the Khronos group.

Simplified model - single node



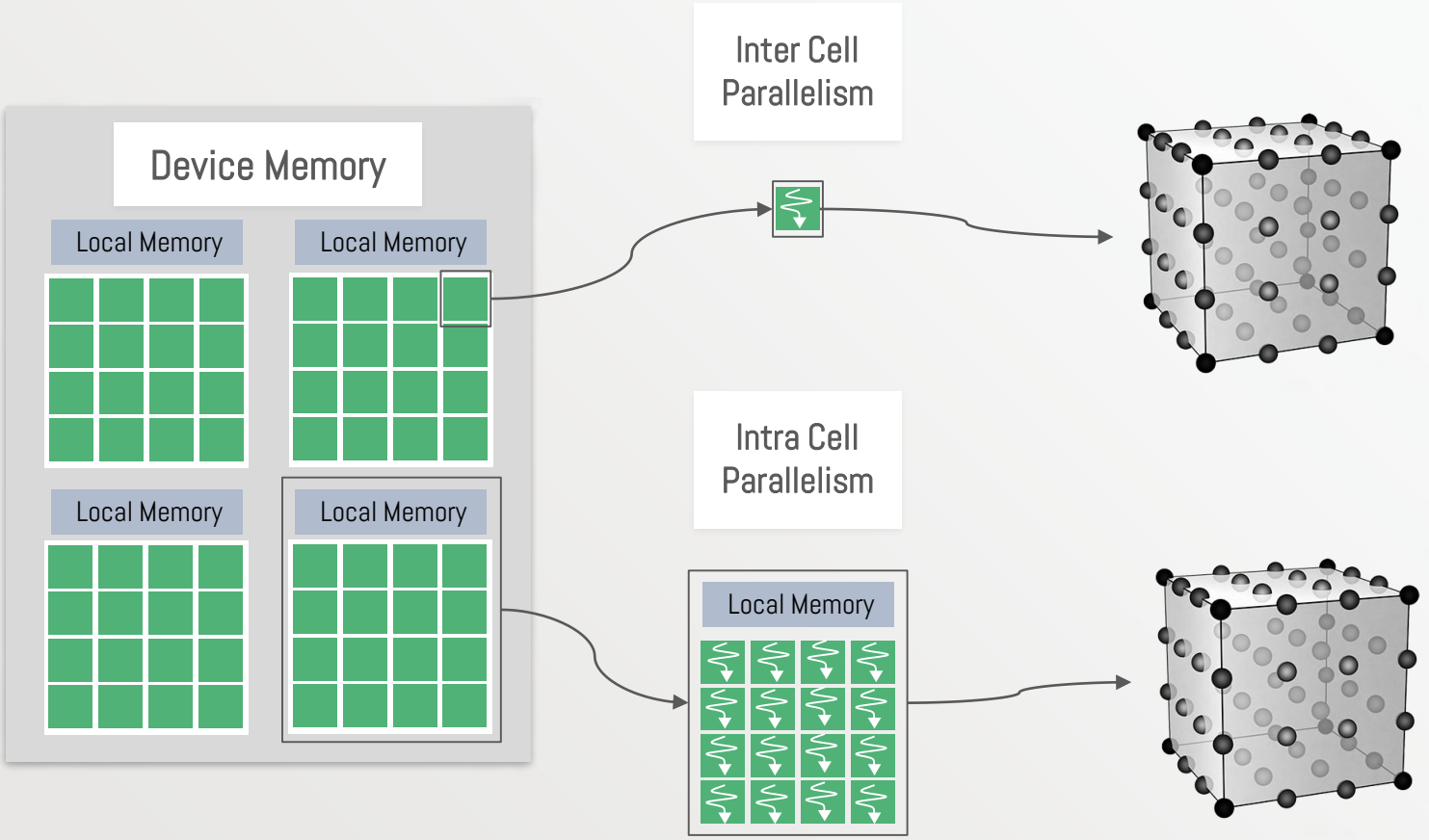
Copy input data from **Host** memory to **Device** memory

Launch kernels for execution on the **Device**

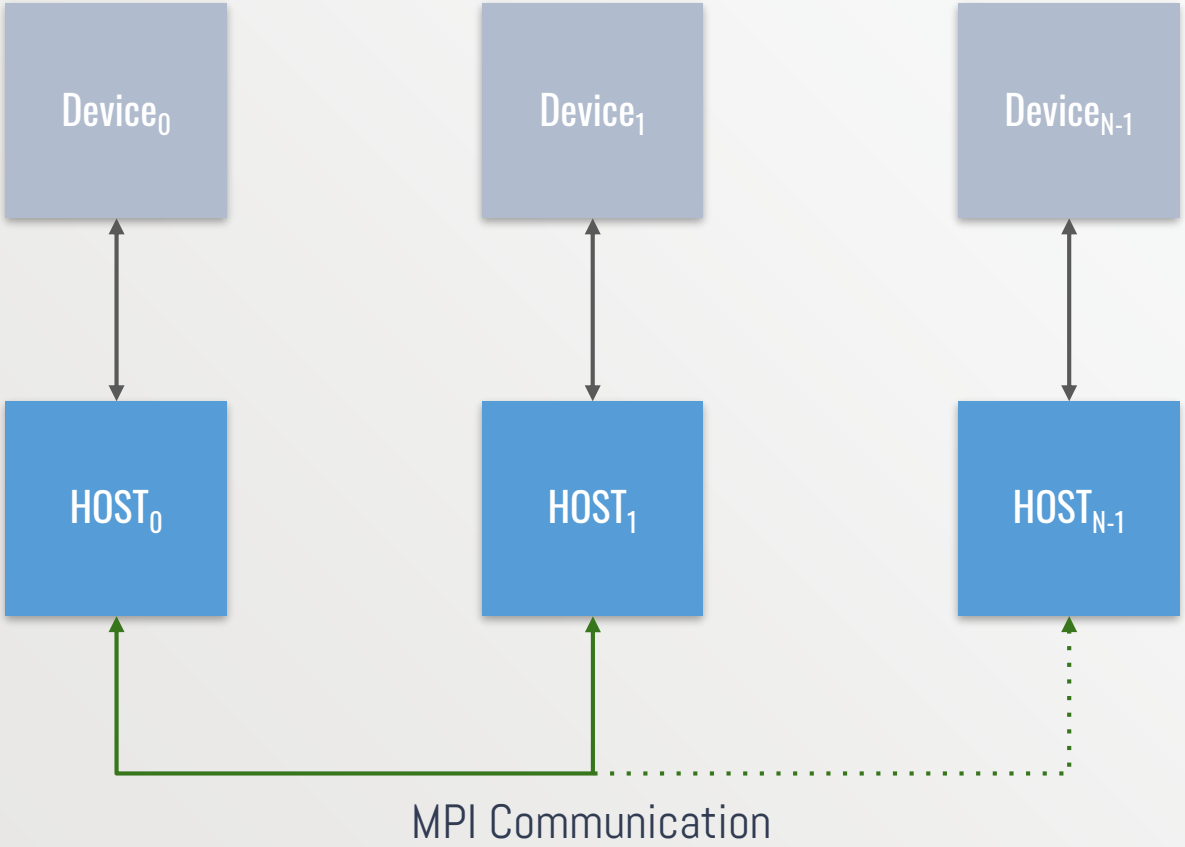
Wait for the execution queue to finish

Copy results back to **Host** from **Device**

Parallelization strategies



Communication among devices



Ceed Benchmarks

https://ceed.exascaleproject.org/bps/
Benchmarks, results, codes

What can we learn by comparison of kernel operations in other codes?

MFEM	libCEED	Nek
https://github.com/mfem	https://github.com/CEED/libCEED	https://github.com/Nek5000

Benchmark description

Bake-off problems: high-order kernels/benchmarks designed to test and compare the performance of high-order codes.

$$-\nabla \cdot (\mu \nabla u) + \gamma u = f$$

$$H\mathbf{u} = A\mathbf{u} + B\mathbf{u} = \mathbf{b}$$

Where $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}$ are non-negative functions of x.

Benchmark	Parameters	Solver
BP1 (Mass)	$\boldsymbol{\mu}$=0, $\boldsymbol{\gamma}$=1	(P)CG
BP3 (Stiffness)	$\boldsymbol{\mu}$=1, $\boldsymbol{\gamma}$=0	(P)CG

CG Solver benchmark: kernels

Memory Bound

Memory or Compute Bound

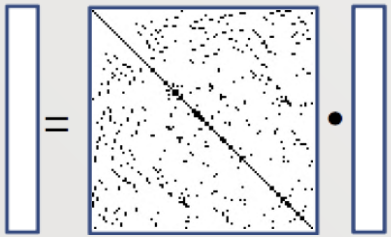
Name	Expression
Vector Copy	$\mathbf{y} = \mathbf{x}$
Vector axpy	$\mathbf{y} = \alpha\mathbf{x} + \mathbf{y}$
Vector Inner Product	$\alpha = \mathbf{x} \cdot \mathbf{y}$
Gather	$\mathbf{x} = Z^T\mathbf{x}^K$
Scatter	$\mathbf{x}^K = Z\mathbf{x}$
Matrix Vector Product	$\mathbf{y} = A \cdot \mathbf{y}$

BLAS, oneMKL, cuBLAS

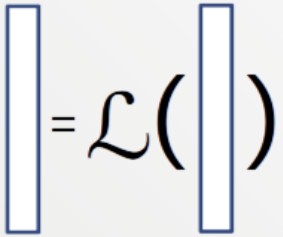
FEM kernel

Matrix-Vector Product Action

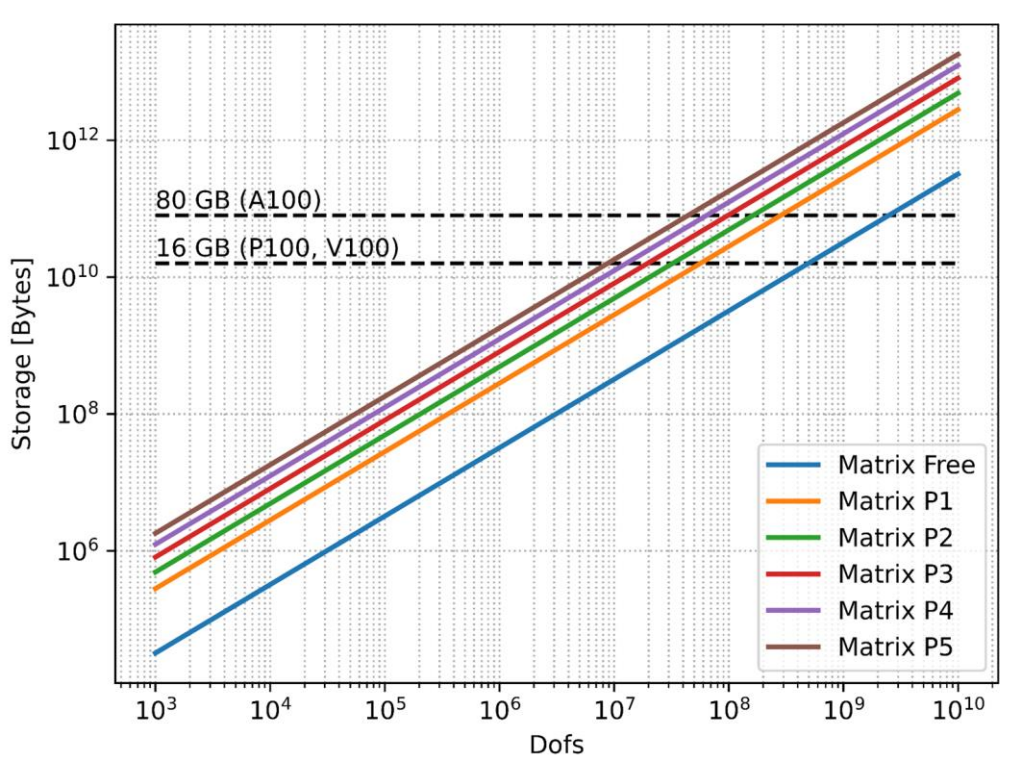
"Full assembly"



"Matrix free"



Matrix-free vs Standard sparse matrices [Local Storage]



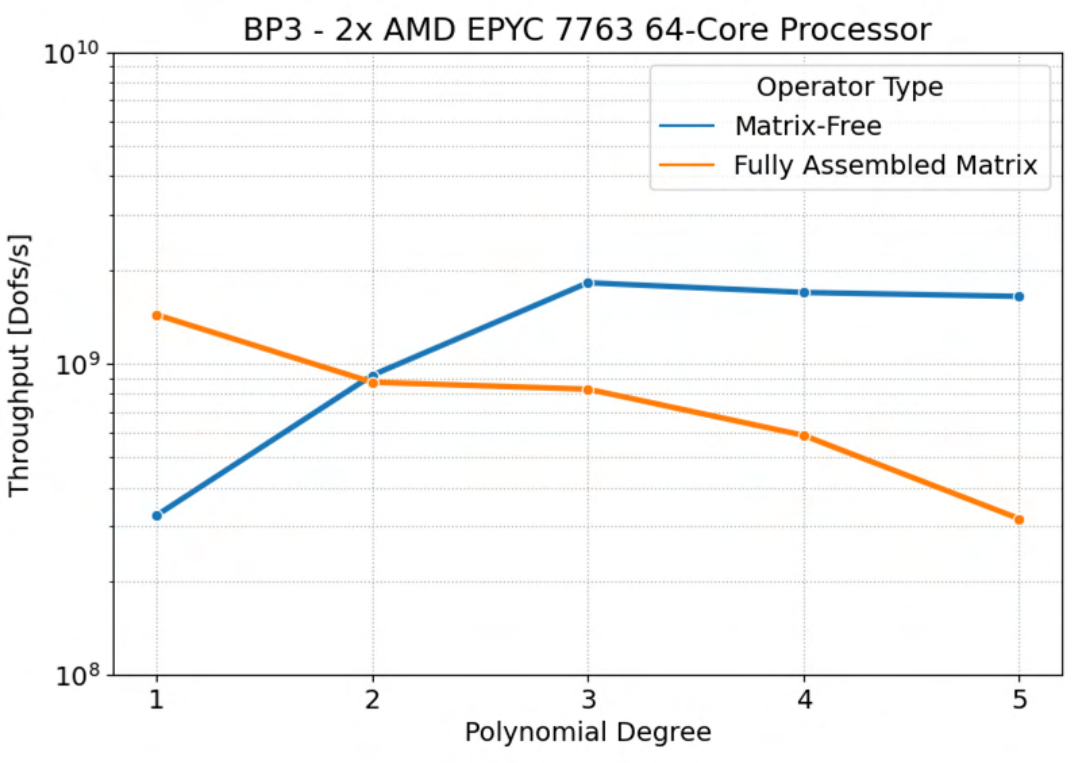
Vectors

ndofs * 8 bytes (double)

CSR Matrix

nnz * 8 bytes (double)
nnz * 4 bytes (int32)
ndofs * 4 bytes (int32)

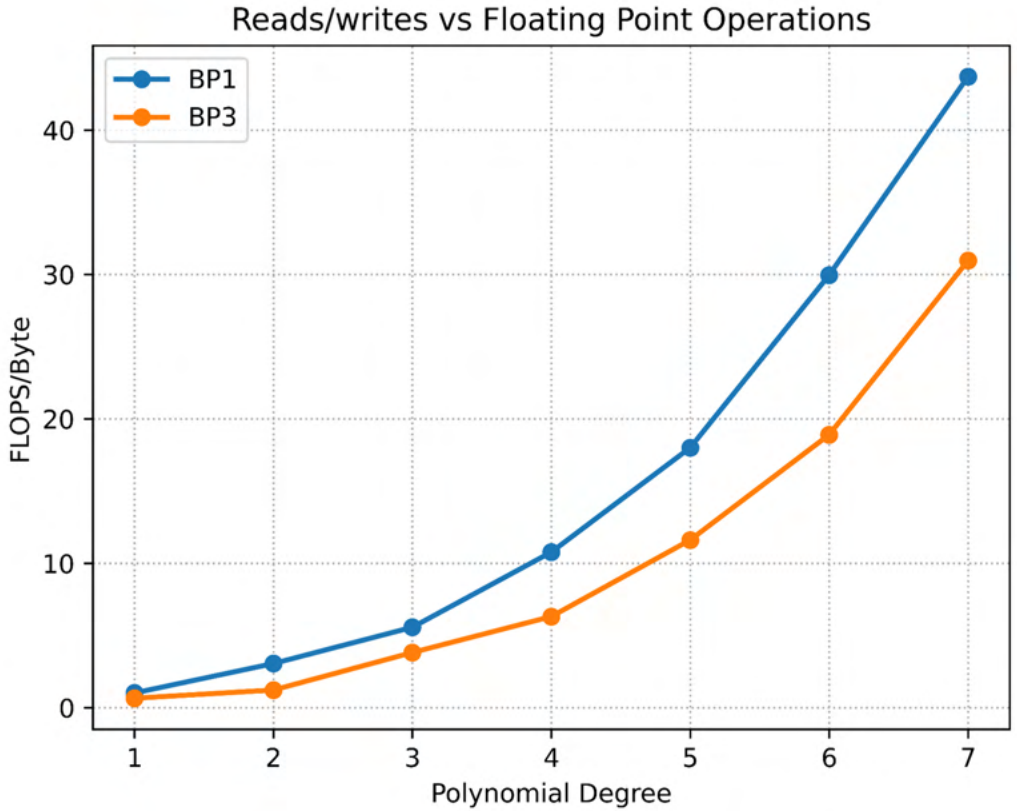
Matrix-free Vs Standard sparse matrices [Throughput]



Matrix Free require only $O(n)$ data movement and storage.

Fully assembled matrix performance falls rapidly at high orders while matrix-free operators scale well.

Floating-point operations (FLOPs) per memory access



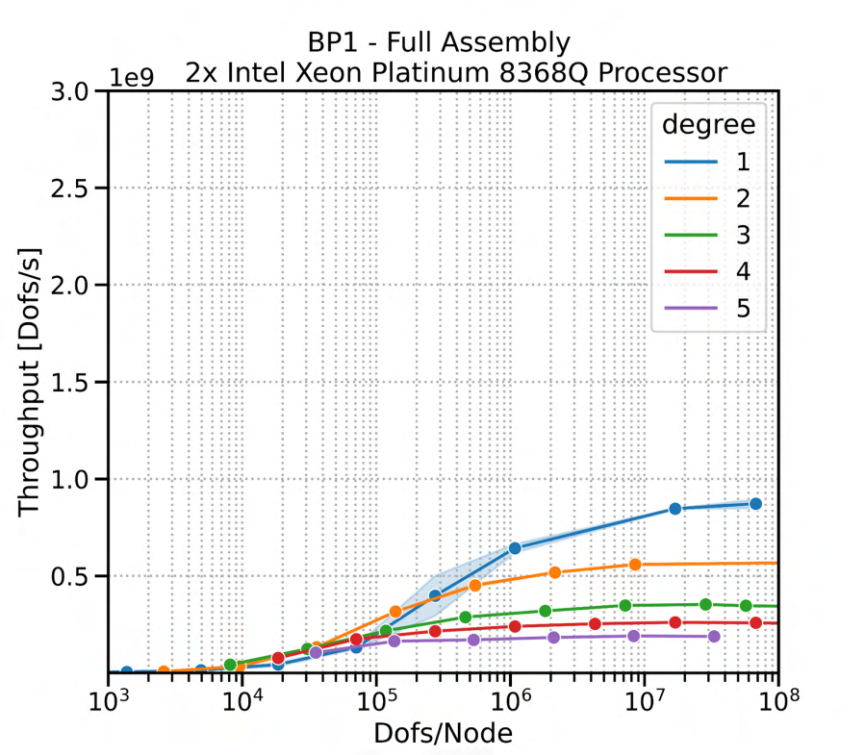
Historically, FLOPs were the performance-limiting factor.

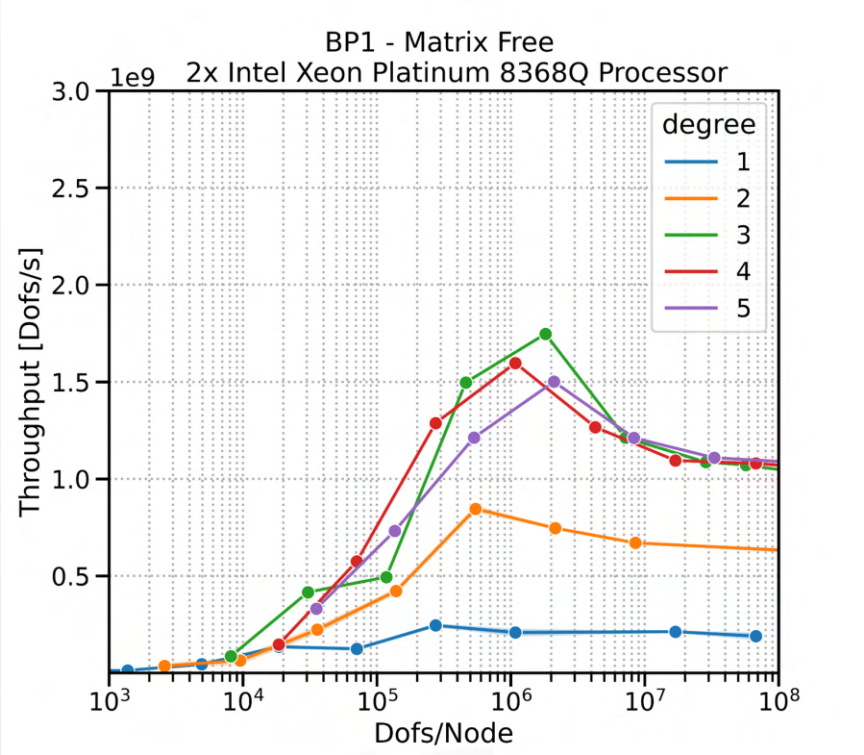
Gap between computer processing speed and memory access.

High order methods:

- Control over arithmetic intensity.

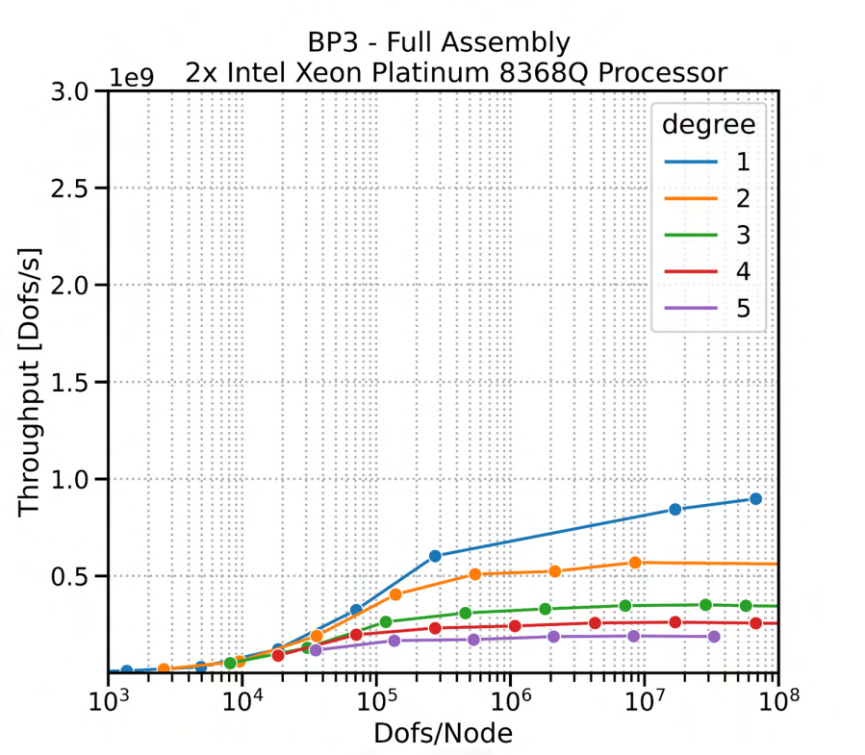
BP1 - Ice Lake node

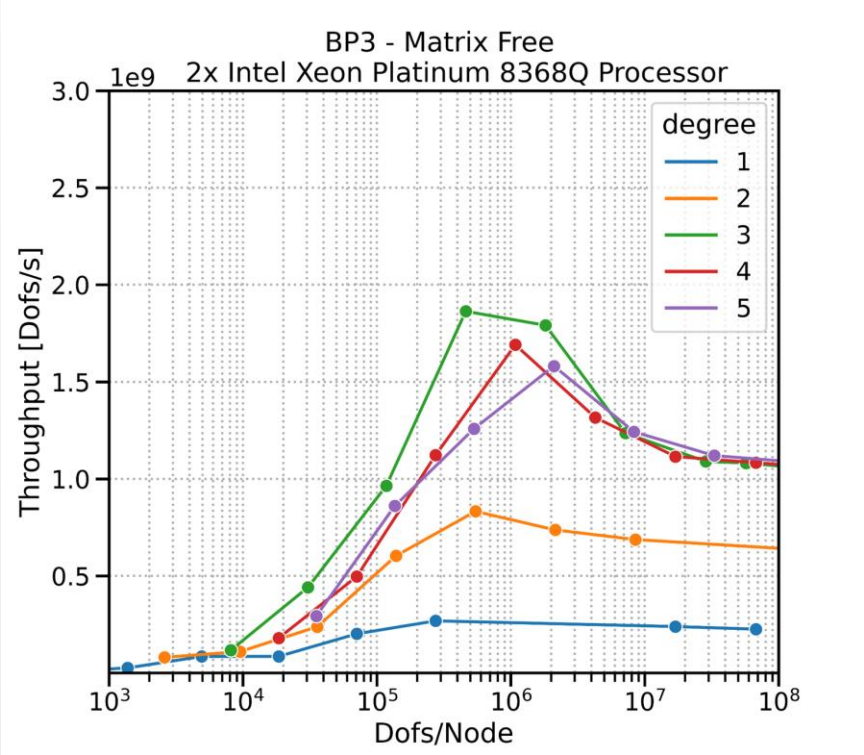




- Local Performance vs Local Problem Size
- Peak* Performance: 1.1 TDOF/s per node

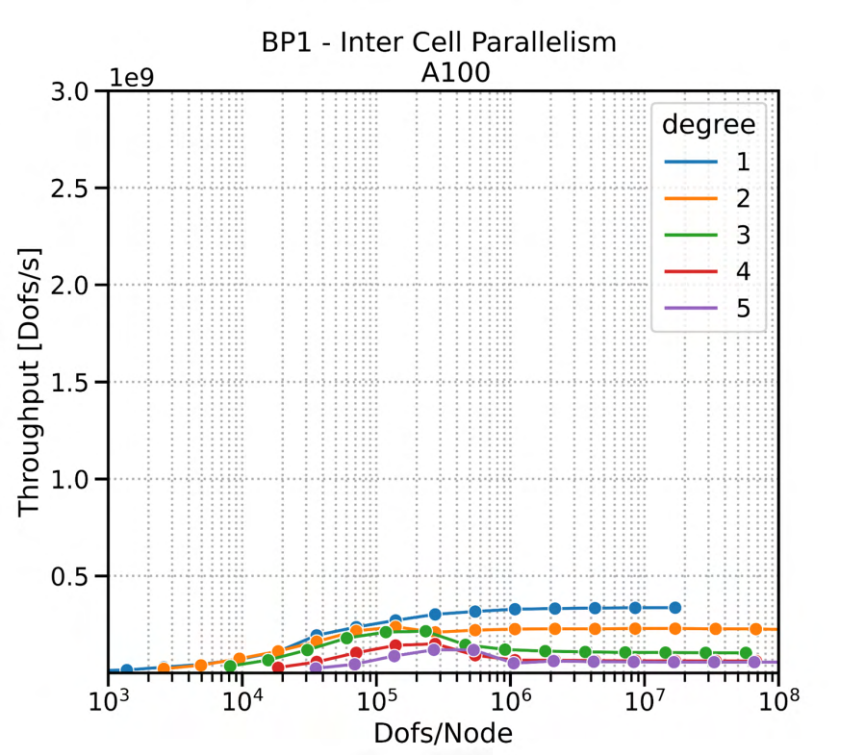
BP3 - Ice Lake node

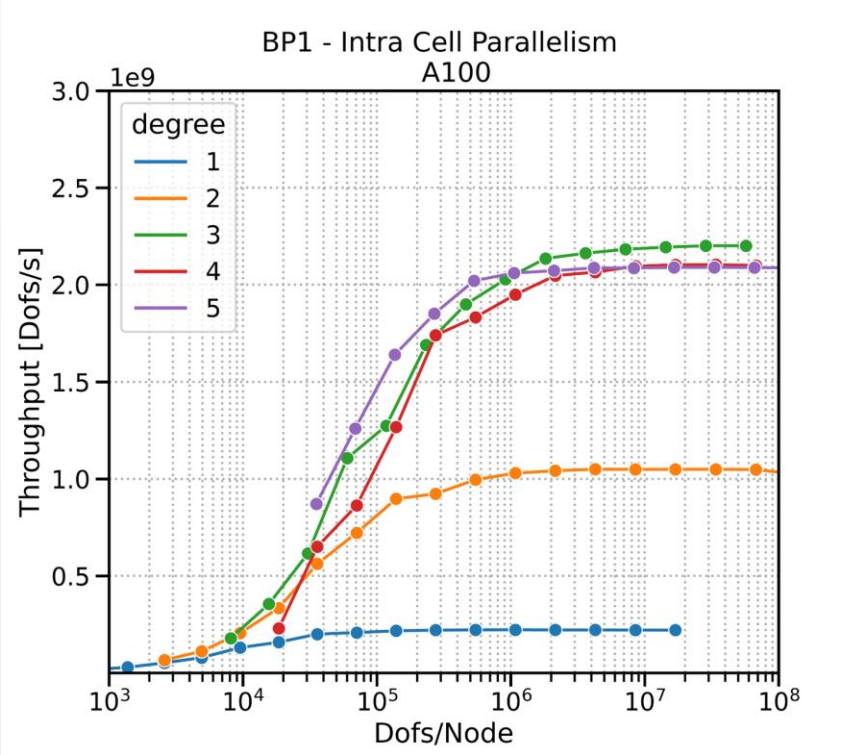




- Local Performance vs Local Problem Size
- Peak* Performance: 1.1 TDOF/s per node

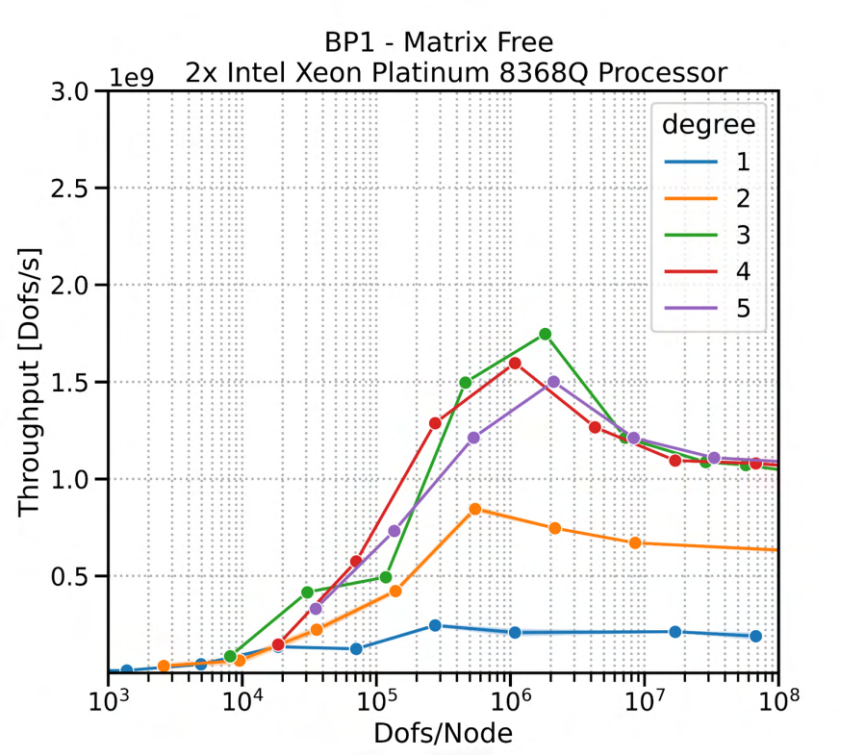
BP1 - Matrix-free - NVIDIA A100-80GB

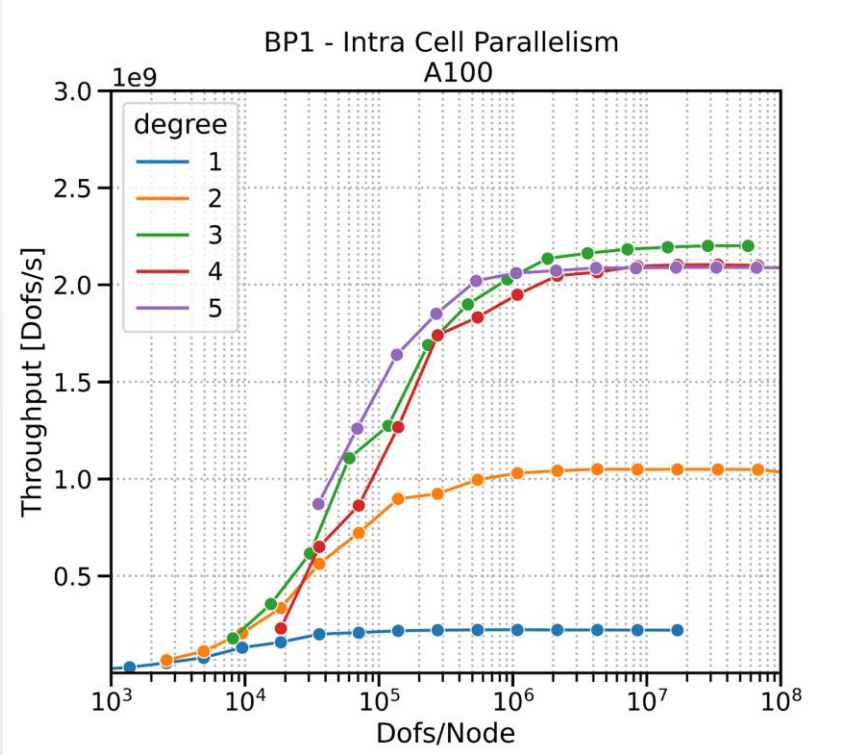




- One cell per thread vs one cell per block
- Peak Performance: 2.2 TDOF/s per node

CPU vs GPU implementations

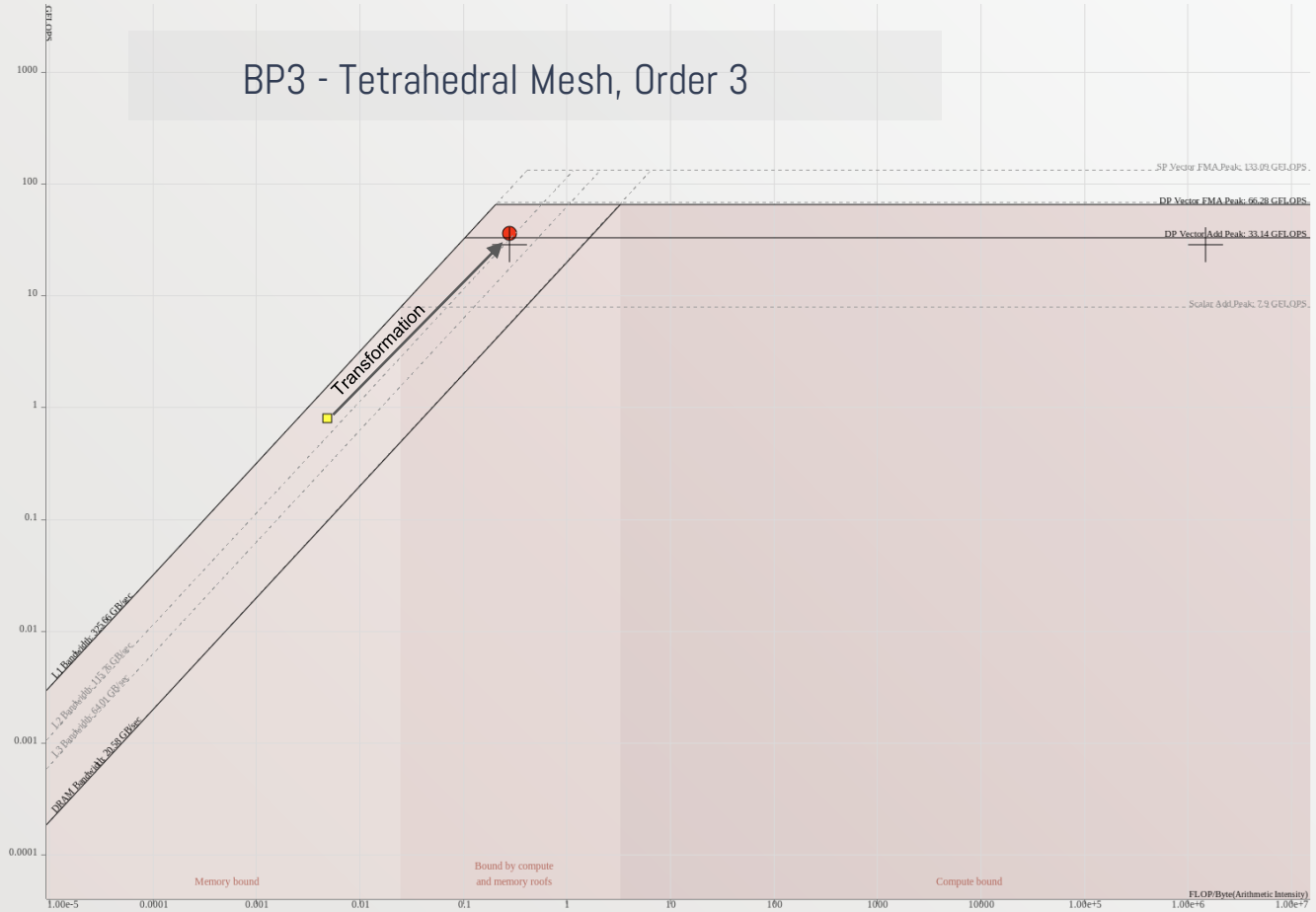




- DP Peak Performance: 4.6 TFLOPS/S
- Best CPU implementation: 1 cell per thread

- DP Peak Performance: 9.7 TFLOPS/S
- Best GPU implementation: 1 cell per thread block

GPU and CPU rooflines - matvec kernel



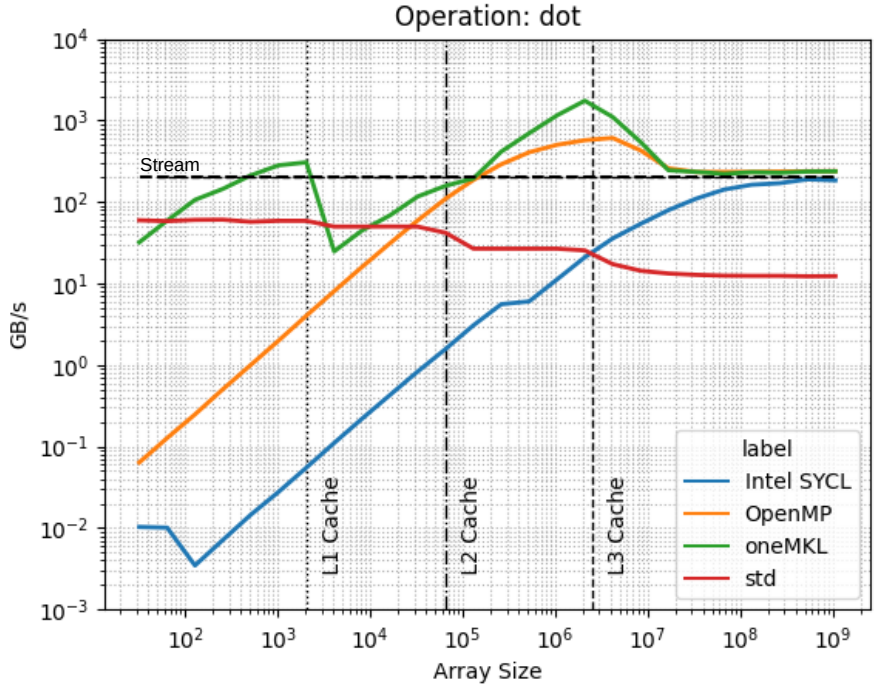
~50% of peak performance with AVX512

Strategies for reducing global memory traffic

Improved memory access pattern

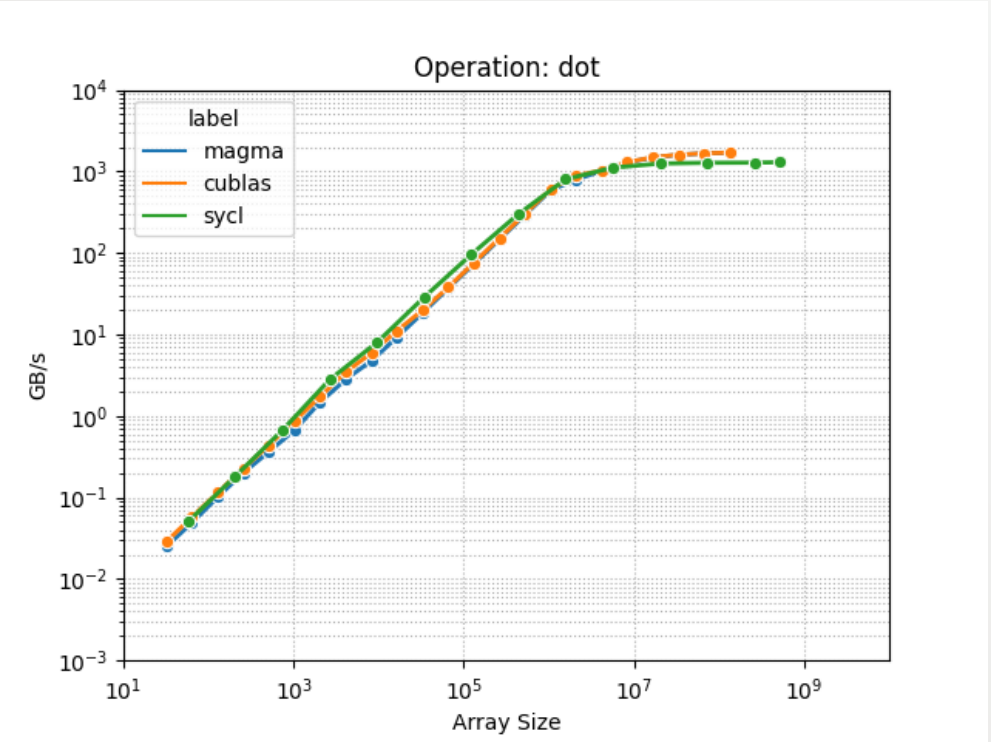
Improved generated code for auto-vectorization

dot: $\alpha = x \cdot y$ (Cascade Lake, 56 cores per node)




```
template <typename T, std::size_t BLOCK_SIZE=32>
T dot(sycl::queue q, std::size_t n, const T* x, T* y) {

 auto padded_length = BLOCK_SIZE * divceil(n, BLOCK_SIZE);
 sycl::range local{BLOCK_SIZE};
 sycl::range global{padded_length};

 auto sum = sycl::malloc_shared<T>(1, q);
 sum[0] = 0.0;

 auto event = q.parallel_for(sycl::nd_range<1>{global, local},
 sycl::reduction(sum,std::plus<T>()),
 [=](sycl::nd_item<1> it, auto& sum) {
   std::size_t i = it.get_global_id(0);
   if (i < n)
     sum += x[i] * y[i];
 });
 event.wait();

 T result = sum[0];

 return result;
}
```

dot: $\alpha = x \cdot y$ (A100 - 80GB)



SYCL code:

```
template <typename T, std::size_t BLOCK_SIZE=32>
T dot(sycl::queue q, std::size_t n, const T* x, T* y) {

 auto padded_length = BLOCK_SIZE * divceil(n, BLOCK_SIZE);
 sycl::range local{BLOCK_SIZE};
 sycl::range global{padded_length};

 auto sum = sycl::malloc_shared<T>(1, q);
 sum[0] = 0.0;

 auto event = q.parallel_for(sycl::nd_range<1>{global, local},
 sycl::reduction(sum,std::plus<T>()),
 [=](sycl::nd_item<1> it, auto& sum) {
   std::size_t i = it.get_global_id(0);
   if (i < n)
     sum += x[i] * y[i];
 });
 event.wait();

 T result = sum[0];

 return result;
}
```

Final observations

The range of problem sizes was chosen to span from the performance-saturated limit (>1M cells per core) to beyond the strong-scale limit (1 cell per core).

- SYCL implementation performed well in the performance-saturated limit (CPU and GPU).
- Small Problems: difficult to hide latency and thread overhead.

Where to go from here?

- In-depth profiling of the finite element GPU kernels.
- Testing on Intel GPUs underway (looks promising).
- Optimized MPI-X communication (hundreds of gpus).
- Code generation for GPUs (generalize to different differential operators).

Paul Calleja (University of Cambridge)

The Modern Cloud Native Heterogeneous Super Computer – A Converged Platform for Simulation, AI & Data Analytics

Unfortunately, due to illness Paul Calleja was unable to present at CIUK 2021.

Abstract: During the presentation we will position research computing infrastructure as a vital national capability driving research innovation and enabling the knowledge economy and industrial competitiveness whilst also providing vital resources in terms “Urgent Computing” for national emergency response activities. We will look at how modern research computing infrastructure is evolving from the traditional monolithic proprietary, highly specialized supercomputers of the past to cloud native heterogenous high performance infrastructure supporting converged AI, simulation and data analytics workloads. We will illustrate this trend by looking at the Cambridge Research Computing Service, its infrastructure, operational model and a selection of use cases and current development / outreach activities that may be of interest.

Bio: Dr Calleja is Director of Research Computing at the University of Cambridge where he oversees one of the UK’s leading large-scale National HPC centres supporting a diverse community of UK frontier science, engineering, and medical research programmes.

Dr Calleja has a strong academic / industrial HPC co-design background focusing on commodity open standards-based solutions. Recently he has pioneered the convergence of OpenStack and Research Computing use-cases, working with industry partners to develop the “Scientific OpenStack”, a software defined supercomputing middleware solution making large scale cloud native supercomputing a reality.

Dr Calleja also heads up the Cambridge Open Exascale Lab a prominent UK academic / industrial collaboration aimed at the development and democratisation of exascale computing solutions.

Alastair Basden (Durham University)

The Durham Intelligent NIC Environment (DINE)

Abstract: We present the DINE cluster which uses NVIDIA Data Processing Unit (DPU) technology in the form of BlueField cards to enable advanced HPC algorithms.

Now entering its second generation, DINE is a 24 node test cluster equipped with AMD Rome processors, 512GB RAM per node and NVIDIA BlueField DPUs. BlueField-2 DPUs contain 8-core ARM processors at >2GHz and offer a 200GBit/s HDR network interconnect which is shared with the host. The DPUs have access to host memory via RDMA.

Here we present the hardware and software configuration of DINE, and discuss how this heterogeneous compute cluster provides insight into future Exascale systems. We provide researcher use cases and describe how codes can be adopted to benefit from the flexibility of the underlying architecture.

Bio: Alastair is the technical manager for the DiRAC Memory Intensive HPC service hosted at Durham.

BlueField in HPC

Experience with the DINE cluster

Alastair Basden and many others
Durham University / DiRAC

DiRAC
High Performance
Computing Facility

Durham
University
Institute for Computational
Cosmology

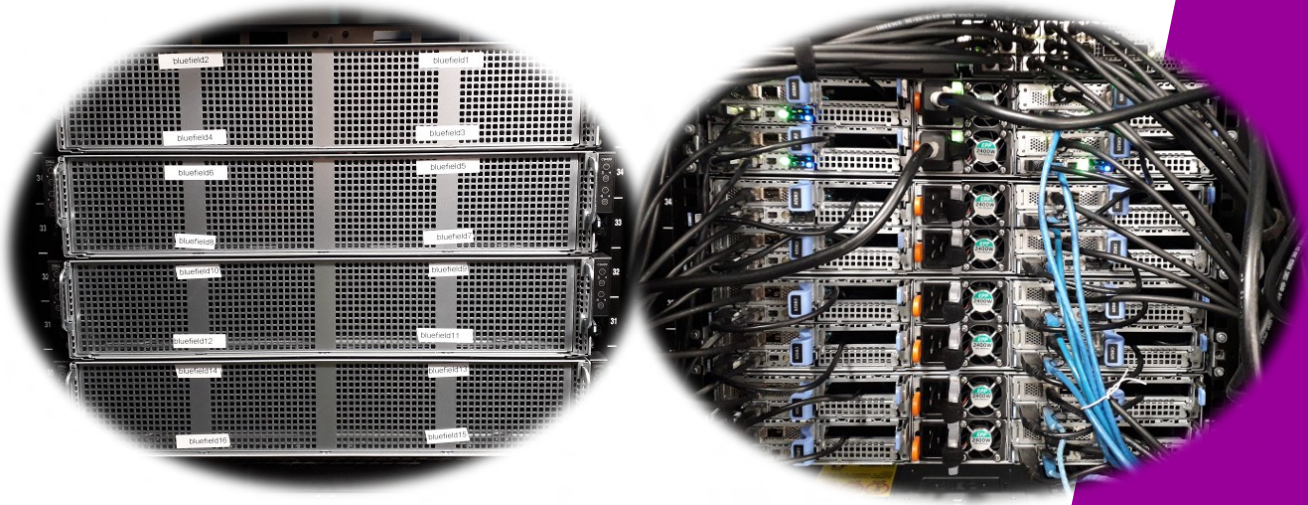
What is BlueField?

- The NVIDIA (Mellanox) Data Processing Unit
  - DPU (previously, known as an Intelligent NIC)
- Offloads data processing from a CPU
  - A programmable NIC
  - ARM cores (Linux)
  - Connect-X NIC chip
  - Hardware accelerators
  - Optional GPU chip
- Primary use cases: Not HPC
- This talk: Using BlueField in HPC)
  - Heterogeneous compute

Credit: nvidia.com

The DINE cluster

- A 16/24 node test cluster
  - Part of the COSMA DiRAC HPC facility, with Durham University investment
  - Dell C6525 half-U (double density servers)
  - 2x 16-core AMD Rome processors, 3GHz, 512GB RAM
- 16/24 BlueField cards
  - 16x BF-1, 25G Ethernet
  - 24x BF-2, 200G IB
- Available to the UK community via ExCALIBUR



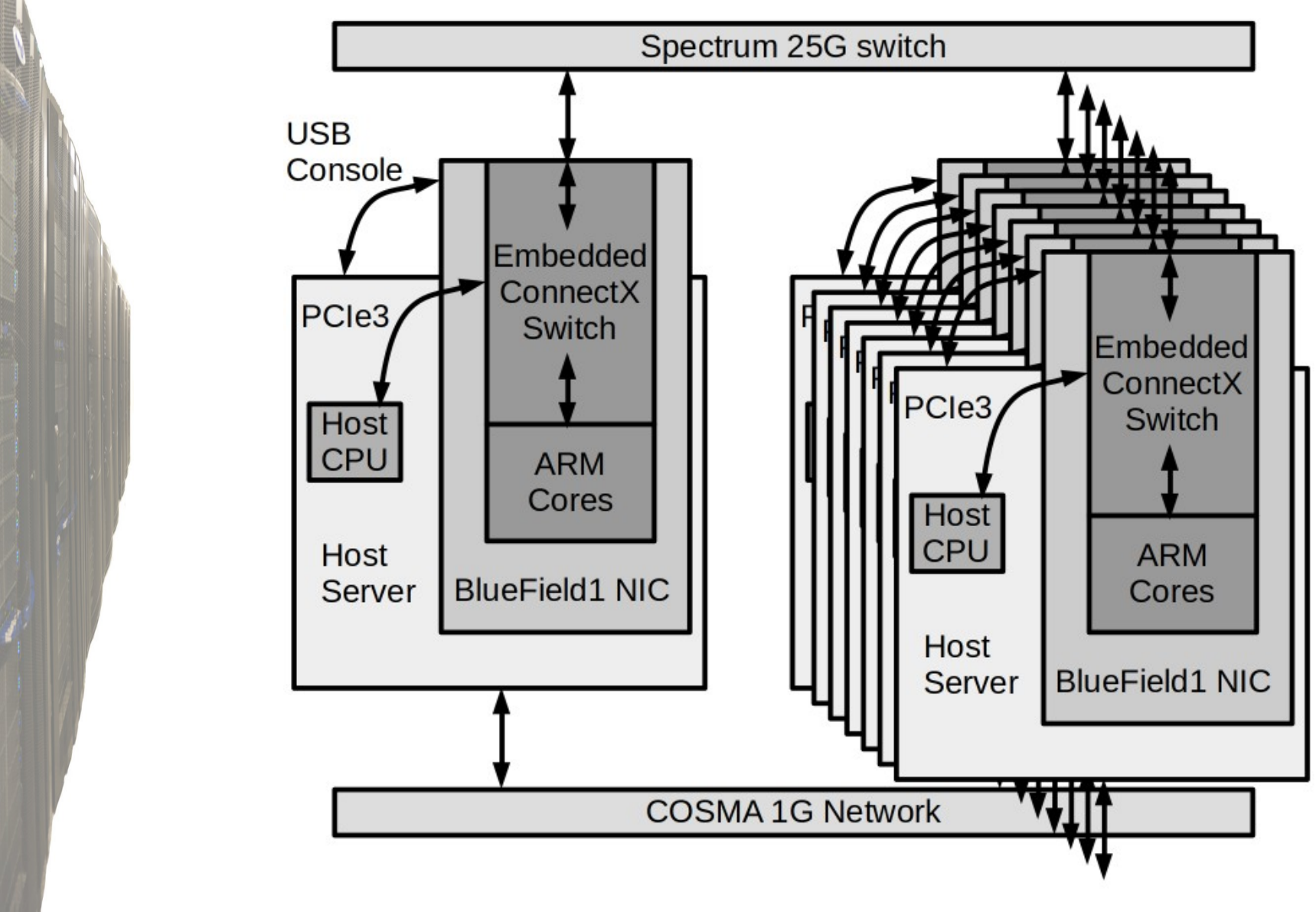
COSMA / DiRAC

- DiRAC: STFC-funded HPC Tier-1 service
- COSMA: The DiRAC Memory Intensive service
  - 4 generations of system in operation
  - COSMA8 has just come online as part of DiRAC-3
    - AMD Rome, Dell 1/2U servers, 1TB RAM/128 cores
  - ~70k cores, 13PB storage, 26PB tape

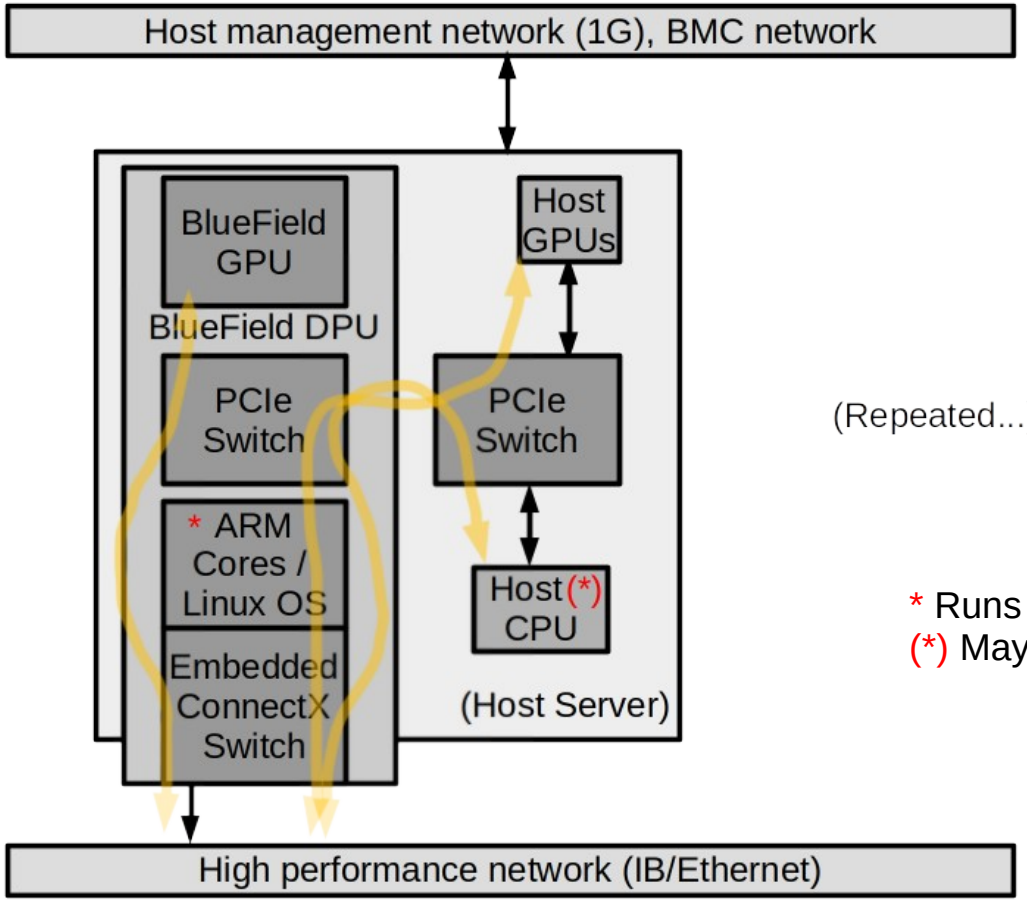
BlueField network

- Host-separated mode:
  - ARM cores run separate Linux OS
  - Have their own MAC address
    - Separate from that of the host
  - But physically share the same network interface
  - Connect-X chip routes between them
  - Look like separate devices on the network

Credit: nvidia.com


Spectrum 25G switch
USB
Console
Embedded
ConnectX
Switch
PCIe3
Host
CPU
ARM
Cores
Host
Server
BlueField1 NIC
COSMA 1G Network

Re-centring around the DPU



(Repeated...)

* Runs MPI tasks
(*) May run MPI tasks

Flexible HPC

- MPI tasks run on the BlueField ARM cores
- GPUs (on BF, or within node) can be used
- Server CPUs can be used (separate MPI tasks)
  - But this is optional: possibly to have no CPU intervention
  - Server CPUs become an accelerator
- Host RAM can be accessed from GPU or ARM
- GPUDirect can also be used
- For traditional CPU intensive tasks:
  - MPI tasks run only on the server CPU (and optionally GPUs)
- Low latency network-bound tasks:
  - MPI tasks on the DPU cores

Heterogeneous compute

- Combining ARM, X86, GPU
- Maximising the ability of BlueField for data orchestration

DPU HPC use-cases

- Data migration
- Task migration and management
- MPI progression

DPU data migration

- DPU can access host memory by RDMA
  - No host processor intervention required
  - This can be used to seamlessly move or repackage data
    - RDMA to remote host

Task migration and management

- Load balancing becomes problematic for Exascale systems
  - Some nodes become congested
  - Others are starved of tasks
- DPU can be used to migrate tasks
  - And return the results
  - Determine where new tasks should run, based on host load
- Ideal for task-based parallelism codes
  - See work on ExaHype

MPI progression

- MPI operations are not always scheduled as expected
  - The MPI progression issue
  - Compute cores can be blocked waiting for their MPI task to complete or be scheduled
    - CPU time wasted
  - Offload MPI operations to the DPU:
    - Data passed to DPU, host CPU can then continue
    - DPU performs the MPI task
    - DPU can ensure MPI progression using dedicated threads

Lessons learned

- Not an out-of-the-box solution
  - Sys-admin/resOps effort is required
- Requires effort from users
  - Have to be willing to compile ARM and x86 binaries
  - Have to give correct command line args when launching an MPI task spanning host and BF
  - Have to understand what they’re doing
  - Not all MPI libraries supported

Now for some detailed information

- Hardware installation
- Network interfaces
- Software configuration
  - Including for the BF devices
- Additional information
- Next steps

Hardware installation

- Insert the BF cards
  - 16 screws per card!
    - Removal of full height face plate
    - Removal of 1st PCIe riser
    - Removal of 2nd PCIe riser to let USB cable through

Hardware thoughts

- A bit of a faff
  - Why not put the USB port on the face plate! Or within the PCIe switch
  - Why not ship with a half-height face place!
- The USB port (serial interface) was essential
- When we later wanted to install a 2nd PCIe card, we modified the face plate to allow the USB cable through it (using some tin snips)

Network interfaces

Hostname	IP address	Host or device	Network interface	Network	Comment
b101 – b116	172.17.178.201 – 216	Host	em1	COSMA 1G	COSMA network
Unnamed	192.168.100.1 – 31 (odd)	Host	tmfifo_net0	Internal PCIe console	Low bandwidth
bluefield101 - 116	192.168.100.2 – 32 (even)	Device	tmfifo_net0	Internal PCIe console	Low bandwidth
bfh101 – 116	192.168.101.1 – 31 (odd)	Host	p1p2	25G	Dedicated Bluefield network
bfd101 – 116	192.168.101.2 – 32 (even)	Device	enp3s0f1	25G	Dedicated Bluefield network
bflocal (only from a host)	192.168.101.2 – 32 (even)	Device	enp3s0f1	25G	This hostname used to access the device within the current server

Software config

- CentOS7 on servers (standard COSMA7 image)
- BlueField drivers including rshim kernel module installed
- Internal tmfifo_net0 interface brought up
- NAT set up on hosts, iptables installed - providing file system access
- sshd config modified
- Network settings for p1p2 interface (The BlueField 25G Ethernet fabric) created
- BlueField device booted with CentOS7 image (.bfb from NVIDIA-Mellanox)
- MOFED installed, OpenIBD service restarted
- bflocal added to /etc/hosts, IP pointing to local card on the 25G network
- 25G IP addresses added to DNS (host and device)

Software config of BF devices

- Initial access:
  - via tmfifo_net0 interface (low bandwidth internal NIC)
  - screen /dev/ttyUSB0 115200, root/centos
  - Set gateway to host IP address to provide a route to rest of cluster.
- Set hostname (hostnamectl)
- Add search criteria to /etc/resolv.conf
- Add Ethernet COSMA file systems to /etc/fstab
- Install kerberos and IPA packages, and set up config
- Reboot host and device
- Enrole device in IPA, create sshd config file, restart sshd
- Switch device into host-separated mode
  - Reboot host server. At this point, the server now sees a boot problem, and requires F1. Probably some non-compliance.
- Change default gateway to the 25G network
- For RDMA, mediated device required, enable some mlxconfig flags
- Install MOFED

Other things

- Haven’t yet built our own bfb file (Arm Linux image), though we have instructions on how to do that.
- Integration of the BlueField device with cluster manager was harder because they weren’t on the same network
  - No direct access for cluster manager
  - Hence creation of a bflocal address on the hosts
    - Allowing effective pdsh access to all BlueField devices
- Some experiments with OpenVSwitch

Next steps

- Upgrade to BlueField-2
  - HDR200
  - 2-node test system almost ready
    - OOB Network interface (direct access to SLURM etc)
    - No F1 prompt on reboot!
- Closer integration with system package manager
  - Updating packages etc
- Closer integration with SLURM
  - A BlueField-only queue?
  - Mixed BlueField-Host queue?
  - Will require direct contact between BF and SLURM
    - Should be easier with BF2

Heterogeneous compute: Other stuff

- GPU acceleration of adaptive atmospheric correction system
  - First “on-sky” demonstration of GPU for telescope control worldwide
- FPGA acceleration of HPC code
  - Xilinx FPGAs tightly coupled to AMD X86 processors
    - Avoiding PCI bottleneck
  - Used in offload mode
    - Dataflow pipeline moved onto FPGA
    - Algorithms include simulation of:
      - Atmosphere, telescope optics, detectors, random noise, data processing
      - Custom Mersenne Twister generator, 2D FFTs
    - Hand-coded VHDL
    - 600x speed-up over CPU-only version

Heterogeneous compute: Not new

- GPU acceleration of adaptive atmospheric correction sys
  - First “on-sky” demonstration of GPU for telescope control wo
- FPGA acceleration of HPC code
  - Xilinx FPGA tightly coupled to processor
    - Avoiding PCI bottleneck
  - Used in offload mode
    - Dataflow pipeline moved onto FPGA
    - Algorithms include simulation of:
      - Atmospheric, telescope optics, detectors, random noise, data processing
      - Custom Mersenne Twister generator, 2D FFTs
    - Hand-coded VHDL
    - 600x speed-up over CPU-only version

2010

2005:
Cray XD1

Conclusions

- Heterogeneous compute: tread carefully
  - Make codes generic, use portable standards
    - Task-based parallelism
  - Share any work, contribute to standards, libraries etc
- DINE cluster provides a research platform for DPUs
  - Hybrid X86/Arm MPI tasks
  - Data orchestration
  - Task migration
  - MPI progression
  - Ask if you wish to use it… (via ExCALIBUR)

Andrew Edmondson (Research Software Group Leader, Advanced Research Computing, University of Birmingham)

Installing and curating software for heterogeneous compute environments

Abstract: The Advanced Research Computing team at the University of Birmingham operates evolving and constantly-changing heterogenous compute environments. This includes the Baskerville Tier 2 HPC, our BlueBEAR Tier 3 HPC, and own-cloud virtual machines. Across these compute systems we currently have five different CPU architectures: Intel (Haswell, Broadwell, Cascade Lake and Ice Lake) and IBM POWER9, running RedHat, CentOS Stream and Ubuntu operating systems. We also operate multiple different NVIDIA GPU models, from K40 to A100.

We manage a large range of over 1,000 applications across all these systems, using EasyBuild and a suite of our own automation and management tools to maintain quality, consistency and minimise the staff time required. These applications can be seen on the automatically-updated websites https://bear-apps.bham.ac.uk/ and https://apps.baskerville.ac.uk.

In this talk I will tell the story of how our processes, systems and tools have evolved over time from a single, homogenous HPC cluster to our current, complex heterogenous environment. I will explain some of the difficulties and challenges we encountered, and how we have overcome them – including the times we have chosen to spend time in order to save time.

Bio: Known as "Ed", I started my career as a software engineer and team leader at QinetiQ, after completing an MMath at the University of Oxford. I left QinetiQ to complete a BA in Theology at Birmingham Christian College after which I worked part-time as a senior developer at ApplianSys working on embedded Linux and Python firmware for network appliances.

I completed a part-time PhD in New Testament Textual Criticism in the Institute for Textual Scholarship and Electronic Editing (ITSEE) at the University of Birmingham supervised by Professor David Parker. The title of my PhD thesis is "An analysis of the coherence-based genealogical method using phylogenetics" and is available online here. I am currently an Honorary Fellow of ITSEE.

In 2016 I joined Advanced Research Computing at the University of Birmingham and founded the Research Software Group (RSG). The RSG supports researchers using ARC's various compute resources, and offers advice, coaching, coding, mentoring and training to researchers and RSEs across campus.

I am an active member of the Society of Research Software Engineering, and was the Programme Chair of the 2019 UK RSE Conference.

UNIVERSITY OF BIRMINGHAM | BEAR
BIRMINGHAM ENVIRONMENT FOR ACADEMIC RESEARCH

Installing and curating software for heterogeneous compute environments

Dr Andrew Edmondson

Advanced Research Computing, University of Birmingham

BEAR services


BlueBEAR HPC


BEAR Research Data Store


BEAR Cloud


CaStLeS (powered by BEAR)


BEAR Gitlab


BEAR Data Transfer

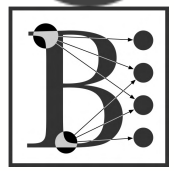

BEAR Archive


BEAR Research Data Network

BEAR Software


Exclusive BEAR


BEAR Database Service


Baskerville Tier 2 HPC

https://intranet.birmingham.ac.uk/bear/services


UNIVERSITY OF BIRMINGHAM | BEAR BIRMINGHAM ENVIRONMENT FOR ACADEMIC RESEARCH

CPU and OS

2012
- BlueBEAR 2 – Sandy Bridge – SL6
- Single, homogeneous cluster

2016
- Added BlueBEAR Haswell – EL7 – heterogeneous cluster
- Launched BEAR Cloud – Haswell – EL7 / U1604

2017
- Broadwell – EL7
- NVIDIA P100s

CPU and OS

2018
- POWER9 – EL7
- NVIDIA V100s

2019
- Cascade Lake – EL7

2020
- Upgrade EL7 => EL8 and U1604 => U2004
- Decommissioned Sandy Bridge nodes

CPU and OS

2021
- BlueBEAR Ice Lake – EL8 – NVIDIA A30 and A100
- Baskerville Tier 2

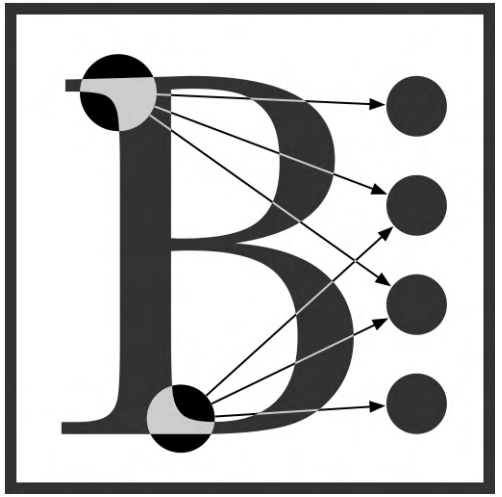
2022
- ?

2023
- ?

Baskerville

- 46 Ice Lake nodes with 512G RAM and quad NVIDIA A100 GPUs
- Total of 184 A100 GPUs
- Novel accelerator technologies:
  - NextSilicon ExCALIBUR project on Baskerville - https://excalibur.ac.uk/projects/novel-hardware-software-architecture-testbed/
  - More planned to come

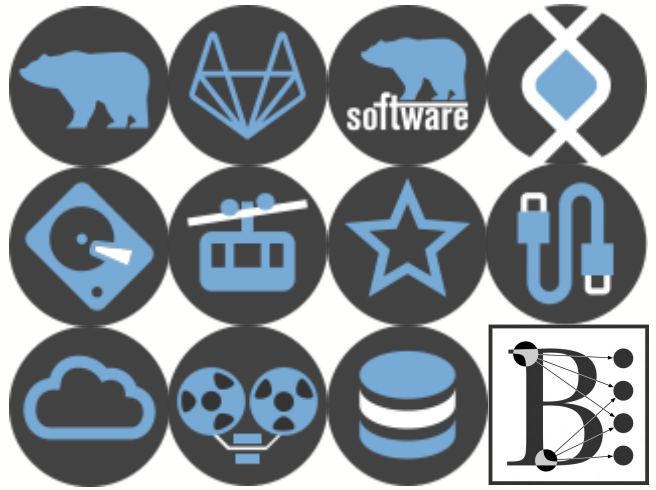
- https://www.baskerville.ac.uk/
- Come and see us in the Research Zone!



UNIVERSITY OF BIRMINGHAM | BEAR BIRMINGHAM ENVIRONMENT FOR ACADEMIC RESEARCH

Simon Thompson

Our infrastructure story from a homogeneous Sandy Bridge cluster to a multi-architecture, multi-OS suite of services: “The Thompson Era”



Who will be next?

https://www.jobs.ac.uk/job/CKY408/research-computing-infrastructure-architect

Closing date 15th December

UNIVERSITY OF BIRMINGHAM | BEAR BIRMINGHAM ENVIRONMENT FOR ACADEMIC RESEARCH

Users of BEAR's compute services

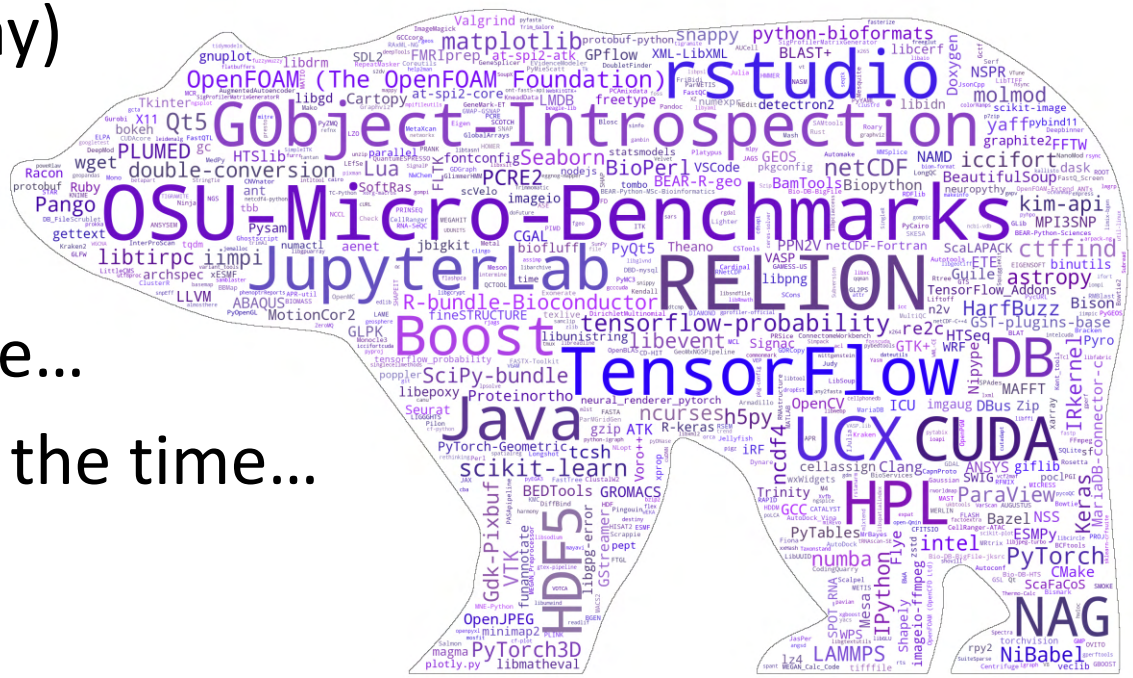
- Wide variety of researchers
  - Traditional HPC users and non-traditional
  - Staff, PhD, PGT, UG

- Slurm-based BlueBEAR HPC (SSH)
- OpenStack BEAR Cloud VMs (SSH/x2go)
- Open OnDemand BEAR Portal
  - Jupyter/Rstudio/MATLAB/…
  - BlueBEAR GUI

UNIVERSITY OF BIRMINGHAM | BEAR BIRMINGHAM ENVIRONMENT FOR ACADEMIC RESEARCH

The challenge we face

- Wide variety of people using BEAR's compute services
- 1,181 unique applications/libraries
- Six different CPU/OS combinations (today)
- 3,122 distinct modules
- Upgrade OS => rebuild software…
- Add new CPU or GPU => rebuild software…
- And new software is being requested all the time…



UNIVERSITY OF BIRMINGHAM | BEAR BIRMINGHAM ENVIRONMENT FOR ACADEMIC RESEARCH

Installing and curating software

- The rest of this talk:
  - EasyBuild: Installing software efficiently
  - Documenting what's installed
  - ReFrame: Testing and finding problems
  - Monitoring and visualising usage
  - Deprecating and removing old software
  - Problems of a changing environment

EasyBuild

"EasyBuild is a software build and installation framework that allows you to manage (scientific) software on High Performance Computing (HPC) systems in an efficient way."

https://easybuilders.github.io/easybuild/

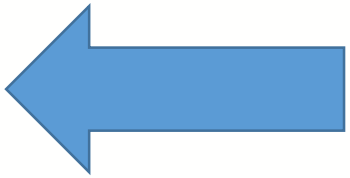
UNIVERSITY OF BIRMINGHAM | BEAR BIRMINGHAM ENVIRONMENT FOR ACADEMIC RESEARCH

EasyBuild

- We started in 2016 – when Haswell arrived
- We contributed lots upstream
  - POWER9 support
  - New applications/versions
  - Testing
  - Dr Simon Branford is now an official EasyBuild Maintainer
- We've benefited hugely from other people's work
- See https://github.com/bear-rsg/easybuild-easyconfigs

EasyBuild

- REALLY helpful for installing the same thing for multiple arch/OS
- E.g. our 2019a environment has:
  i. EL7-cascadelake
  ii. EL7-haswell
  iii. EL7-power9
  iv. EL7-sandybridge
  v. Ubuntu16.04-haswell
  vi. EL8-cas
  vii. EL8-has
  viii. EL8-p9
  ix. U20-has

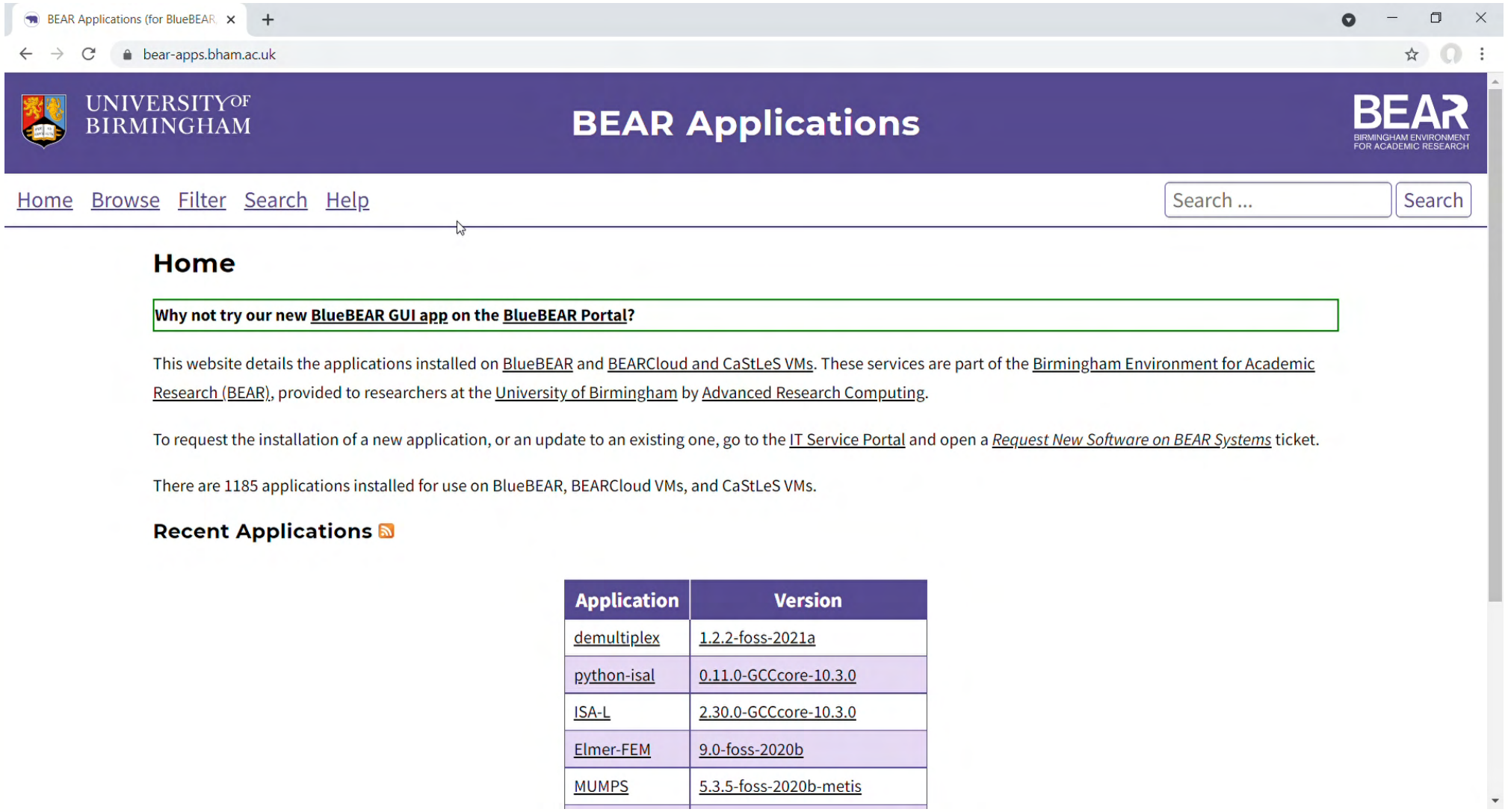


Upgraded OS
Changed naming convention
Sandy Bridge decommissioned

Documentation

- We used to maintain our user-facing docs manually…
  - Out of date
  - Time consuming

- Now we use https://bear-apps.bham.ac.uk
  - Django-based website
  - API for generating docs from modules
  - “Spend time to save time…”


BEAR Applications (for BlueBEAR
bear-apps.bham.ac.uk
UNIVERSITY OF BIRMINGHAM
BEAR Applications
BEAR
BIRMINGHAM ENVIRONMENT FOR ACADEMIC RESEARCH
Home Browse Filter Search Help
Search ... Search
Home
Why not try our new BlueBEAR GUI app on the BlueBEAR Portal?
This website details the applications installed on BlueBEAR and BEARCloud and CaStLeS VMs. These services are part of the Birmingham Environment for Academic Research (BEAR), provided to researchers at the University of Birmingham by Advanced Research Computing.
To request the installation of a new application, or an update to an existing one, go to the IT Service Portal and open a Request New Software on BEAR Systems ticket.
There are 1185 applications installed for use on BlueBEAR, BEARCloud VMs, and CaStLeS VMs.
Recent Applications
Application
Version
demultiplex
1.2.2-foss-2021a
python-isal
0.11.0-GCCcore-10.3.0
ISA-L
2.30.0-GCCcore-10.3.0
Elmer-FEM
9.0-foss-2020b
MUMPS
5.3.5-foss-2020b-metis

ReFrame – HPC testing

"ReFrame is a powerful framework for writing system regression tests and benchmarks, specifically targeted to HPC systems…"

https://reframe-hpc.readthedocs.io

- Nightly tests – can detect all kinds of problems with software, systems, filesystems, package updates, ...
- Single-node stress tests
- Methods to generate a single test that runs on all subsets in a group of nodes – e.g. to detect MPI/networking problems.

Monitoring usage

- Imagine you're upgrading the OS, and you know lots of software won't work any more (due to major glibc changes, for example)…
  - Do you have to rebuild and fix everything?

- Imagine you've been running the same cluster for 10 years
  - Do you have to keep supporting all the old software?

- What are people using? Do we want to ask them to change? Can we deprecate old software? Can we delete old software?

Monitoring usage

- We wrote a Django-based website (for admin use)
  - Log all module loads to disk (using module headers)
  - Ingest all module loads into the database via cron
  - “Spend time to save time…”

Modules DB

bearadmin.bham.ac.uk/modules/index

UNIVERSITY OF BIRMINGHAM

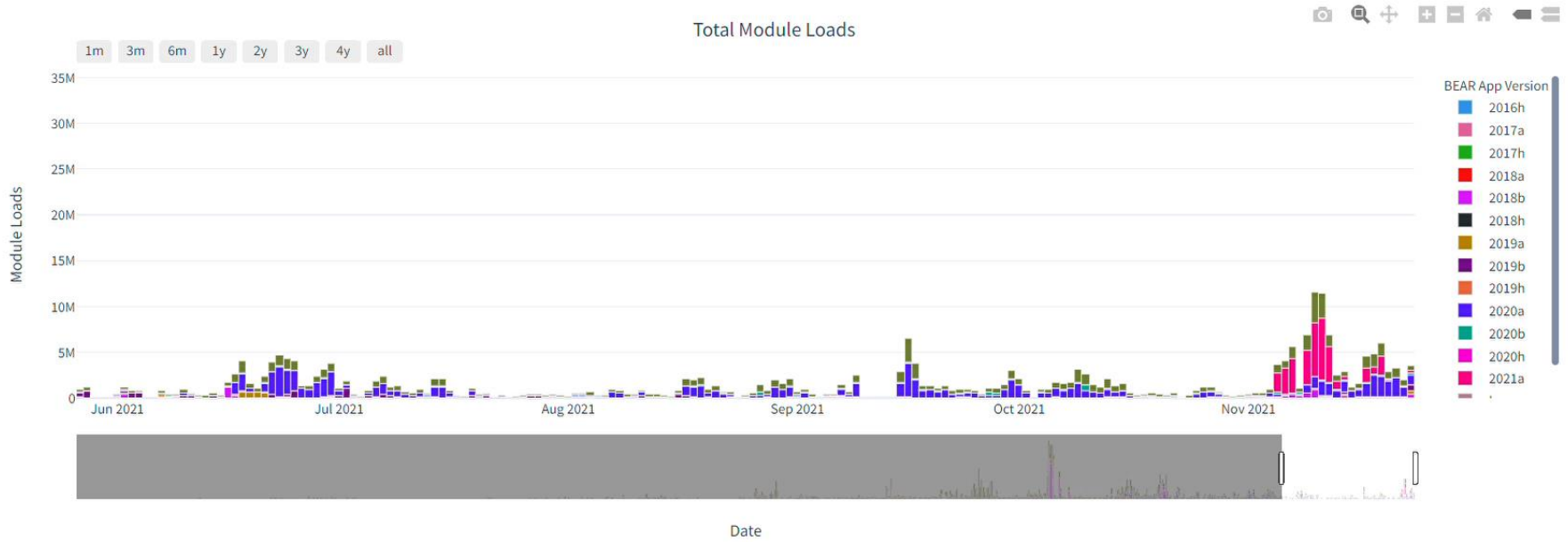
BEAR Admin

BEAR
BIRMINGHAM ENVIRONMENT FOR ACADEMIC RESEARCH

Modules | Deprecated | Users | Top Modules

Search ... | Search | Logout

BEAR Apps Versions



Top

Problems of a changing environment

Example: EL7 to EL8 and Ubuntu 16.04 to Ubuntu 20.04 OS upgrades

- Mid 2020 – “New OS coming”
  - EL7 final scheduled release was 7.9
  - EL8 needed to support latest CUDA on POWER9, for TensorFlow 2.3 etc.
  - EL8 planned for Baskerville, and support for future hardware
  - U1604 out of support in 2021
- Dec 2020 – CentOS support model changed!
  - Change of plans! Switched from CentOS Stream 8 to RedHat 8.
  - Upgraded GPU nodes only
- April 2021
  - Upgraded remaining compute nodes to RedHat 8
  - Upgraded all VMs to CentOS Stream 8, or Ubuntu 2004

Problems of a changing environment

Example: EL7 to EL8 and Ubuntu 16.04 to Ubuntu 20.04 OS upgrades

- Virtually everything needed to be re-installed
  - System library changes – e.g. glibc
- Many old versions wouldn't re-install on EL8 – simple patches but time-consuming!
- Deprecated everything installed before 2018
  - Found who was using old software (using our modules database)
  - Talked to them and helped them to move to new versions, or installs of the old versions with newer compilers, libraries etc.
  - New policy and automatic messaging (next slides)

Deprecating and removing old software

Applications Support and Retention Policy

- Applications installed on the BEAR systems are organised by the year of installation.
- All mention of years in this policy refer to installation date.
- Applications from the current year and the previous two years will usually be supported.
- **Supported applications may be removed at short notice if they can no longer run on the BEAR systems.**
- Applications installed more than two years ago will usually be deprecated.
- **All deprecated applications will warn the user that they are deprecated, each time they are loaded.**
- Any deprecated application may be removed without warning.
- After an application has been deprecated it may become unsupported, instead of being removed.

https://bear-apps.bham.ac.uk/help

UNIVERSITY OF BIRMINGHAM | BEAR BIRMINGHAM ENVIRONMENT FOR ACADEMIC RESEARCH

Deprecating and removing old software

```
$ module load ScaLAPACK/2.0.2-gompi-2018a-OpenBLAS-0.2.20
------------------------------ WARNING --------------------------------
This module is deprecated:
 * ScaLAPACK/2.0.2-gompi-2018a-OpenBLAS-0.2.20

Please see https://intranet.birmingham.ac.uk/bear-apps-versions for
details of the 2020 OS upgrade for BEAR systems.

This module will NOT be available after the OS upgrade.

Please see https://bear-apps.bham.ac.uk for alternative versions.
-----------------------------------------------------------------------
```

UNIVERSITY OF BIRMINGHAM | BEAR BIRMINGHAM ENVIRONMENT FOR ACADEMIC RESEARCH

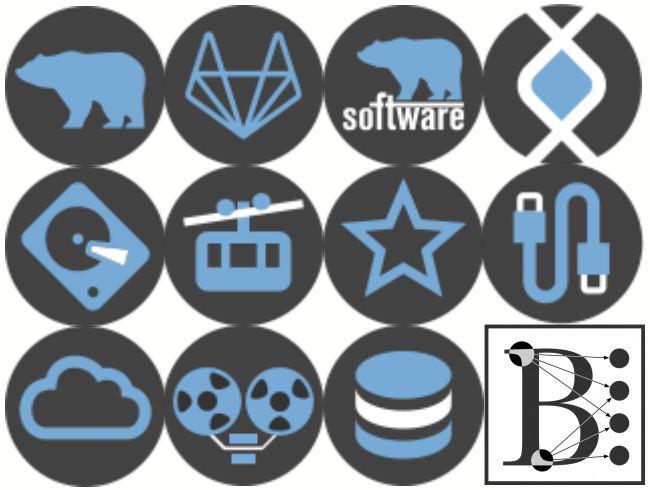
Questions?

UNIVERSITY OF BIRMINGHAM | BEAR BIRMINGHAM ENVIRONMENT FOR ACADEMIC RESEARCH

Come and lead our architecture, infrastructure and systems team!

https://www.jobs.ac.uk/job/CKY408/research-computing-infrastructure-architect

Closing date 15th December



UNIVERSITY OF BIRMINGHAM | BEAR BIRMINGHAM ENVIRONMENT FOR ACADEMIC RESEARCH

Dr Elizabeth Bent (Senior Portfolio Manager, UKRI-EPSRC)

Harnessing Exascale Computing – an ExCALIBUR overview

Abstract: ExCALIBUR is a UK research programme that aims to deliver the next generation of high-performance simulation software for the highest-priority fields in UK research. The programme brings together experts across the research landscape to address the efficiency of simulation codes that will transform capability within UK science and develop expertise for the next generation of supercomputers. This presentation will provide an overview of the programme to date and provide an insight into the science highlights so far.

Bio: Elizabeth is a Senior Portfolio Manager in the Research Infrastructure Team. She manages the UK Research and Innovation (UKRI) aspects of the ExCALIBUR Strategic Priority Fund programme including the ExCALIBUR Steering Committee. She joined EPSRC in July 2017 as a member of the Energy Team where she worked in a range of research areas and convened the Energy Programme Scientific Advisory Committee.

EXCALIBUR 10

HARNESSING EXASCALE COMPUTING – AN EXCALIBUR OVERVIEW

Elizabeth Bent

CIUK 2021, 9 December

Met Office

UKRI UK Research and Innovation

UK Atomic Energy Authority

Strategic Priority Fund (SPF)

The SPF builds on the vision of a ‘common fund’ set out in Sir Paul Nurse’s review.

The Strategic Priorities Fund (SPF) is being led by UKRI to:

- drive an increase in high quality multi and interdisciplinary research and innovation
- ensure that UKRI’s investment links up effectively with government research priorities and opportunities
- and ensure the system responds to strategic priorities and opportunities.

Two waves of programmes

34 themes of research

Harnessing Exascale Computing

Exascale Computing ALgorithms & Infrastructures for the Benefit of UK Research (ExCALIBUR

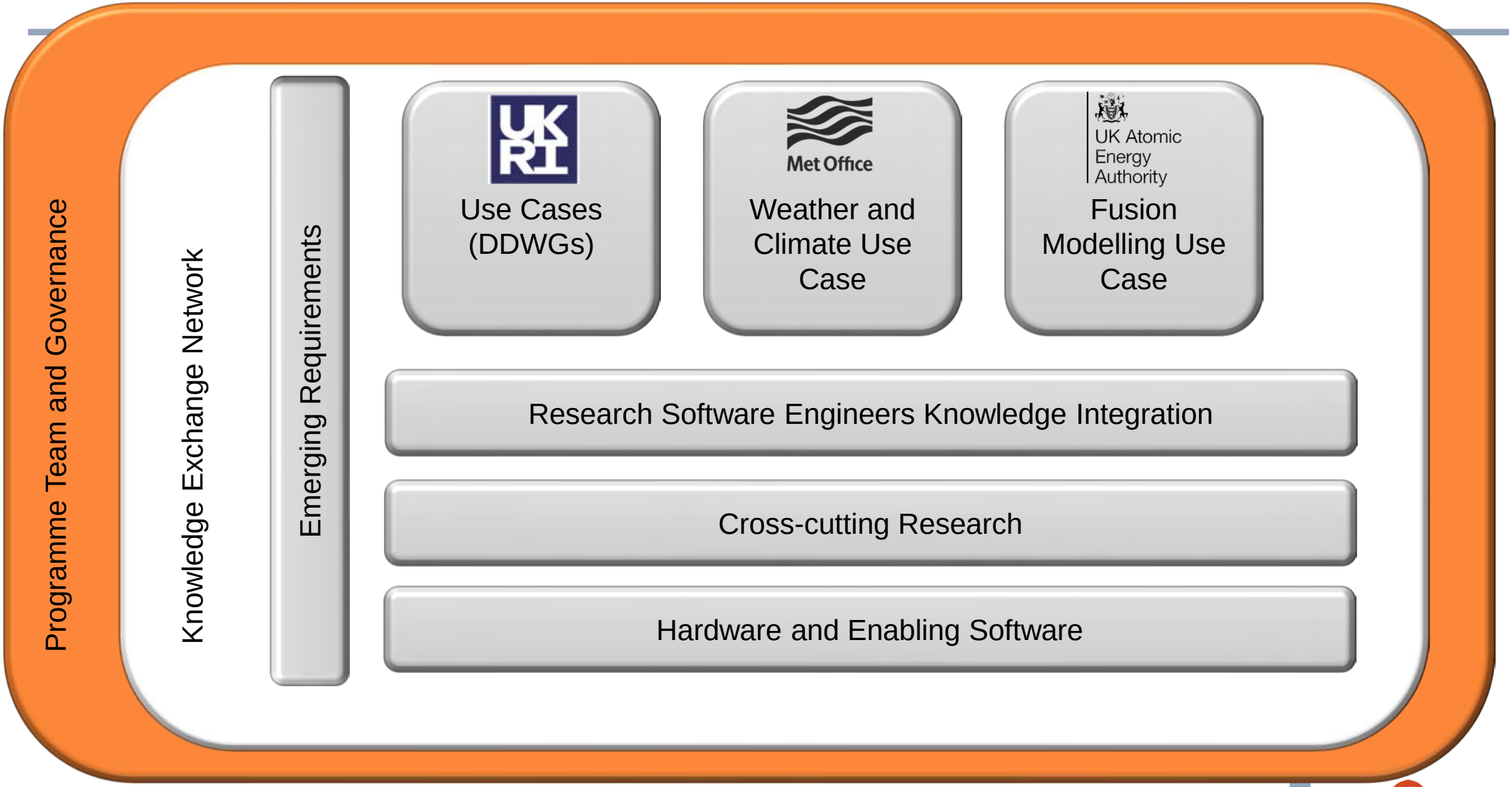
- Lead Delivery partners: Met Office (PSREs) + EPSRC (UKRI)
  - Delivery Partners: UKAEA, STFC, NERC, MRC
- ***Aiming to redesign high priority simulation codes and algorithms to fully harness the power of future supercomputers, keeping UK research and development at the forefront of high-performance simulation science***

Separation of Concerns: Each concern is addressed by distinct parts of the software. Maths of problem separated from computer science of implementation.

Co-design: Holistic design of entire system. Innovative collaborations between mathematicians, domain scientists and computer scientists.

Data Science: Research to design new workflows adapted to managing & analysing vast volumes of data ingested and produced by simulations.

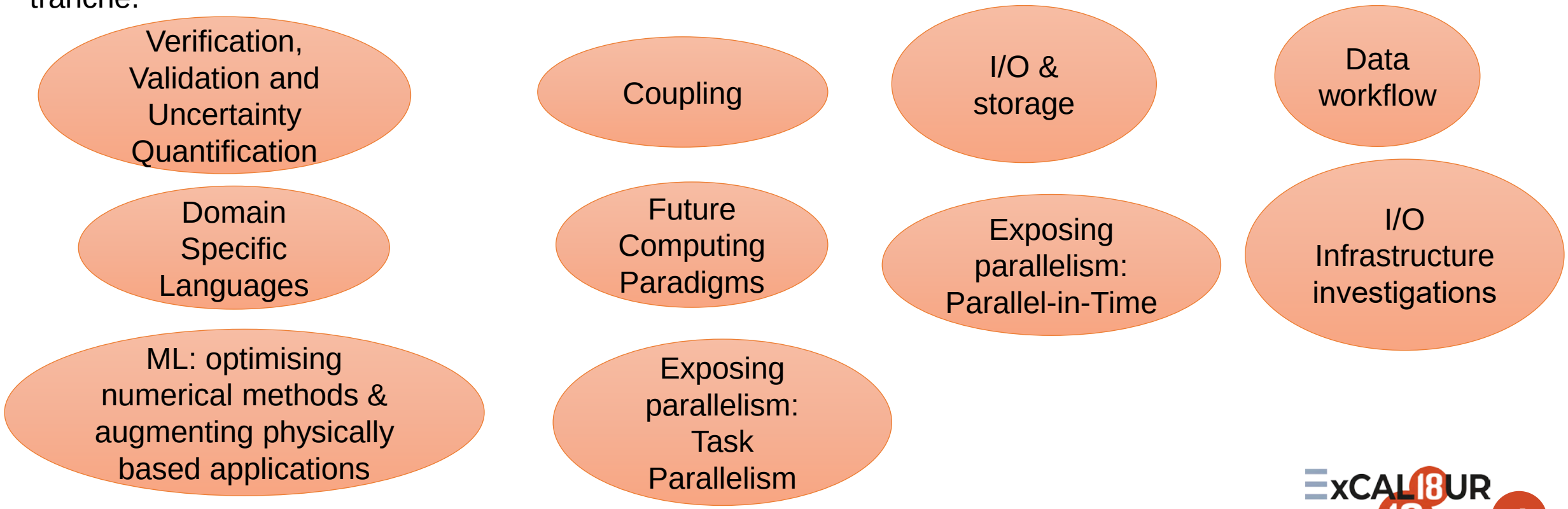
Investment in People: Improved RSE career development driven by professional forward-looking approach to scientific software design of simulation codes.



What is Cross-cutting Research?

- Co-ordinated approach addressing known technology/infrastructure issue if resolved will lead to significant progress across range of exascale software development challenges

- We invited the wider community to help scope the themes of research that should be addressed in this tranche.



Cross-cutting Research Awards

Principal Investigators

Containers
- Dr Stuart Whitehouse, Met Office

Coupling
- Professor Garth Wells, University of Cambridge

Data workflow
- Professor Bryan Lawrence, University of Reading

Domain Specific Languages
- Dr Tobias Grosser, University of Edinburgh

Exposing parallelism: Parallel-in-Time
- Dr Jemma Shipton, University of Exeter and Imperial College London

Exposing parallelism: Task Parallelism
- Professor Tobias Weinzierl, Durham University and STFC Hartree

Future Computing Paradigms
- Dr Vivien Kendon, Durham University

Future Computing Paradigms
- Professor Jason McEwen, University College London

Verification, Validation and Uncertainty Quantification
- Professor Peter Challoner, University of Exeter

Verification, Validation and Uncertainty Quantification
- Professor Peter Coveney, University College London

I/O & storage
- Professor Bryan Lawrence, University of Reading, NCAS and University of Cambridge

I/O Infrastructure investigations
- Dr Stuart Whitehouse, Met Office

Machine learning: optimising numerical methods and augmenting physically based applications
- Dr Amy Krause, University of Edinburgh, EPCC

Workflow Design and Analysis
- Dr Stuart Whitehouse, Met Office

Knowledge Exchange

Identify and lead on opportunities for Knowledge Exchange at project level

Increase the awareness of your project and the ExCALIBUR programme

Collaborate with relevant academic, industrial and international communities

Collaborate as a network with other ExCALIBUR Knowledge Exchange co-ordinators to deliver activities and facilitate knowledge sharing

Knowledge Exchange Co-ordinators

Containers, I/O Infrastructure investigations, Workflow Design and Analysis
- Dr Stuart Whitehouse, Met Office

Coupling
- Dr Chris Richardson, University of Cambridge

Data workflow
- Dr Fanny Adloff, University of Reading

Domain Specific Languages
- Dr Nick Brown, University of Edinburgh

Exposing parallelism: Parallel-in-Time
- Dr Jemma Shipton and Professor Beth Wingate, University of Exeter

Exposing parallelism: Task Parallelism
- Dr Marion Weinzierl, Durham University

Future Computing Paradigms
- Dr John Buckeridge, London South Bank University

Future Computing Paradigms
- Dr Harpreet Dhanoa and Dr Jeremy Yates, University College London

Verification, Validation and Uncertainty Quantification
- Dr Derek Groen, Brunel University

Verification, Validation and Uncertainty Quantification
- Dr James Salter, University of Exeter

I/O & storage
- Dr Fanny Adloff, University of Reading, NCAS

Machine learning: optimising numerical methods and augmenting physically based applications
- Dr Amy Krause, University of Edinburgh, EPCC

UKRI Use Cases

A phased approach

Design and Development Working Groups

- ELEMENT - Exascale Mesh Network
- Materials And Molecular Modelling Exascale Design And Development Working Group
- Gen X: ExCALIBUR working group on Exascale continuum mechanics through code generation
- Exascale Computing for System-Level Engineering: Design, Optimisation and Resilience.
- Massively Parallel Particle Hydrodynamics for Engineering and Astrophysics
- Benchmarking for AI for Science at Exascale (BASE).
- Lattice Field Theory at the Exascale Frontier
- ExaClaw: Clawpack-enabled ExaHyPE for heterogeneous hardware
- ExCALIBUR-HEP (= High Energy Physics)
- Turbulent Flow Simulations at the Exascale: Application to Wind Energy and Green Aviation

A co-ordinated range of activities, which aims to develop simulation code with a focus on an application or applications pre-identified by the relevant communities as benefitting from exascale software development.

Announcement of awarded projects will take place in the new year.

RSE Knowledge Integration

Developing expertise

- Create an evolving training curriculum encompassing both technical and professional development.
- Work towards growth in number of the UK RSE Community across both academia and industry.
- Focus on the development of activities to facilitate both the cross fertilisation of knowledge and the movement of people within and between academia and industry.

Phase 1: Landscape Review

Phase 2: Funding Opportunity

Training and skills embedded in call requirements

Landscape Review

- The skills required by RSEs in HPC.
- The future training needs of RSEs.
- Challenges faced in developing these skills and growing the number of RSEs in the UK with a specific focus on HPC.
- The importance of establishing a career path for RSEs that does not rely on the conventional academic metrics.

Hardware and Enabling Software Group

- H&ES is an initiative set up to trial prototypes and testbeds of a number of potential next generation supercomputing systems and supporting software such as compilers for the use of ExCALIBUR projects and the wider UKRI community.

- Jointly led by Rev Dr Jeremy Yates and Professor Simon McIntosh-Smith. Programme Co-ordinator Martin Hamilton

    - Delivers annual calls to support novel hardware/software testbeds for prototyping and development

    - Lead benchmarking activities
        - Surveying ExCALIBUR projects to understand their exascale benchmarking requirements
        - Working with ExCALIBUR projects to collate and package exascale benchmarks ready for novel architecture

Highlights and Updates

Highlights

- Delivery of numerous workshops, training and scoping activities held by DDWGs since 2019 to develop their use cases
  - Code performance series, Podcast series, Firedrake training, Deep Dive talks
- Series of Cross-cutting workshops and kick off meetings held
  - Task Parallelism: Performance Analysis Methodology workshop
- H&ES have secured general access for ExCALIBUR projects to the first wafer scale accelerator system in Europe (EPCC with Cerebras and HPE)
- First cluster based on NVIDIA Arm Developer Kit in the EMEA region (Leicester with NVIDIA and Arm)

Upcoming opportunities

- Emerging Requirements – to prepare communities and understand their software requirements
- UKRI Use Cases Phase 2 – ensure a balance across the portfolio of UKRI Use Cases supported

Contact us

Visit our website:
www.excalibur.ac.uk

Sign up to Knowledge Exchange Annoucements mailing list:
https://jiscmail.ac.uk/excalibur-ke-annouce

Programme Team:
spfprogrammeoffice@metoffice.gov.uk

elizabeth.bent@epsrc.ukri.org

Rob Akers (UKAEA)

ExCALIBUR – exploiting the exascale to bottle a star

Abstract: The UK Atomic Energy Authority (UKAEA) has been set a grand challenge – to help UK achieve Net Zero by delivering fusion energy to grid in the 2040's. Fusion is the process that powers our Sun. Recreating the fusion process here on Earth by confining a hot, thermonuclear plasma inside a "magnetic bottle" or "tokamak" however is challenging, and has often been referred to as the "holy grail" of power production. Since the advent of modern computing and with a growing understanding of the complex physics that describes the fusion plasma, our ability to design a viable fusion reactor concept has for a long time been heralded an "exascale challenge". The exascale itself is, of course, nearly upon us. Rob will briefly describe one of the most challenging areas of tokamak plasma modelling being tackled through the ExCALIBUR programme – the so-called plasma "exhaust", a region of the machine where hot plasma leaves the confined plasma core and encounters the tokamak first wall. This part of the machine is a well established, multi-physics, multi-scale problem. We have therefore assembled a rainbow team of UK researchers and developers to co-design a brand new, "actionable", scalable and performance portable platform that will be capable of exploiting the world's first exascale supercomputers. The project connects across half a dozen ExCALIBUR Cross Cutting (XC) Theme projects together with a Hardware and Enabling Software (H&ES) programme necessary for systems co-design. As we enter the commercial fusion and exascale era, supercomputing and interdisciplinary collaboration will be key to success – a very demanding timeline necessitates that the reactor design process and our journey up the engineering "S-curve" must take place "in-silico" rather than in the real world via test-based design. The ExCALIBUR programme, with its underpinning four pillars as a core element of UKAEA's mission couldn't, therefore, be more timely.

Bio: Rob is Department Manager and Programme Lead for Advanced Computing at UKAEA. He is leading a rapidly growing team of experts into the exascale and AI era of supercomputing with a mission to "accelerate the delivery of commercial fusion energy by exploiting advances in data science and extreme scale computing". Programmes managed by Rob and his group leaders include the £5M ExCALIBUR High Priority Fusion use case (project NEPTUNE), the Advanced Computing Programme within UKAEA's EPSRC Fusion Grant (focusing upon low TRL research), a new collaboration with STFC HNCDI to grow an Extreme Scale Computing Centre to Advance Fusion Energy at Hartree Centre, Daresbury Laboratory: https://www.hpcwire.com/2021/08/19/uks-new-extreme-scale-computing-center-to-advance-fusion-energy/, collaborations with the University of Cambridge and the Cambridge/Intel Open Exascale Lab (https://www.exascale.hpc.cam.ac.uk) and a new partnership with the University of Manchester focusing upon Fusion Digital Engineering: https://www.gov.uk/government/news/fusion-research-partnership-agreed-between-ukaea-and-the-university-of-manchester).

Rob gained a PhD. in High Energy Physics from the University of Manchester in 1995 (based at CERN) and has worked for UKAEA ever since. His early career was as a plasma modeller and experimentalist, working on the pioneering START Spherical Tokamak and then the EPSRC flagship tokamak MAST. Around a decade ago he became interested in GPU programming after visiting the Machine Evaluation Workshop (predecessor to CIUK) where every stand in the vendor space seemed to be showcasing GPU solutions. After a decade of programming CUDA he must now be content to let his team enjoy delivering

technical work while he instead concentrates on how to scale operations in order to deliver the world's first commercial fusion reactor, a machine called STEP: https://step.ukaea.uk.

EXCALIBUR – EXPLOITING THE EXASCALE TO BOTTLE A STAR

Rob Akers, UKAEA

CIUK – Thursday Dec 9th 2021, 14:20-14:40

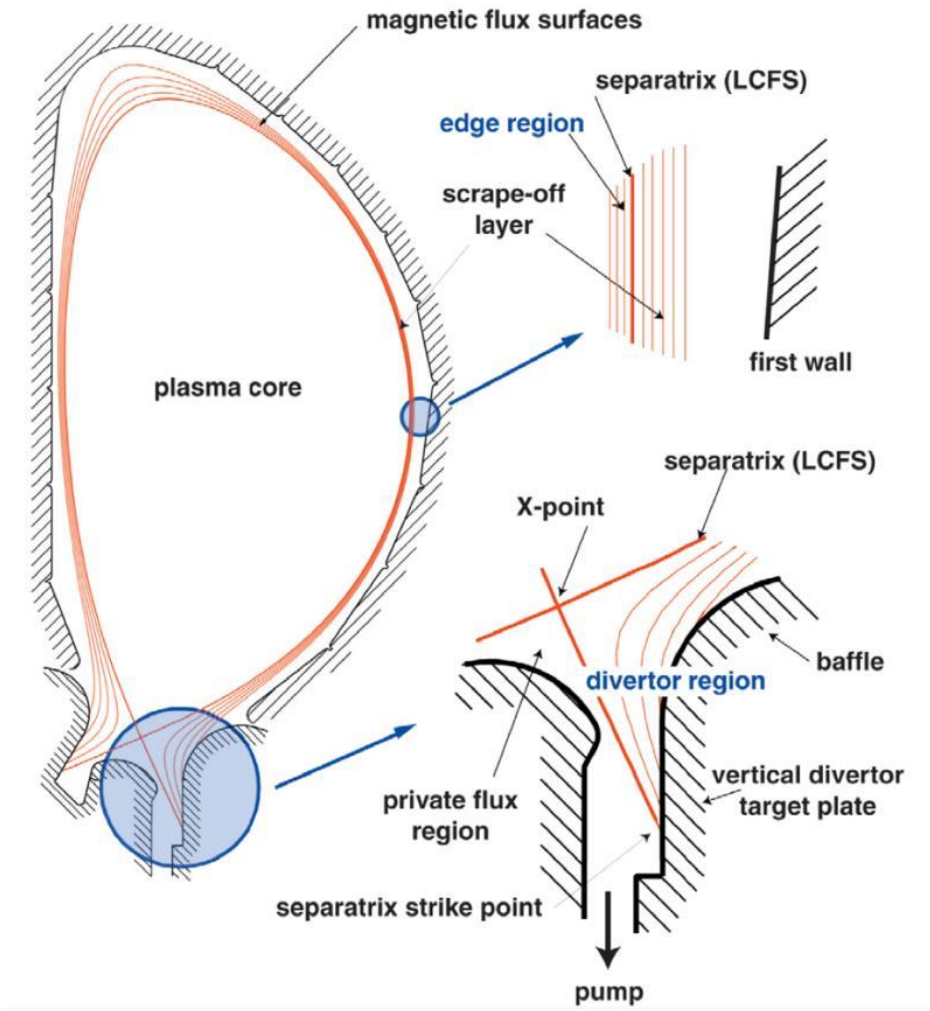
Why “ExCALIBUR”?

….build a UK wide “rainbow” team – take a multi / interdisciplinary approach

- Legacy codes tend not to be “actionable” **- VVUQ**
- Codes usually designed for “science”, not “engineering” **– MOR** methods not built in
- Codes designed in isolation – not designed for **“coupling”**
- Codes are always designed for one architecture – inflexible – not designed for **emerging architectures** or with **performance portability** in mind
- Codes are nearly always incredibly hard to adapt – many started off as PhD. projects – lack of **DSL based APIs**

NEPTUNE High Priority Use Case

Neutrals and Plasma Turbulence Numerics for the Exascale

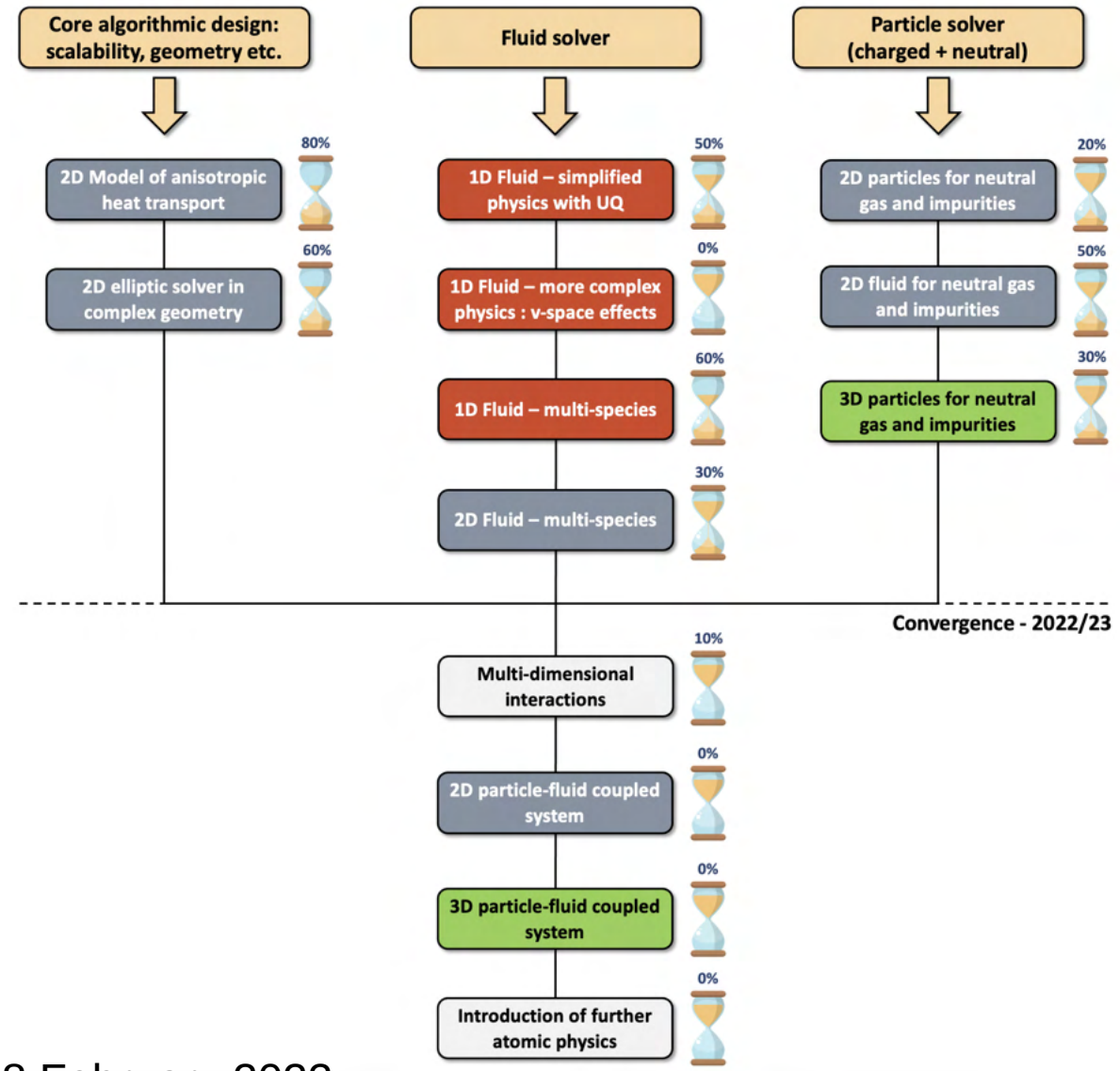


Modelling the plasma edge or 'exhaust'

- A long established exascale **grand-challenge, Multi-physics, Multi-scale** problem
- Complexity – **turbulence**, **atomic physics** etc.
- Incomplete mathematics ($1M Millennium Prize)
- For plasma, kinetic effects can't be ignored – requires **coupled fluid + particles**

Requires an interdisciplinary rainbow team...

Development by Proxyapps



NEPTUNE

1. Fluid solver: High order spectral-hp
2. …coupled to FEM-PIC
3. Performance Portability (SYCL/OneAPI)
4. Next gen Preconditioners (MCMC and Structure preserving methods)
5. Built in UQ and Model Order Reduction
6. Time stepping – parallel in time?
7. DSL front end (Julia?)
8. Converge Proxyapp based research in 2022 to start building full code/library

Nigel Wood (Met Office)

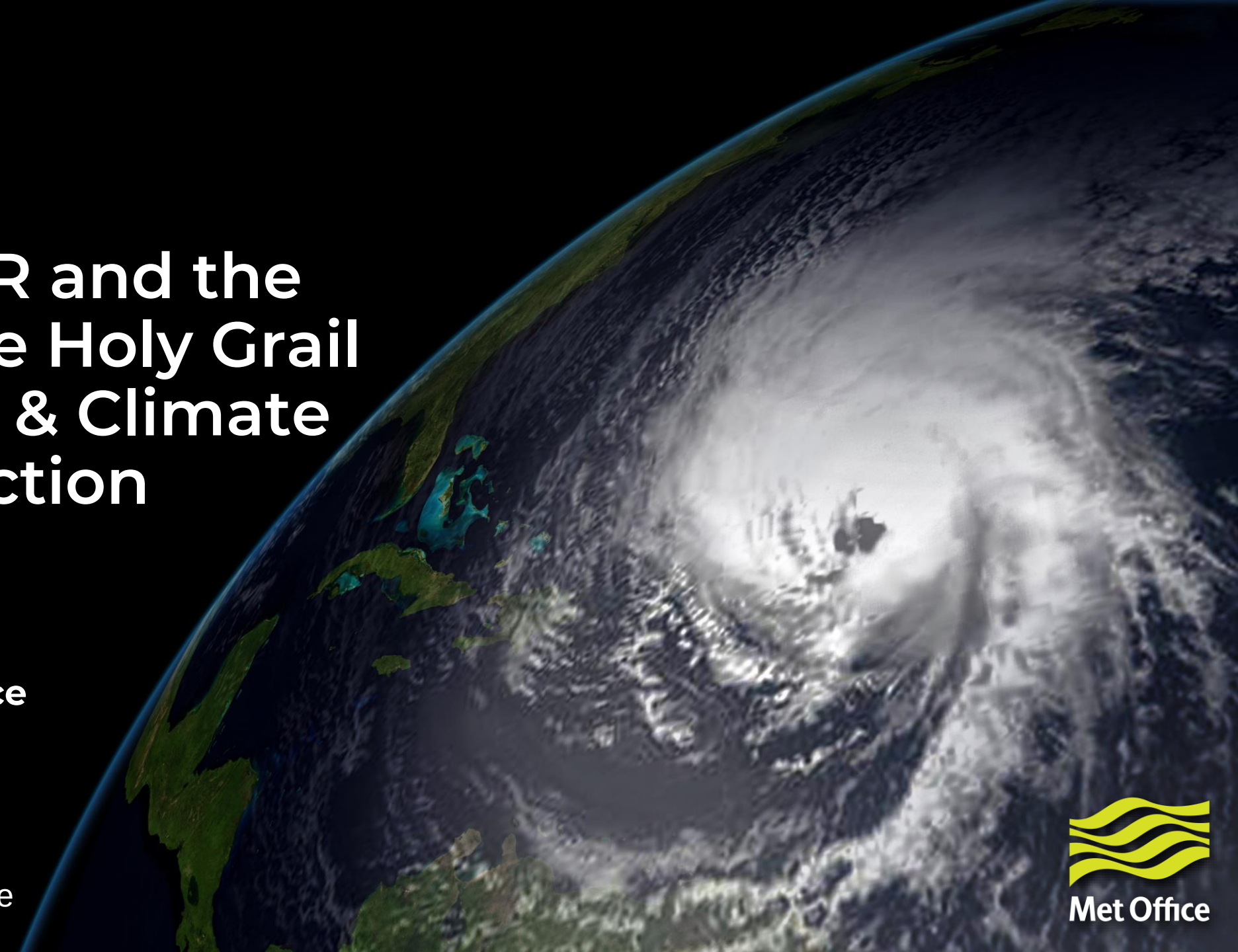
ExCALIBUR and the quest for the holy grail of weather & climate prediction

Abstract: At the heart of the UK's national capability in weather & climate prediction is a complex, unified modelling system which comprises an observation processing system, a data assimilation system, a dynamical model of the atmosphere, a number of marine modelling systems, a land surface model, a chemistry model, a variety of subgrid-scale parametrisations, as well as all the infrastructure to manage the data flow into, through, and out of the coupled system of models. I will briefly introduce this system and discuss why being able to fully exploit Exascale computers is so important to us in our quest to continually improve the accuracy of weather & climate predictions. This will explain what might be referred to as the holy grail of weather & climate prediction capability. For most of the components of the system, exploiting Exascale machines requires a redesign of the software infrastructure. For some of the components, a complete redesign of the algorithmic approach is also required. Co-design, application of the principle of a separation of concerns, data science, and software engineering expertise are all key to the success of this venture; they are also the pillars of ExCALIBUR. I will outline the important role that ExCALIBUR has in preparing the path to our holy grail.

Bio: I am the Senior Science Supplier for the Met Office's Next Generation Modelling Systems programme. This is one of the Met Office's corporate strategic actions which aims to reformulate and redesign our complete weather and climate research and operational/production systems, including oceans and the environment, to allow us and our partners to fully exploit future generations of supercomputer for the benefits of society. I am also the Met Office Senior Science Supplier for the Exascale Computing Algorithms and Infrastructure for the Benefit of the UK (ExCALIBUR) Project which has similar aims to NGMS but applied more broadly across the UK's supercomputing landscape. Both the Next Generation Modelling Systems programme and ExCALIBUR are targeting application on supercomputers of the mid-2020s and beyond.

After studying maths at university, I joined the Met Office to work on parametrising the effects on large-scale flows of unresolved turbulent flow over hills. This led to the award of my PhD from Reading University in 1992. At the turn of the millennium I migrated from the area of physical parametrisations to the Met Office's Dynamics Research group becoming responsible for the design and implementation of our operational dynamical core. Since that went operational in 2014 my attention has increasingly turned to ensuring that our modelling systems are ready to fully exploit the supercomputers of the future. This has included leading the GungHo project to develop a highly scalable dynamical core that at least matches the accuracy of our current one. GungHo and the associated infrastructure project, LFRic, were the precursors to the Next Generation Modelling Systems Programme.

ExCALIBUR and the Quest for the Holy Grail of Weather & Climate Prediction



Nigel Wood, Met Office

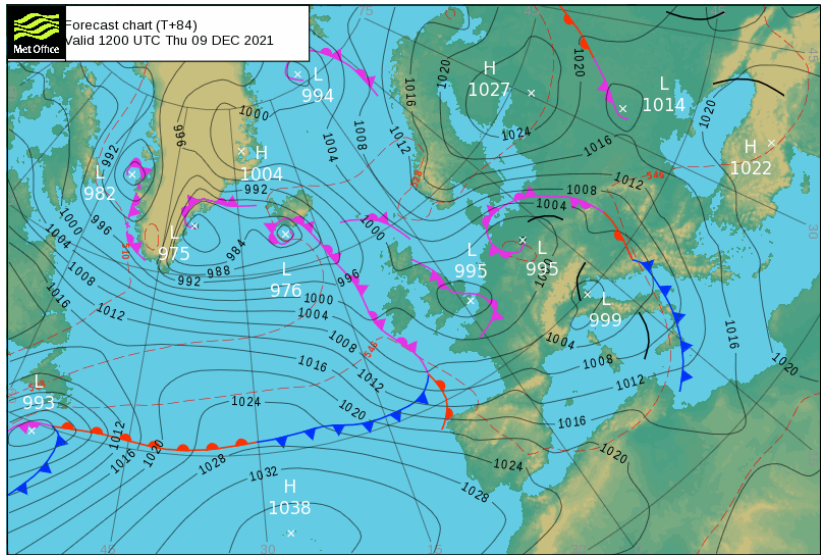
December 2021

UK's Weather & Climate prediction system

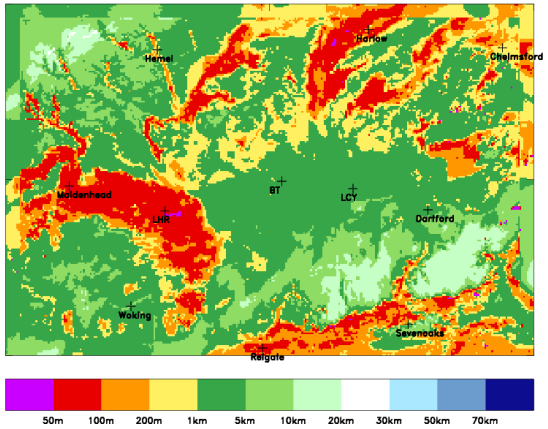
The challenge of a unified approach

- Operational forecasts
  - ➢Global (resolution approx. 10km)
  - ➢Regional (resolution approx. 1.5km)

10 km



300 m

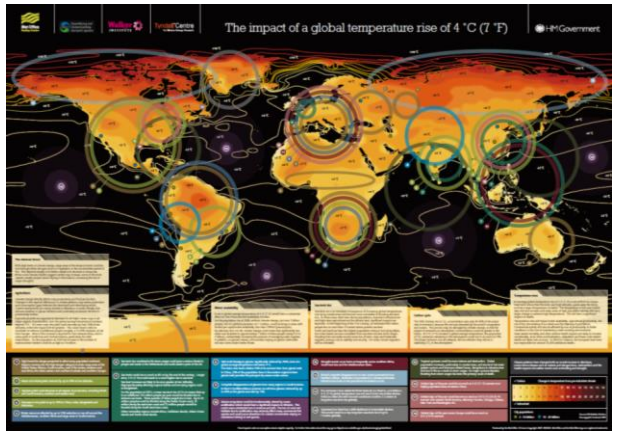


- Seasonal predictions
  - ➢Resolution approx. 60km

Unified ⇒ Same solver, same parametrisations, same code base for all

- Global and regional climate predictions
  - ➢Global resolution around 120km
  - ➢Regional around 4-1.5km
  - ➢Run for 10-100-… years

300 km



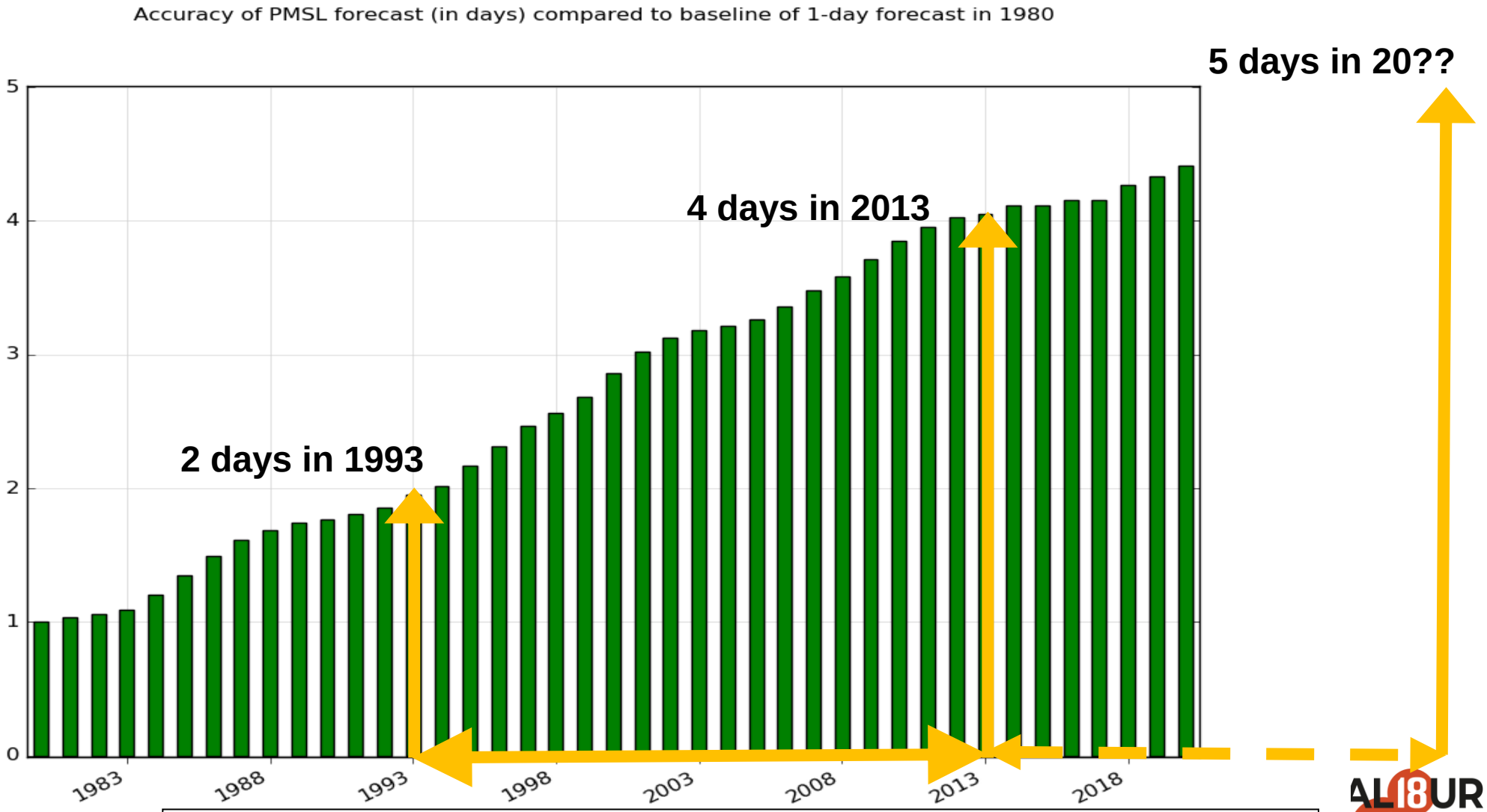
EXCALIBUR

“The quiet revolution”*: ≈ 1 day’s lead time per decade

108 hour forecast today is as accurate as the 24 hour forecast was in 1980

Met Office

“…impact of Numerical Weather Prediction among greatest of any area of science… comparable to simulation of human brain and evolution of early universe”*



*Bauer, Thorpe, Brunet (2015) Nature

How do we achieve this?

What is the holy grail?

Our challenge

Why Exascale?

- Currently we simulate the world's weather at 10 km intervals
- To complete the 7 day forecast in 1 hour needs a petascale machine (16 Pflops) (we use 19,000 cores)

- To get to 5 km means 2x2 more cells and a 2 times smaller interval in time
  - **O(10)** increase in compute power & data
- To get to 1 km means 10x10 more cells and a 10 times smaller interval in time
  - **O(1000)** increase in compute power & data

GungHo/LFRic/PSyclone

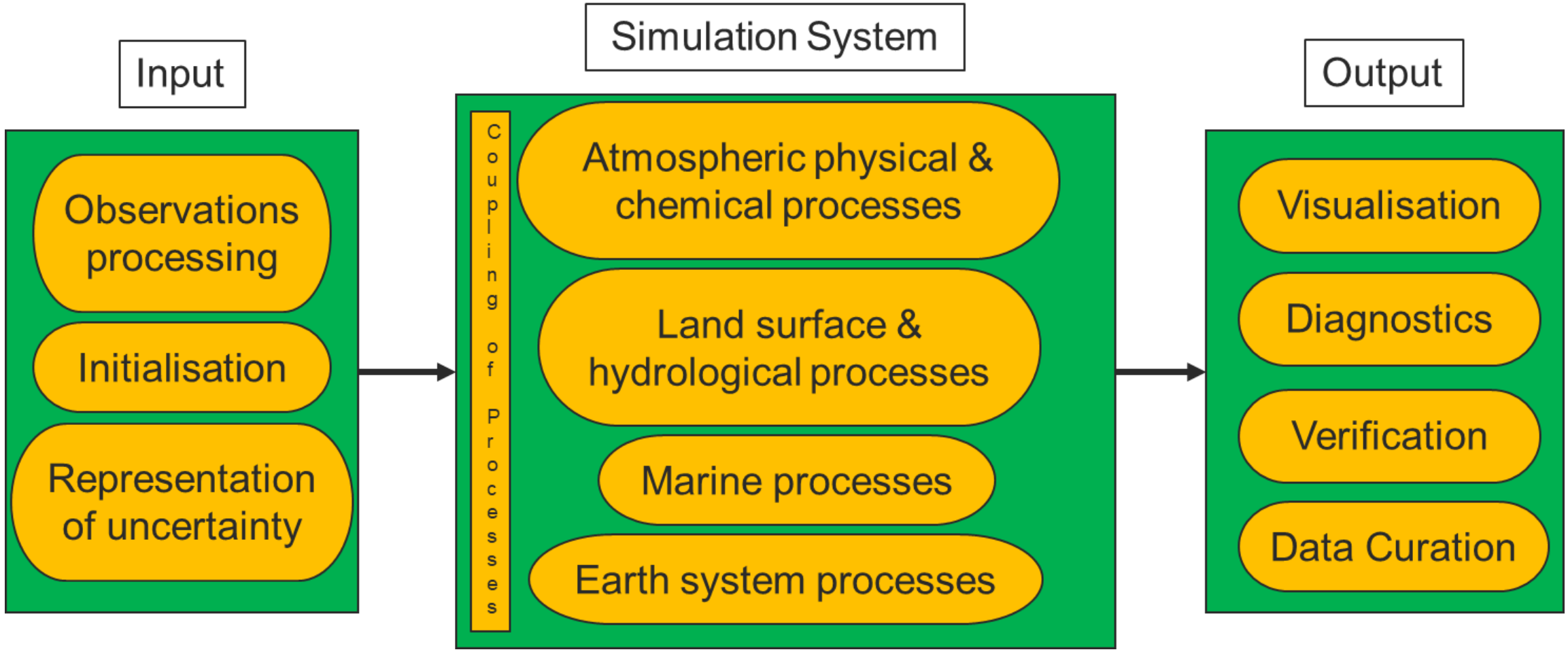
Not ExCALIBUR funded but key to delivery of next generation capability

March 2021 saw delivery of first capability of new GungHo/LFRic/PSyclone based global atmosphere model

But a lot more to do…

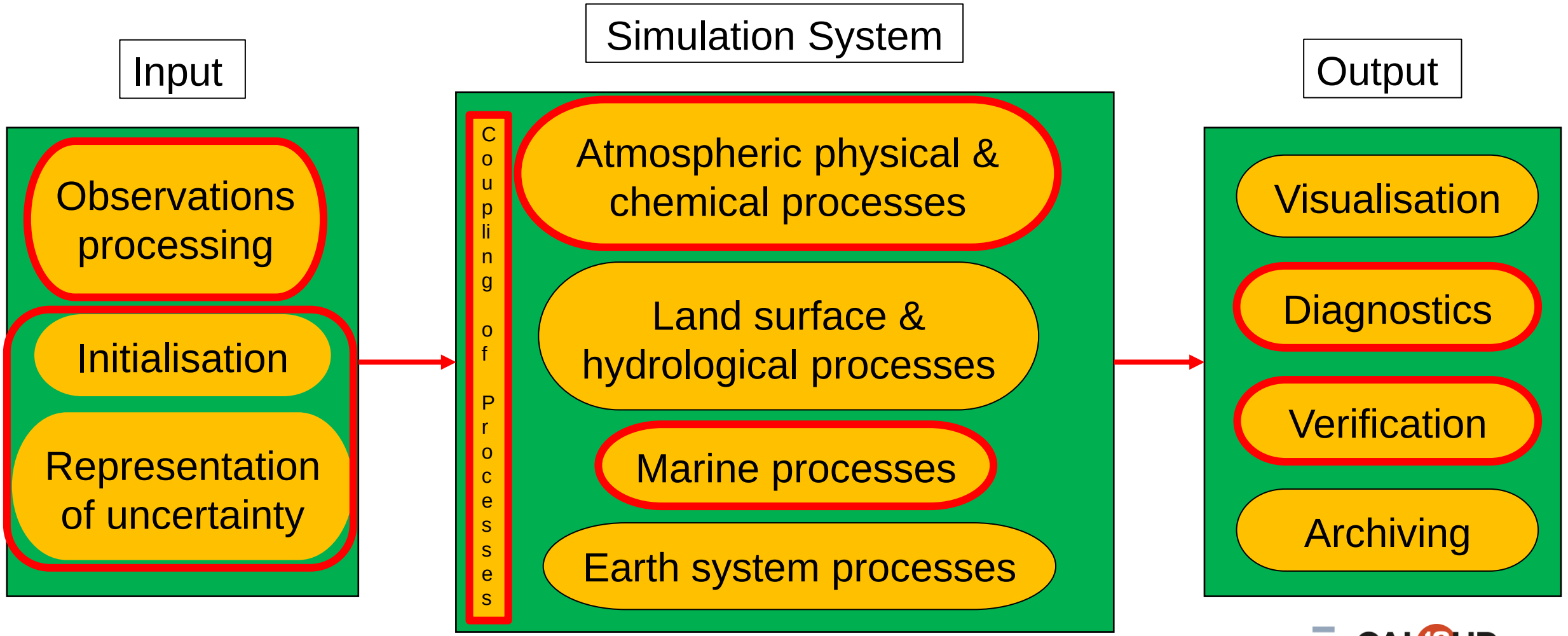
Weather & Climate prediction system

Schematic



Weather & Climate prediction system

Activities



Some example activities

Marine Systems (NEMO) design system

Delivering extension of PSyclone DSL to existing marine systems without rewriting code (fruition of GOCEAN POC)

Atmospheric Model data layout and memory access design system

Delivering mixed precision capability & flexibility in memory layout

Atmospheric observation pre-processing and assimilation

Delivering new flexible framework deployable across different architectures

Performance Portability for Existing Marine-Systems Models

NEMO

MEDUSA

NEMOVAR

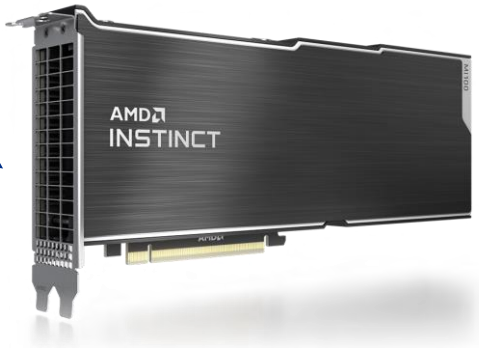
WAVEWATCH III

PSyclone

OpenMP

OpenACC

OpenMP



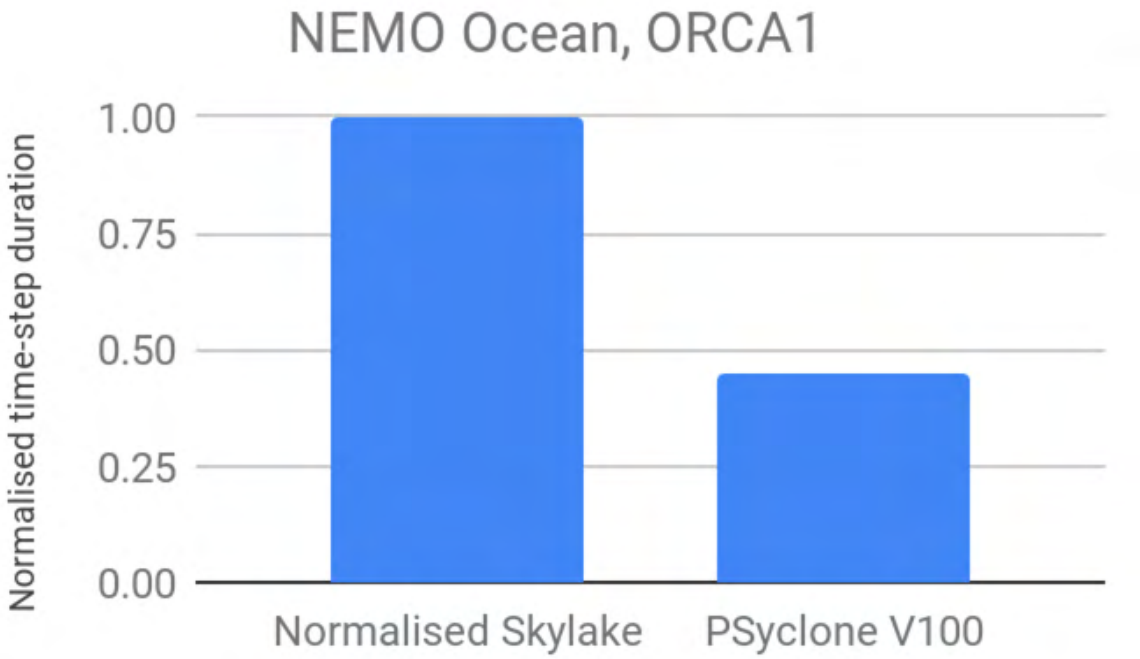
UKRI Science and Technology Facilities Council
Hartree Centre

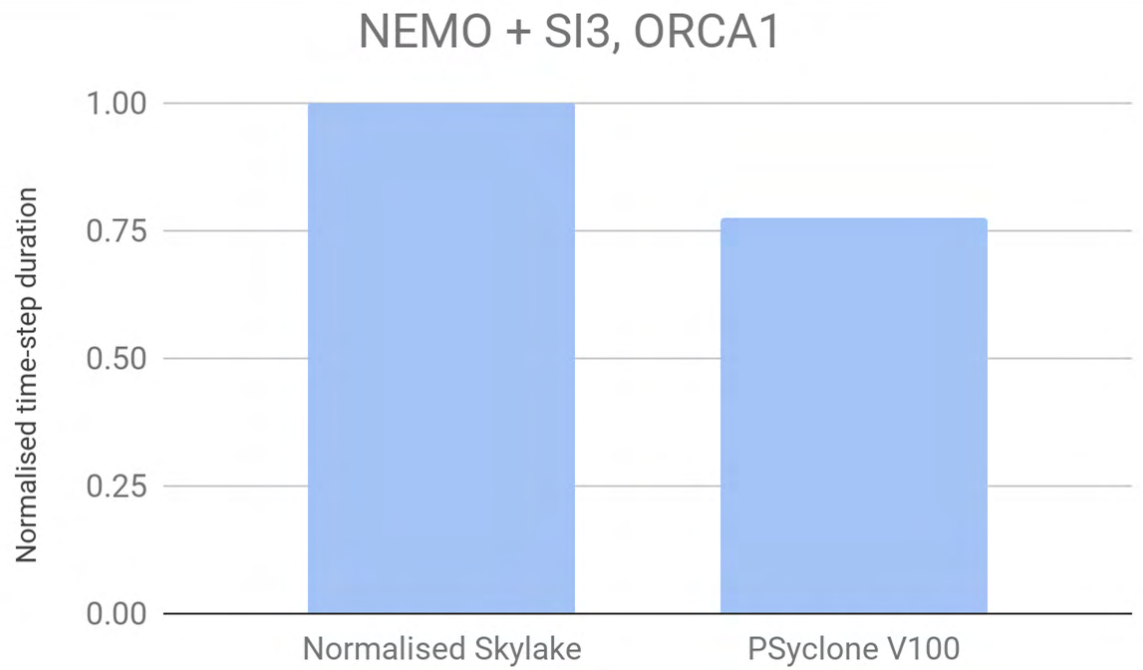
National Oceanography Centre

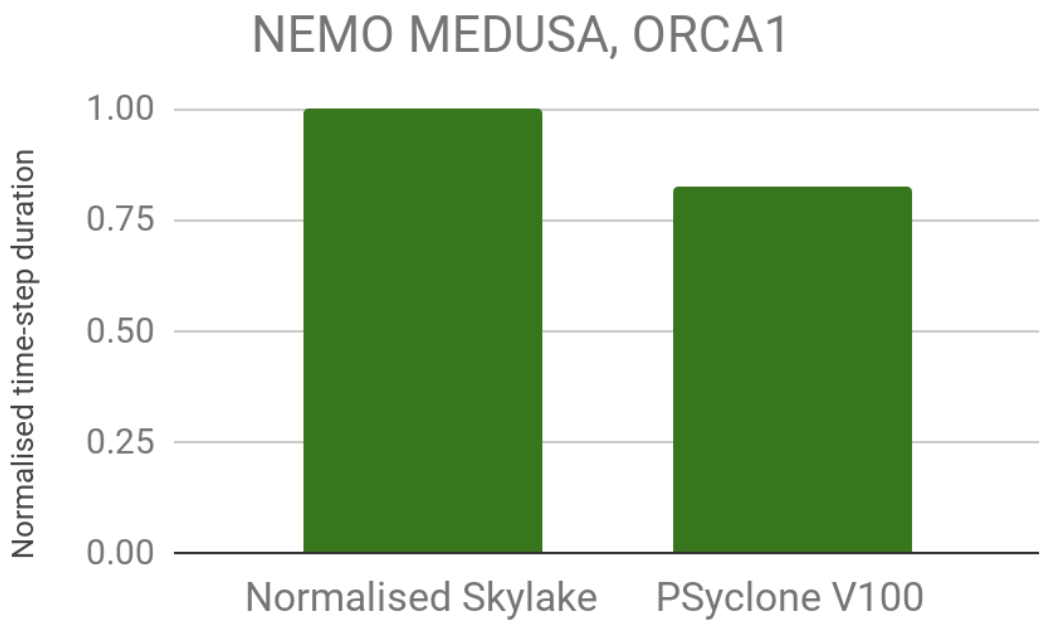
National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

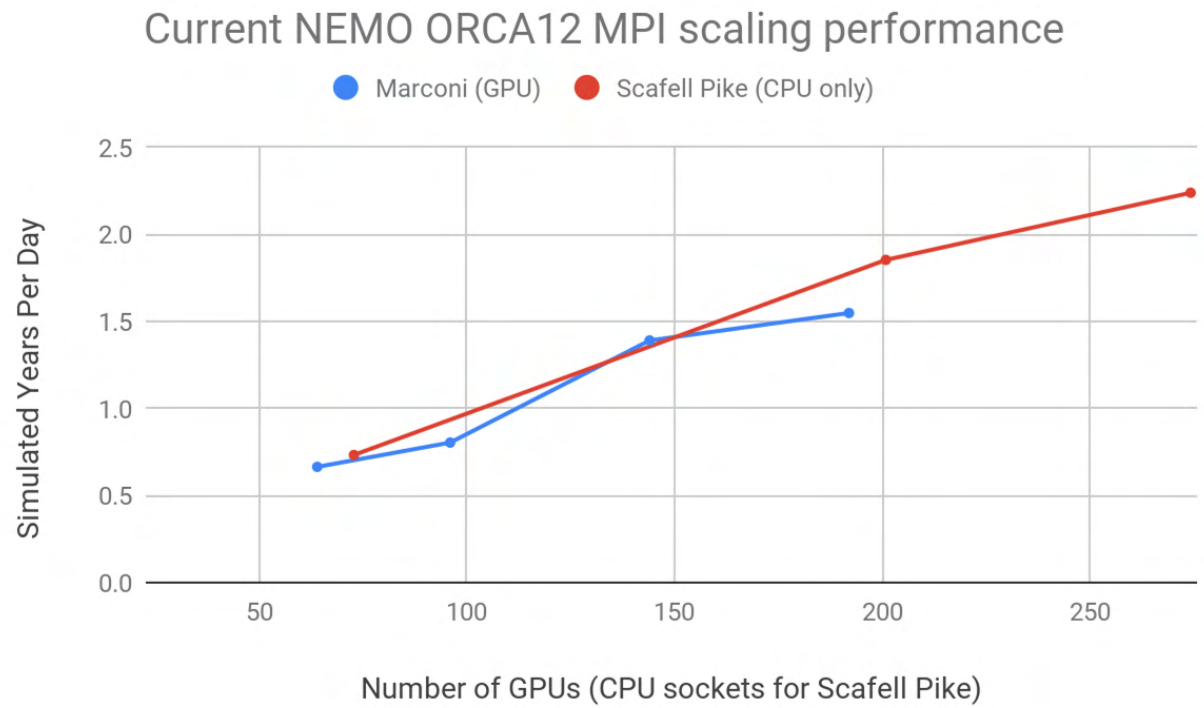
Met Office

NVIDIA

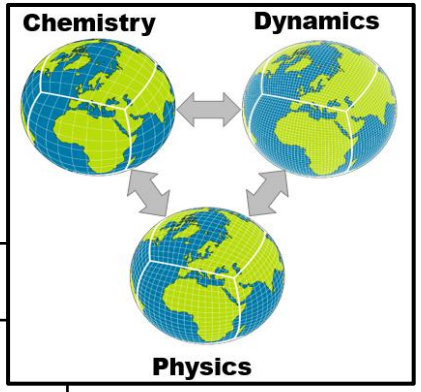

NEMO Ocean, ORCA1
Normalised time-step duration
1.00
0.75
0.50
0.25
0.00
Normalised Skylake
PSyclone V100


NEMO + SI3, ORCA1
Normalised time-step duration
1.00
0.75
0.50
0.25
0.00
Normalised Skylake
PSyclone V100

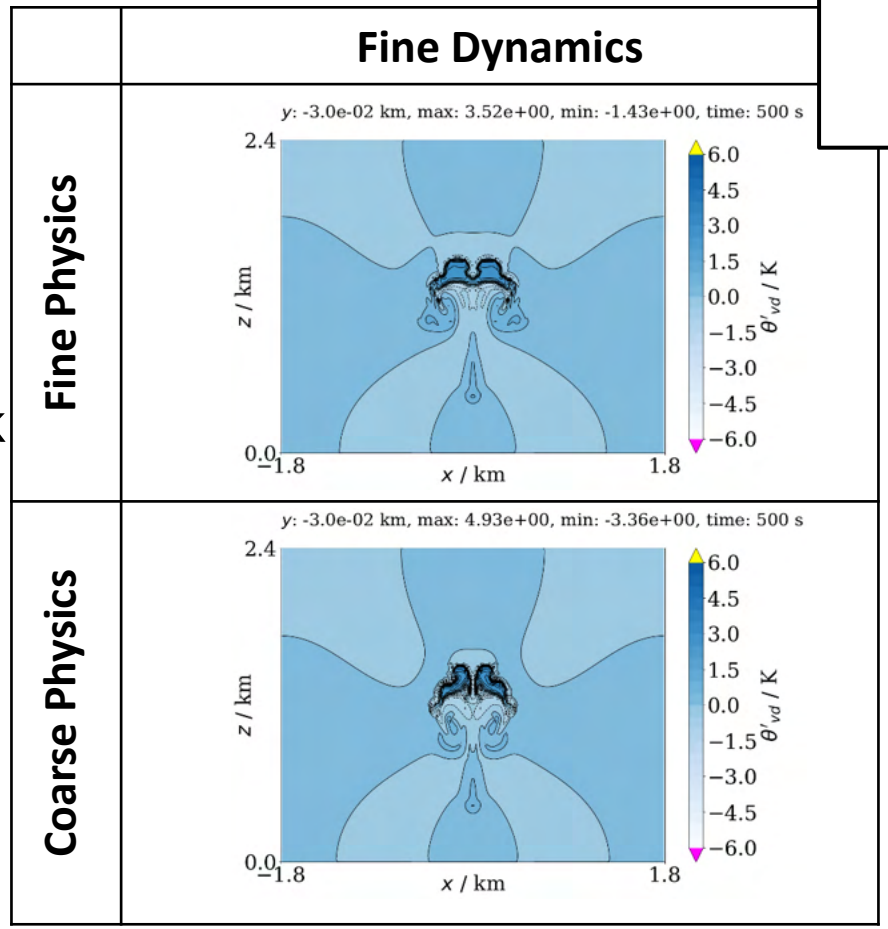

NEMO MEDUSA, ORCA1
Normalised time-step duration
1.00
0.75
0.50
0.25
0.00
Normalised Skylake
PSyclone V100


Current NEMO ORCA12 MPI scaling performance
Marconi (GPU)
Scafell Pike (CPU only)
Simulated Years Per Day
2.5
2.0
1.5
1.0
0.5
0.0
50
100
150
200
250
Number of GPUs (CPU sockets for Scafell Pike)

Atmospheric Model data layout, memory access design system, and spatial decoupling of processes



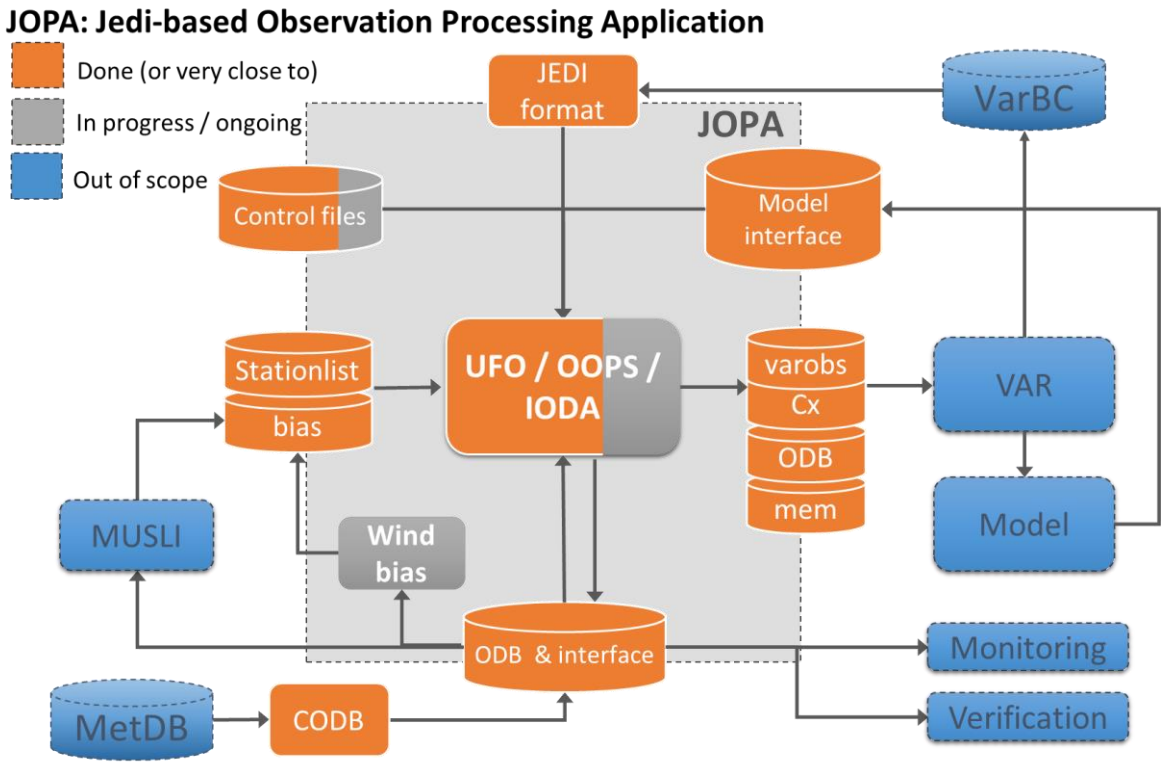
- **Only do what is needed:** Example miniapp implementation capable of spatial splitting of combinations of transport, dynamics and UM physics parametrizations.
- **Only do it to the accuracy needed:** framework in place for mixed-precision, but suspected compiler in Gnu means refactoring needed to demonstrate benefit.
- **Do it using the optimal data layout:** Implementation of the "i-first" data transpose to the microphysics code in the LFRic basic-gal model shows a **4x speed up** in that part of the model.



A moist bubble test showing the impact of running the physics (condensation/cloud scheme) at a coarser resolution.

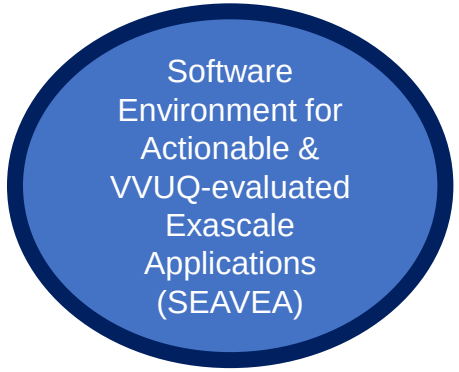
Next-Generation observation processing

Key infrastructure developments achieved



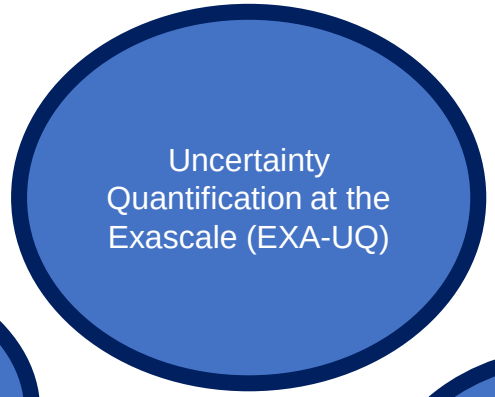
- Automatic Schema
- Global UM interface
  (incl. background error / Rho-Theta level)
- VarObs/CX, interfaces
- Multiple UFO filters (thinning, Composite operator, where clause , etc …)
- ODB full backend to IODA
- Comparison suite
- Generic auxiliary files interface (incl. netcdf / csv)
  used by obs. error, VarBC coeff., Static bias, station list
- Improved ObsSpace Group (Derived Observation)
- Generic (& user defined) QC diagnostics flag

Cross-Cutting Activities

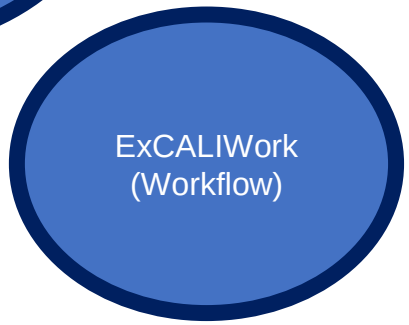


UCL Brunel

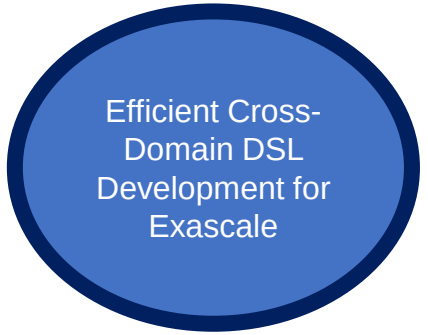
Reading & Cambridge NCAS



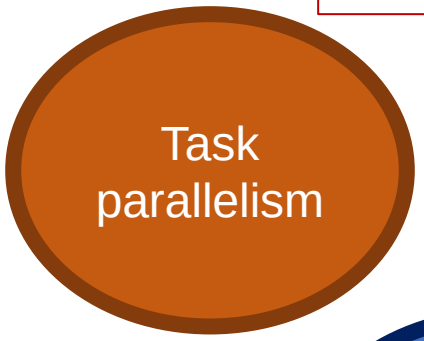
Exeter



Reading NCAS; StackHPC; DDN

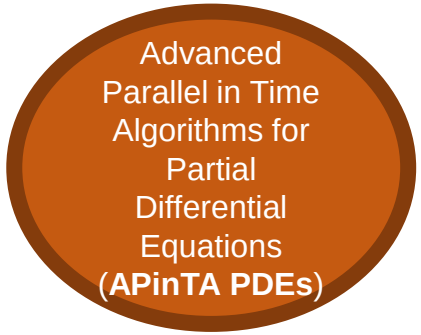


Edinburgh
Imperial College London



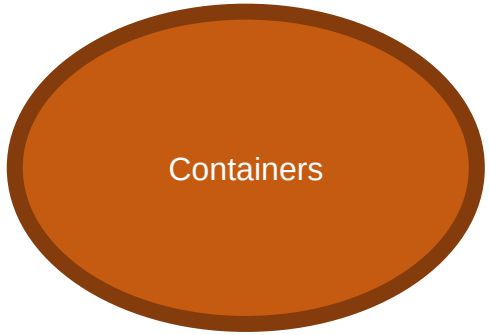
Durham & STFC Hartree

Met Office

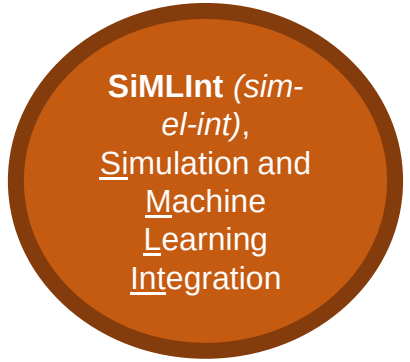


Exeter & Imperial

Met Office



Met Office



EPCC

Thank you! Questions?

See https://excalibur.ac.uk/ for more

Martyn Guest (ARCCA, Cardiff University)

Performance of Computational Chemistry Codes. An Analysis of Molecular Dynamics and Electronic Structure Applications on Multi-core Processors

Abstract: This session will overview application performance on a variety of clusters, focusing on the Intel Ice Lake and AMD EPYC Milan family of processors. Using the Intel Skylake Gold 6148 as the baseline, an assessment is made across a variety of Ice Lake (8358, 8352Y, 8368Q and 8360Y) and Cascade Lake SKUs (e.g., the 9242-AP, 8280, 6252 and 6248), with system interconnects from both Mellanox and Intel. Attention is also focused on the AMD Milan (7713, 7513) and Rome SKUs (e.g., the 7702 and 7502). Our analysis is based on the familiar parallel benchmark performance using popular chemistry community codes – from Molecular Dynamics (DL_POLY, Gromacs and LAMMPS), Quantum Chemistry (GAMESS-UK) and Materials Science (VASP, CASTEP). The benefit of the Intel® oneAPI Toolkit is demonstrated throughout this analysis. To best capture a 'like for like' comparison amidst the extensive array of core densities, our analysis in now based on both a “node-by-node” and the more traditional “core-by-core” consideration.

Bio: Professor Martyn Guest has led a variety of high performance and distributed computing initiatives in the UK. He spent three years as Senior Chief Scientist and Group Leader of the HPC Chemistry Group at PNNL, before returning to the UK as Associate Director of Daresbury’s Computational Science and Engineering Department. He joined Cardiff University in April 2007 and is their Director of Advanced Research Computing as well as Technical Director of both the HPC Wales and successor Supercomputing Wales programme.

Martyn’s research interests cover the development and application of computational chemistry methods. He is lead author of the GAMESS-UK electronic structure program and has written or contributed to more than 250 journal articles.

COMPUTING INSIGHT UK 2021
Science and Technology Facilities Council
9 - 10 DECEMBER 2021
Manchester Central, UK
www.stfc.ac.uk/ciuk

Performance of Computational Chemistry Codes

An Analysis of Molecular Dynamics and Electronic Structure Applications on Multi-core Processors.

Jose Munoz, Christine Kitchen & Martyn Guest

Advanced Research Computing @ Cardiff (ARCCA) & Supercomputing Wales

Introduction and Overview

- Presentation part of our ongoing assessment of the performance of parallel application codes in materials & chemistry on high-end cluster systems.
- Focus here on systems featuring **current processors from AMD** (EPYC Milan SKUs – the 7713 & 7513 etc.) and **Intel** (Ice Lake & Cascade Lake-AP).
  - Baseline cluster: the Skylake (SKL) **Gold 6148/2.4 GHz** and **AMD EPYC Rome 7502 2.5Gz** cluster – "Hawk" – at Cardiff University.
  - Four Intel Xeon Ice Lake clusters, the 32-core Platinum **8358** (2.6 GHz) and **8352Y** (2.2 GHz), the 38-core **8368Q** (2.6 GHz), the 36-core **8360Y** (2.4GHz) plus other Cascade Lake and Cascade Lake-AP systems.
  - Two AMD EPYC Milan clusters featuring the 64-core **7713 2.0 GHz**.
  - Variety of **single-node systems** with varying SKUs (**Boston HPC**)
  - Consider performance of both synthetic and **end-user applications**. Latter include molecular simulation (**DL_POLY, Gromacs & Lammps**), materials modelling (**VASP**) and electronic structure (**GAMESS-UK**).
  - Scalability analysis by **processing elements (cores)** and by **nodes** (ARM Performance Reports). Baselined against **P100** & **V100** NVIDIA GPUs.

Methodology and Approach

1. Provide guidance based on evaluating performance that a **standard user** would experience on the systems
2. Target performance regime – **mid-range clusters**. No real effort invested in optimising the applications having used standard implementations when available
3. All benchmarks run on systems in general production i.e. not dedicated to this exercise – used standard Slurm job schedulers
4. CIUK'21 preparation one of **the most challenging to date**. Over ambitious target to evidence performance across a variety on MPI versions using Intel Parallel Studio XE e.g. 2018/4, 2019/5, 2019/12 and 2020/4 variants.
   - Host of problems encountered, particularly on **AMD Rome and Milan** systems
   - Working code using 2019/5 on AMD Rome systems failed, as did all attempts to use earlier compilers/MPI libraries
   - 2019/12 worked in some cases but still led to failures with codes hanging at arbitrary core counts
   - Saviour was **Intel oneapi** – no such issues when deploying that environment. Issues remain unresolved but **chewed up considerable time**!

AMD EPYC Rome multi-chip package

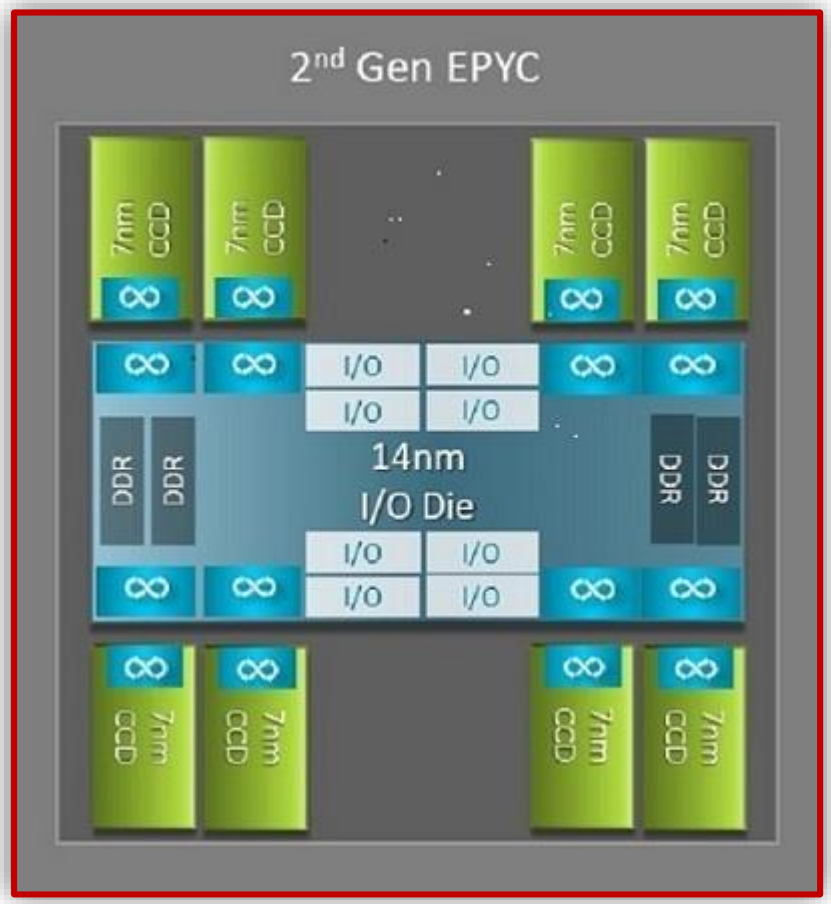


Figure. Rome multi-chip package with one central IO die and up to eight-core dies.

- In Rome, each processor is a multi-chip package comprised of up to 9 **chiplets** as shown in the Figure.
- There is **one central 14nm I/O die** that contains all the I/O and memory functions – memory controllers, Infinity fabric links within the socket and inter-socket connectivity, and PCI-e.
- There are **eight memory controllers per socket** that support eight memory channels running DDR4 at 3200 MT/s. A single-socket server can support up to 130 PCIe Gen4 lanes. A dual-socket system can support up to **160 PCIe Gen4 lanes**.

AMD EPYC Rome multi-chip package

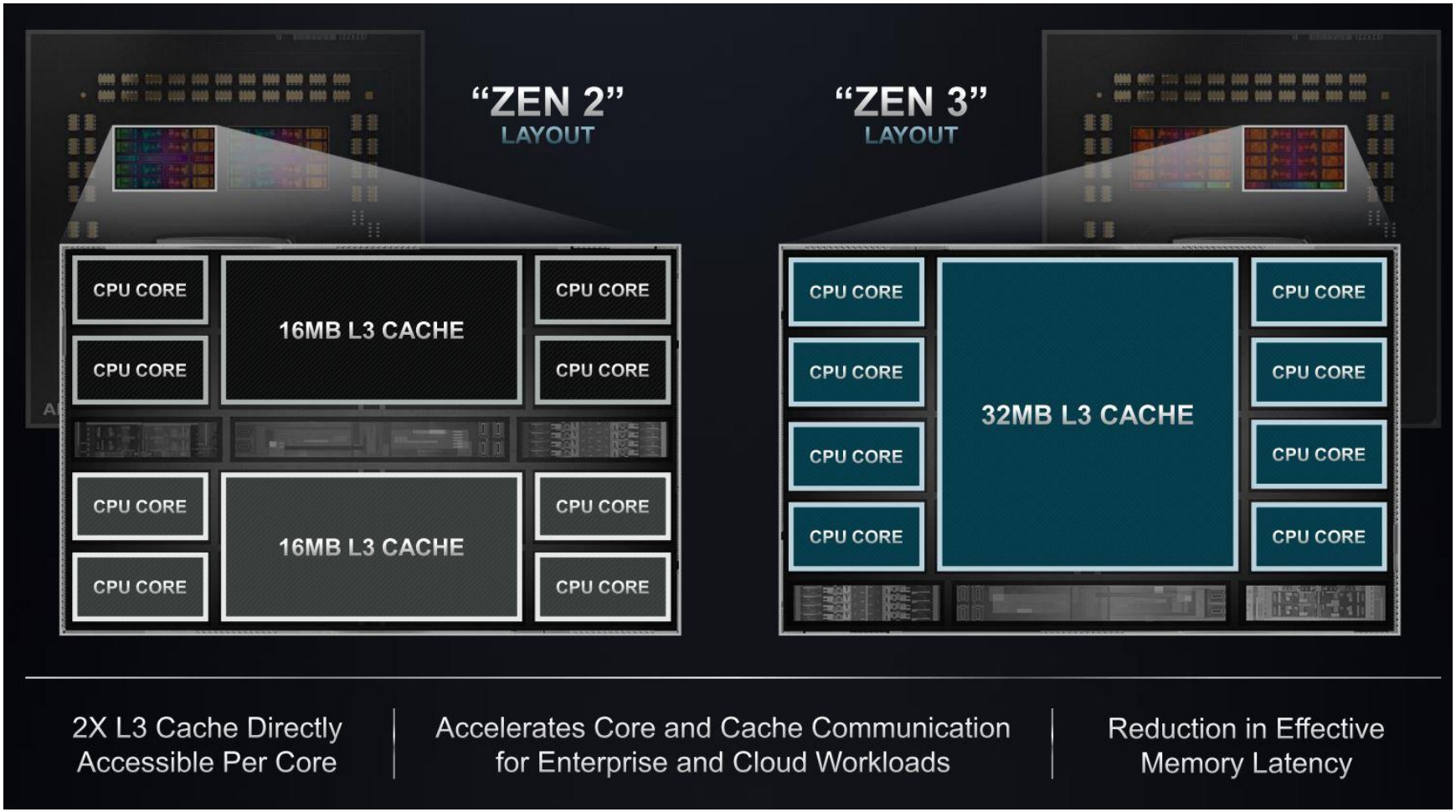
Figure A CCX with four cores and shared 16MB L3 cache

- Surrounding the central IO die are up to **eight 7nm core chiplets**. The core chiplet is called a **Core Cache die** or CCD.
- Each CCD has CPU cores based on the **Zen2 micro-architecture**, L2 cache and 32MB L3 cache. The CCD itself has two Core Cache Complexes (CCX), each CCX has up to four cores and 16MB of L3 cache.
- The figure shows a CCX.
- The different Rome CPU models have different numbers of cores, but all have one central IO die.

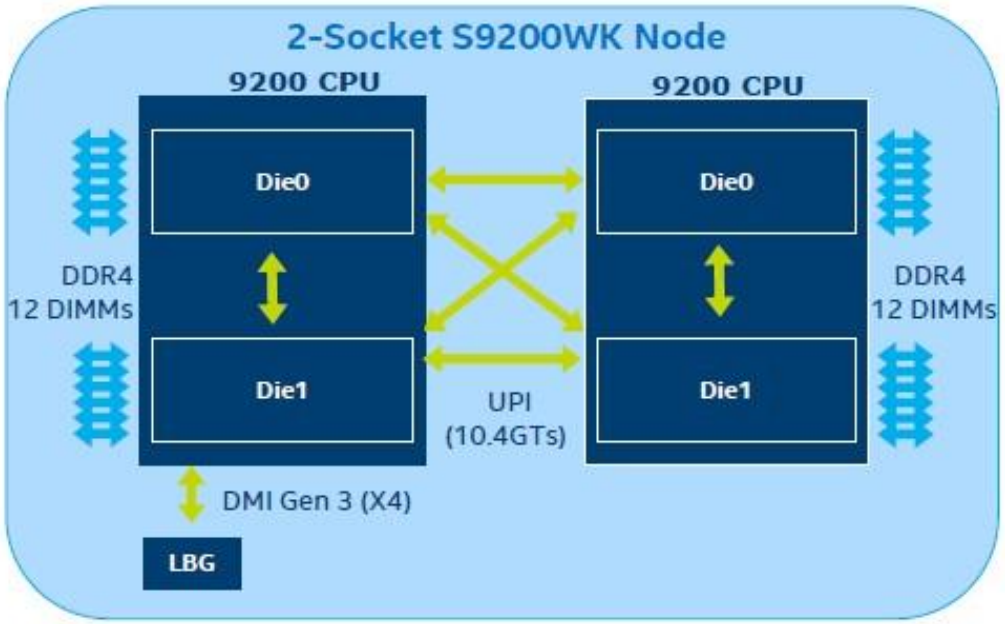
Rome CPU models evaluated in this study

CPU	Cores per Socket	Config	Base Clock	TDP
7742	**64c**	**4c per CCX**	**2.25 GHz**	**225W**
7502	**32c**	**4c per CCX**	**2.5 GHz**	**180W**
7452	32c	4c per CCX	2.35 GHz	155W
7402	24c	3c per CCX	2.8 GHz	180W

AMD EPYC Milan multi-chip package



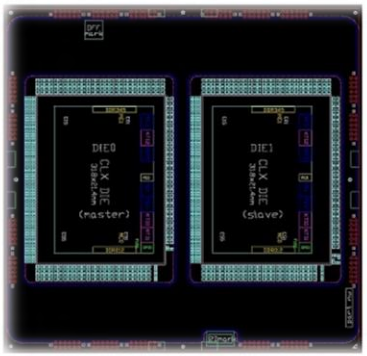
Intel Xeon Platinum 9200 Processors

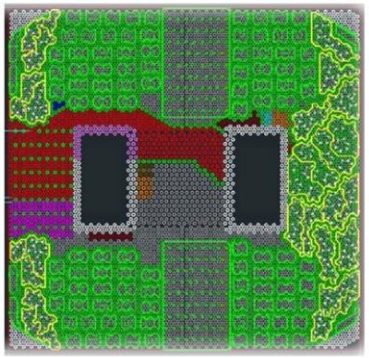


- Intel® Xeon® Platinum 9200 Processors consist of two die in a BGA package
- Multi-chip processor with single hop latency for any of the CPU die to memory in a 2S node
- Key IO/mem features include:
  - 12 ch DDR4 2933 MT/s per CPU
  - 4 UPI x20 wide at 10.4GTs per CPU
  - x80 PCIe G3 lanes per 2S Node in Intel® Server Systems S9200WK*

* Intel® Server Systems S9200WK supports 2 x16 Gen3 slots (per 1U node); 4 x16 Gen3 slots (per 2U node)

Core Count	2-Socket STREAM-TRIAD (GB/s)	Per Core STREAM-TRIAD (GB/s)
	Up to	Up to
56	407	3.6
48	404	4.2
32	397	6.2

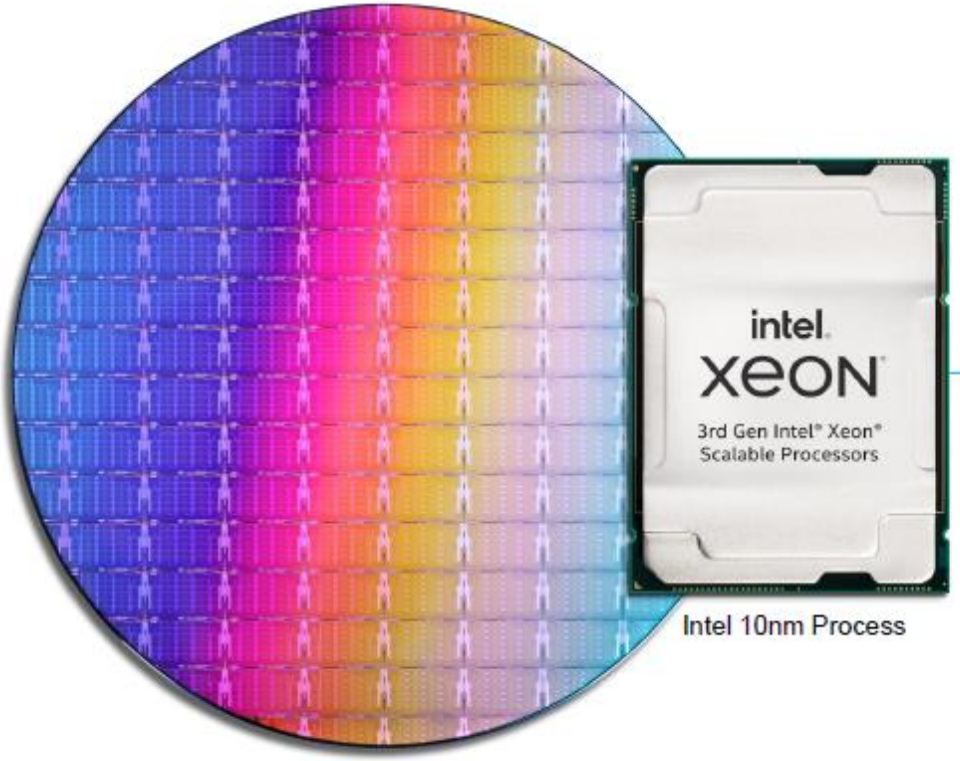




Intel 3rd Generation Xeon Scalable Processors

3rd Gen Intel® Xeon® Scalable processors

Performance made flexible



Up to 40 cores
per processor

20% IPC improvement
28 core, ISO Freq, ISO compiler

1.46x average performance increase
Geomean of Integer, Floating Point, Stream Triad, LINPACK
8380 vs. 8280

1.74x AI inference increase
8380 vs. 8280 BERT

2.65x average performance increase
vs. 5-year-old system
8380 vs. E5-2699v4

Intel Xeon Cascade Lake and Ice Lake

	Cascade Lake (per core)	Ice Lake (per core)
Out-of-order Window	224	352
In-flight Loads + Stores	72 + 56	128 + 72
Scheduler Entries	97	160
Register Files – Integer + FP	180 + 168	280 +224
Allocation Queue	64/thread	70/thread; 140/1 thread
L1D Cache (KB)	32	48
L2 Unified TLB (STLB)	1.5K	2K
STLB-IG Page support	**16**	**1024 (shared w/4K)**
STLB-IG Page support	**16**	**1024 (shared w/4K)**
Mid-level Cache (MB)	1	1.25

Performance of Computational Chemistry Codes

Systems, Software and Installation

Baseline Cluster System

Supercomputing Wales "Hawk" Cluster Configuration	
"Phase-1" - Intel Skylake Partition	201 nodes, totalling 8,040 cores, 46.080 TB total memory. • CPU: 2 x Intel(R) Xeon(R) Skylake Gold 6148 CPU @ 2.40GHz with 20 cores each; RAM: 192 GB, 384GB on high memory and GPU nodes; GPU: 26 x nVidia P100 GPUs with 16GB of RAM on 13 nodes. • Mellanox IB/EDR infiniband interconnect.
"Phase-2" AMD Rome Partition	64 nodes, totalling 4,096 cores, 32 TB total memory. • CPU: 2 x AMD EPYC Rome 7502 CPU @ 2.50GHz with 32 cores each; RAM: 512 GB, and GPU nodes; GPU: 30 x nVidia V100 GPUs with 16GB of RAM on 15 nodes
Researcher Funded Partitions	• 4,616 cores – Intel Skylake dedicated researcher expansion • 2,064 cores – Intel Broadwell and Haswell Raven migrated sub-system nodes

The available compute hardware is managed by the **Slurm job scheduler** and organised into 'partitions' of similar type/purpose.

Intel Xeon Cascade Lake Clusters

Cluster / Configuration
Dell Zenith cluster at the Dell Technologies HPC & AI Innovation Lab – Intel Xeon sub-systems with Mellanox HDR interconnect fabric running Slurm
• Intel **Xeon Gold 6230 Processor / 2.10 GHz**; # of CPU Cores: **20**; # of Threads: 40; Max Turbo Frequency: 3.90 GHz Base Clock: **2.10 GHz;** Cache 27.5 MB**;** Default TDP / TDP: 125W; Mellanox HDR **200Gb/s**
• Intel **Xeon Gold 6252 Processor / 2.10 GHz**; # of CPU Cores: **24**; # of Threads: 48; Max Turbo Frequency: 3.70 GHz Base Clock: **2.10 GHz;** Cache 35.75 MB**;** Default TDP / TDP: 150W; Mellanox HDR **200Gb/s**
• Intel **Xeon Gold 6248 Processor / 2.50 GHz**; # of CPU Cores: **20**; # of Threads: 40; Max Turbo Frequency: 3.90 GHz Base Clock: **2.50 GHz;** Cache 27.5 MB**;** Default TDP / TDP: 150W; Mellanox HDR **200Gb/s**
• Intel **Xeon Platinum 8280 Processor / 2.70 GHz**; # of CPU Cores: **28**; # of Threads: 56; Max Turbo Frequency: 4.00 GHz Base Clock: **2.70 GHz;** Cache 38.5 MB**;** Default TDP / TDP: 205W; Mellanox HDR **200Gb/s**
Cascade Lake-AP cluster at Intel HPC Laboratory with Intel OPA interconnect fabric running Bright release 8.1, NVMe Lustre filesystem.
• 48 CLX-AP nodes (Cascade Lake Advanced Performance) 9242 processors / 2.3 GHz; ; # of CPU Cores: **48**; # of Threads: 96; Max Turbo Frequency: 3.80 GHz Base Clock: **2.30 GHz;** Cache 71.5 MB**;** Default TDP / TDP: 350W

Intel Xeon Ice Lake Clusters

Cluster / Configuration
Dell Zenith cluster at the Dell Technologies HPC & AI Innovation Lab – Intel Xeon sub-systems with Mellanox HDR interconnect fabric running Slurm
• 42 nodes × Intel **Xeon Platinum 8358 Processor / 2.60 GHz**; # of CPU Cores: **32**; # of Threads: 64; Max Turbo Frequency: 3.40 GHz Base Clock: **2.60 GHz;** Cache 48 MB**;** Default TDP / TDP: 250W; Mellanox HDR **200Gb/s**
• 60 nodes × Intel **Xeon Platinum 8352Y Processor / 2.20 GHz**; # of CPU Cores: **32**; # of Threads: 64; Max Turbo Frequency: 3.40 GHz Base Clock: **2.20 GHz;** Cache 48 MB**;** Default TDP / TDP: 205W; Mellanox HDR **200Gb/s**
Ice Lake clusters at Intel's OpenHPC Laboratory with Intel OPA interconnect fabric running Bright release 8.1, NVMe Lustre filesystem.
• 4 nodes × Intel **Xeon Platinum 8368Q Processor / 2.60 GHz**; # of CPU Cores: **38**; # of Threads: 76; Max Turbo Frequency: 3.70 GHz Base Clock: **2.60 GHz;** Cache 57 MB**;** Default TDP / TDP: 270W; **Cornelis OPA-2**
• 4 nodes × Intel **Xeon Platinum 8360Y Processor / 2.40 GHz**; # of CPU Cores: **36**; # of Threads: 72; Max Turbo Frequency: 3.50 GHz Base Clock: **2.40 GHz;** Cache 54 MB**;** Default TDP / TDP: 270W; **Cornelis OPA-2**
Atos Genji cluster at the Atos HPC, AI & QLM Benchmarking Centre with Mellanox HDR interconnect fabric running Slurm
• 32 nodes × Intel **Xeon Platinum 8358 Processor / 2.60 GHz**; # of CPU Cores: **32**; # of Threads: 64; (see above)

AMD EPYC Rome and Milan Clusters

Cluster / Configuration
Dell Minerva cluster at the Dell Technologies HPC & AI Innovation Lab – AMD EPYC Rome and Milan sub-systems with Mellanox HDR interconnect fabric running Slurm
• **62 nodes × AMD EPYC Rome 7702 / 2.00 GHz**; # of CPU Cores: 64; # of Threads: 128; Max Boost Clock: 3.35 GHz Base Clock: **2.00 GHz; L3** Cache 256 MB**;** Default TDP / TDP: 200W; Mellanox HDR-100 **200Gb/s**
• **145 nodes × AMD EPYC Milan 7713 / 2.00 GHz**; # of CPU Cores: 64; # of Threads: 128; Max Boost Clock: 3.35 GHz Base Clock: **2.25 GHz; L3** Cache 256 MB**;** Default TDP / TDP: 225W; Mellanox EDR **100Gb/s**
SPARTAN cluster at the Atos HPC, AI & QLM Benchmarking Centre – AMD EPYC Rome system with Mellanox ConnectX-6 HDR100 interconnect fabric
• **240 × AMD EPYC Rome 7742 / 2.25 GHz;** # of CPU Cores: 64; # of Threads: 128; Max Boost Clock: 3.35 GHz Base Clock**: 2.25 GHz; L3** Cache 256 MB; Default TDP / TDP: 225W; **Mellanox ConnectX-6 HDR 100 InfiniBand:** : Memory: 256GB DDR4 2677MHz RDIMMs per node: **DDN lustre 7990 Storage, NFS**
Atos Genji cluster at the Atos HPC, AI & QLM Benchmarking Centre with Mellanox HDR interconnect fabric running Slurm
• **16 nodes × AMD EPYC Milan 7713 / 2.00 GHz**; # of CPU Cores: 64; # of Threads: 128; Max Boost Clock: 3.35 GHz Base Clock: **2.25 GHz; L3** Cache 256 MB**;** Default TDP / TDP: 200W; Mellanox EDR **100Gb/s**

The Performance Benchmarks

- The ***Test suite*** comprises both **synthetics & end-user applications**. Synthetics limited to **IMB** benchmarks (*http://software.intel.com/en-us/articles/intel-mpi-benchmarks*) and **STREAM**
- Variety of “open source” & commercial end-user application codes:

DL_POLY, GROMACS & LAMMPS (molecular dynamics)

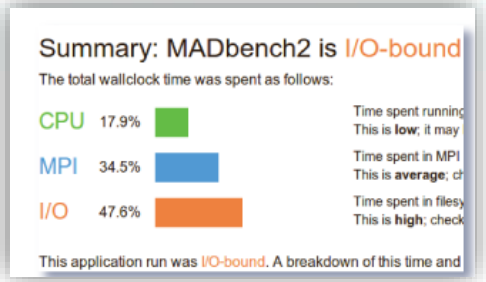
VASP (ab initio Materials properties)

GAMESS-UK (molecular electronic structure)

- These stress various aspects of the architectures under consideration and should provide a level of insight into why particular levels of performance are observed e.g., ***memory bandwidth and latency, node floating point performance and interconnect performance (both latency and B/W) and sustained I/O performance***.

Analysis Software - Allinea|ARM Performance Reports

Provides a mechanism to characterize and understand the performance of HPC application runs through a single-page HTML report.



- Based on Allinea MAP's adaptive sampling technology that keeps data volumes collected and **application overhead low**.
- Modest application slowdown (ca. 5%) even with 1000's of MPI processes.
- **Runs on existing codes: a single command added to execution scripts.**
- If submitted through a batch queuing system, then the submission script is modified to load the Allinea module and add the 'perf-report' command in front of the required mpirun command.

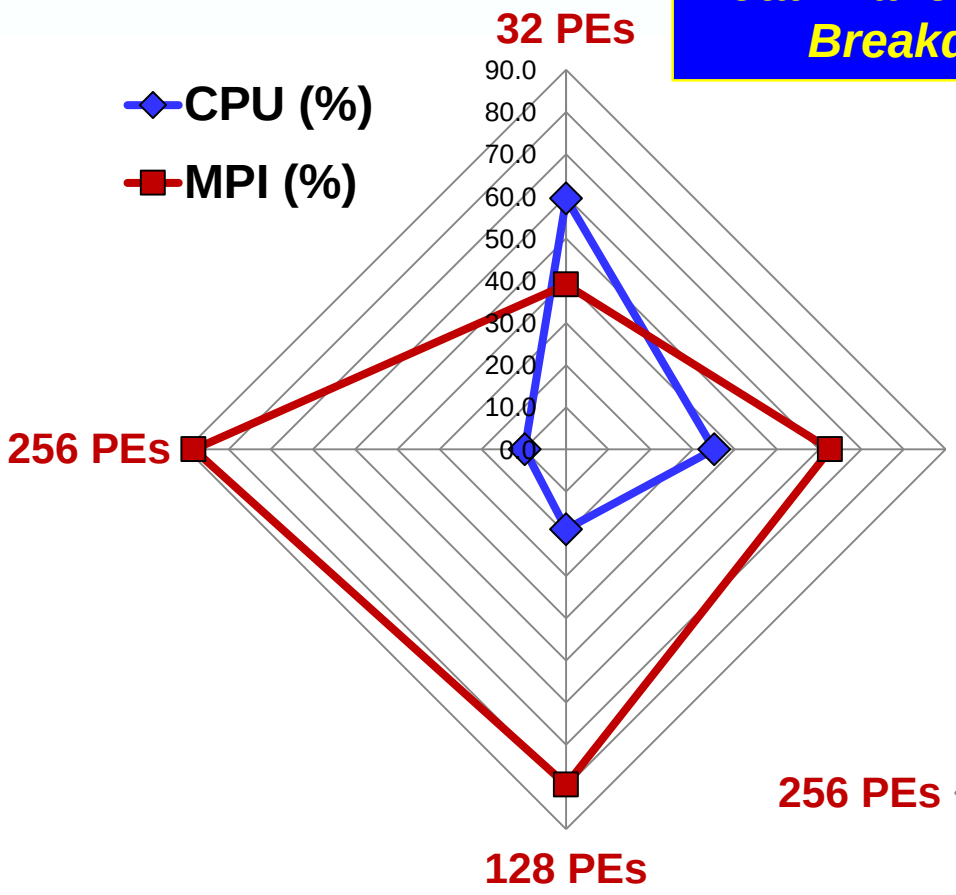
  perf-report mpirun $code

- ***A Report Summary:*** This characterizes how the application's wallclock time was spent, broken down into CPU, MPI and I/O
- All examples from the **Hawk Cluster (SKL Gold 6148 / 2.4GHz)**

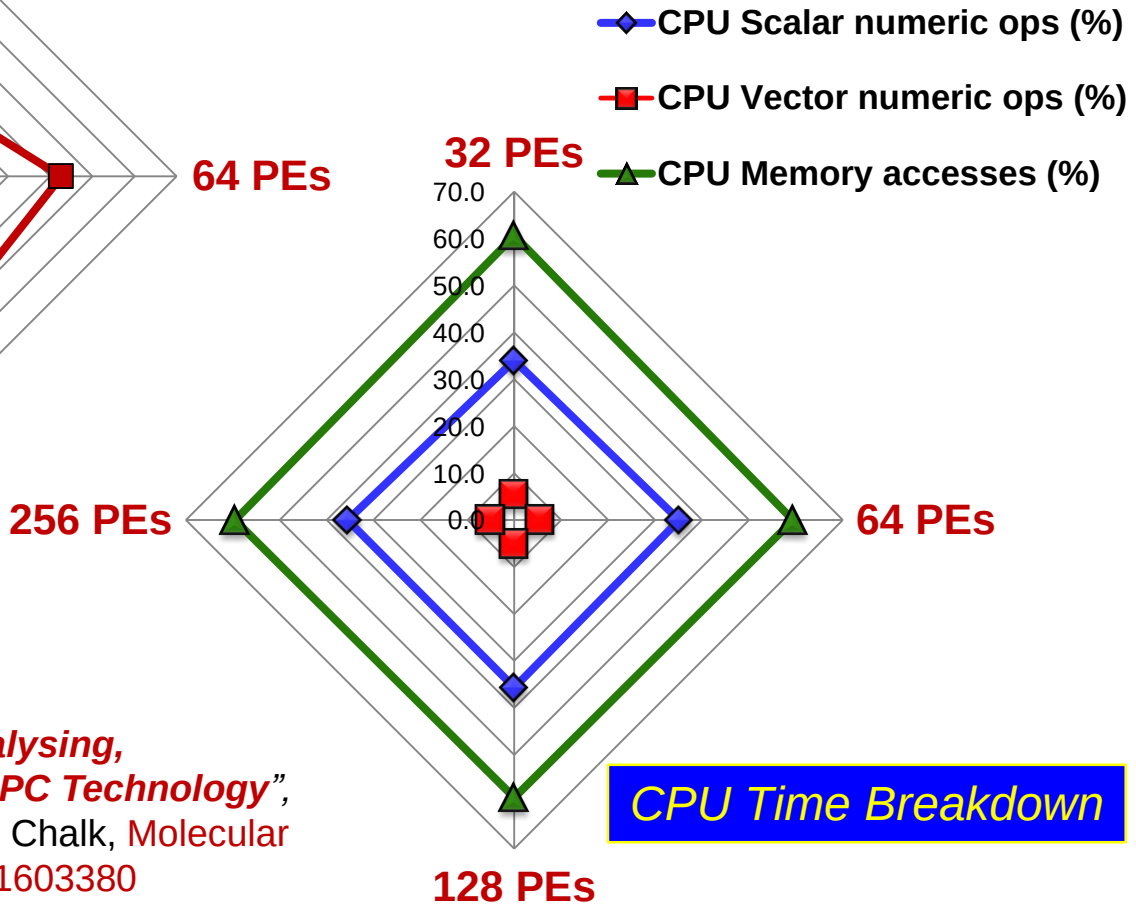
DLPOLY4 – Performance Report

Total Wallclock Time Breakdown

Performance Data (32-256 PEs)



Smooth Particle Mesh Ewald Scheme



"DL_POLY - A Performance Overview. Analysing, Understanding and Exploiting available HPC Technology", Martyn F Guest, Alin M Elena and Aidan B G Chalk, Molecular Simulation, (2019) 10.1080/08927022.2019.1603380

CPU Time Breakdown

EPYC - Compiler and Run-time Options

STREAM (AMD Minerva Cluster):

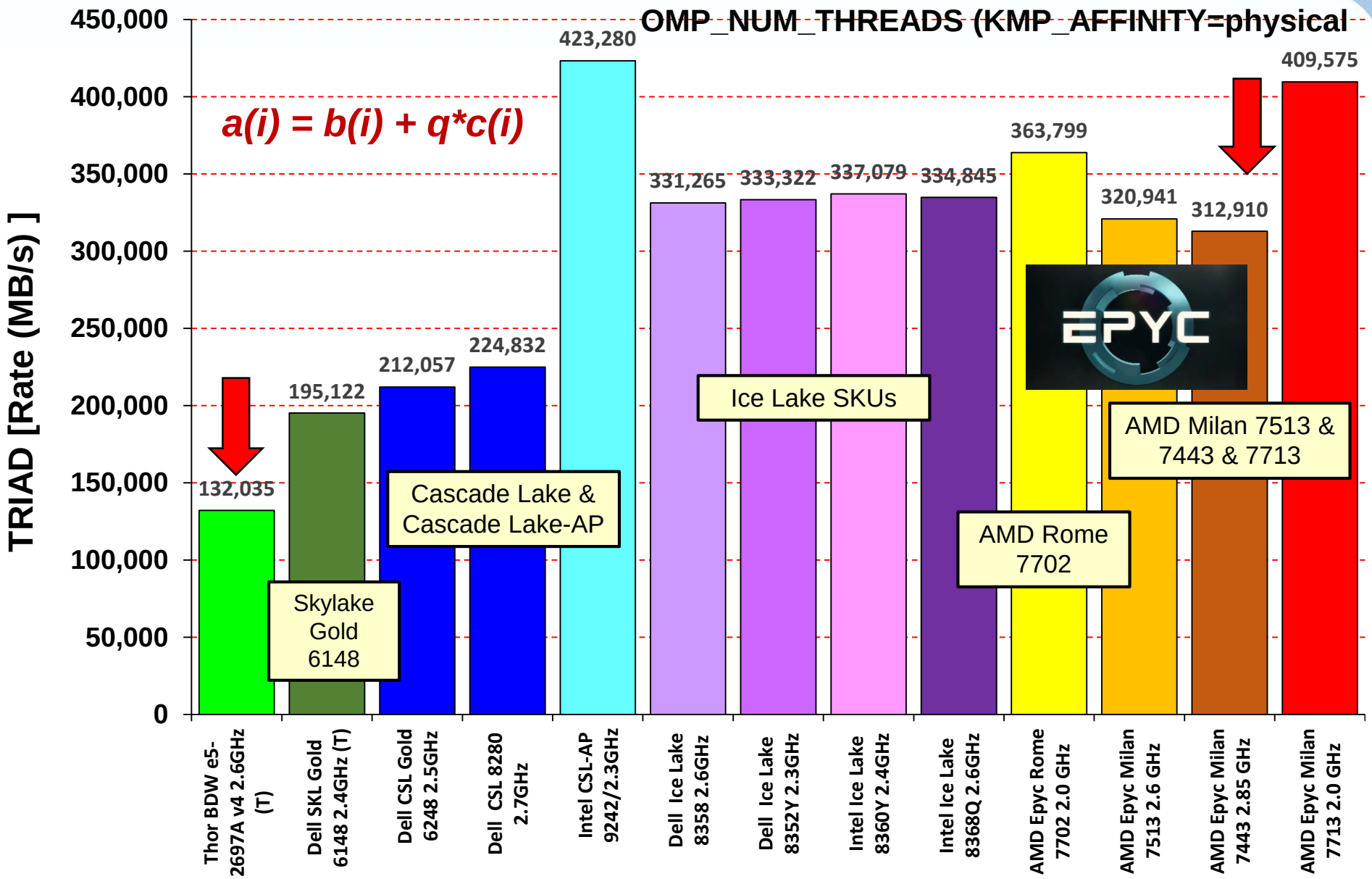
```
icc stream.c -DSTATIC -Ofast -march=core-avx2 -DSTREAM_ARRAY_SIZE=2500000000 -
DNTIMES=10 -mcmodel=large -shared-intel -restrict -qopt-streaming-stores always
-o streamc.Rome
icc stream.c -DSTATIC -Ofast -march=core-avx2 -qopenmp -
DSTREAM_ARRAY_SIZE=2500000000 -DNTIMES=10 -mcmodel=large -shared-intel -restrict
-qopt-streaming-stores always -o streamcp.Rome

# Version of Intel compiler to use and way to source it
source /opt/intel/compilers_and_libraries_2020.2.254/linux/bin/compilervars.sh -
ofi_internal=1 intel64

# Increasing use of oneAPI: e.g., source /opt/intel/oneapi/setvars.sh

# Use this latest version of Intel MKL,  further versions do not allow the
setting of AVX2 on non-Intel processors.
source /opt/intel/compilers_and_libraries_2019.6.324/linux/mkl/bin/mklvars.sh
intel64

# When using IntelMPI on AMD Rome/Milan
export I_MPI_FABRICS=shm:ofi
export I_MPI_SHM=clx_avx2
export FI_PROVIDER=mlx

# ON AMD Rome/Milan when using Intel MKL
export MKL_DEBUG_CPU_TYPE=5
```

Compilation:

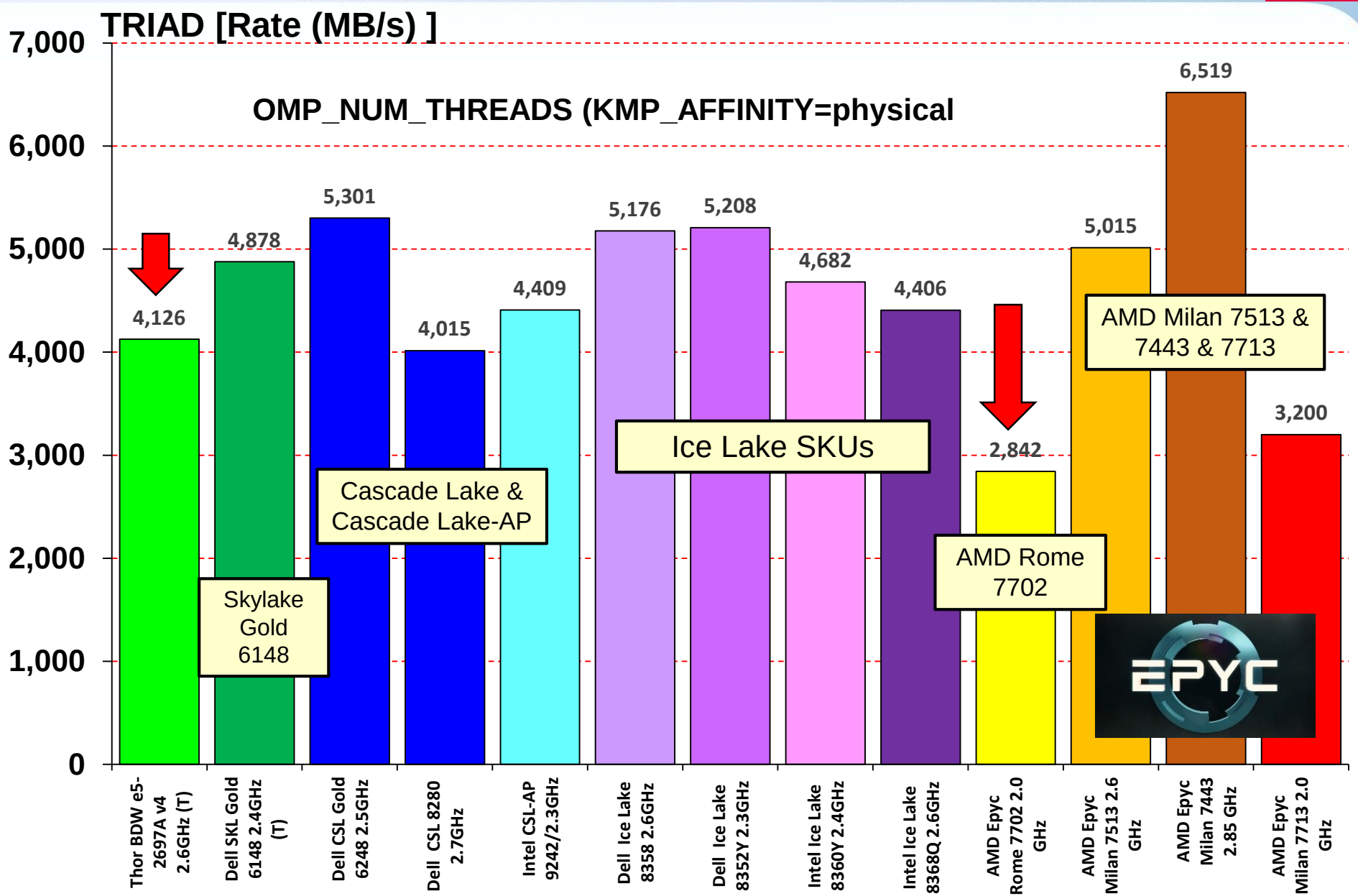
INTEL SKL: –O3 –xCORE-AVX512

AMD EPYC: –O3 –march=core-avx2 -align array64byte -fma -ftz -fomit-frame-pointer

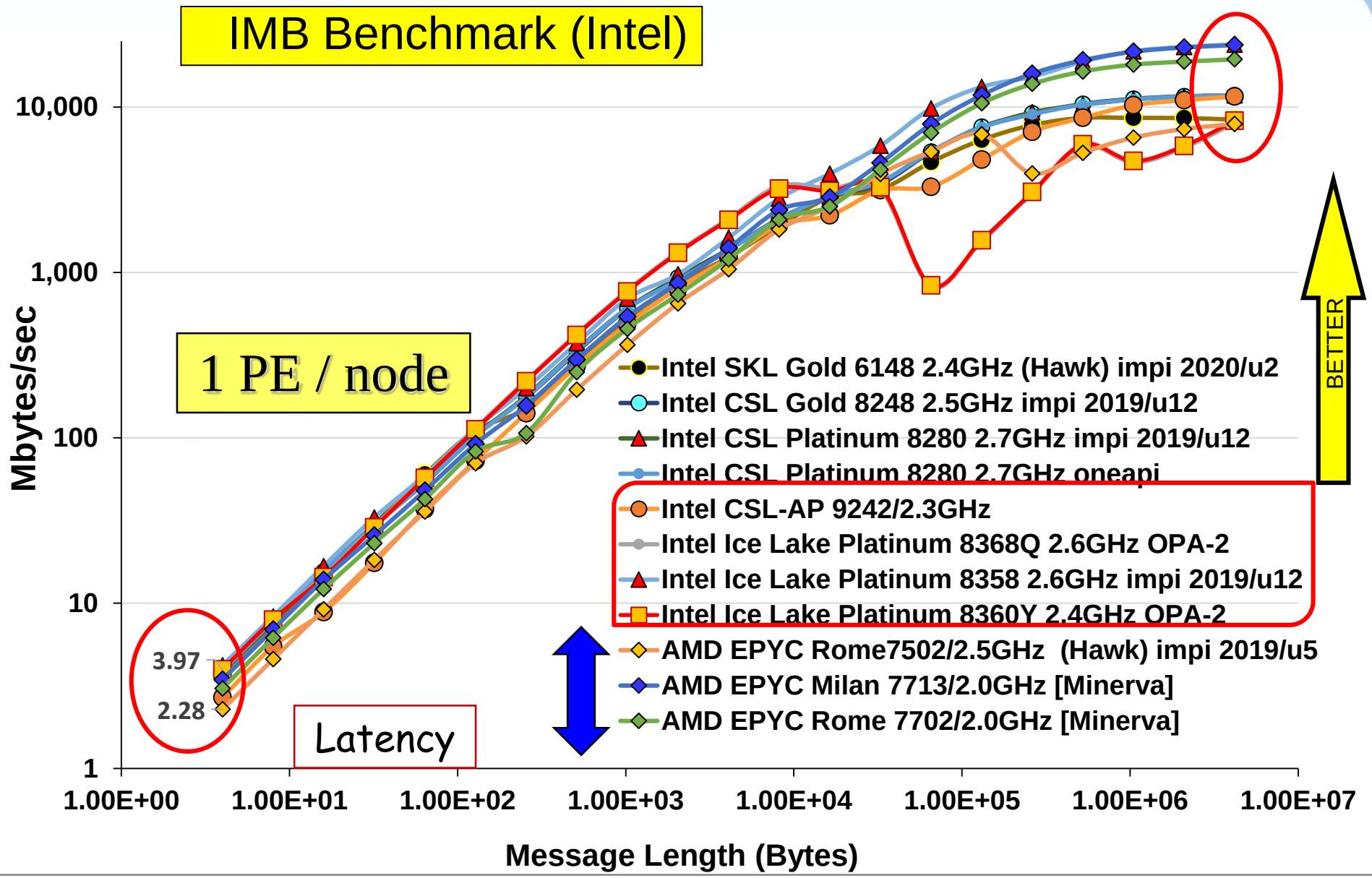
Memory B/W – STREAM performance



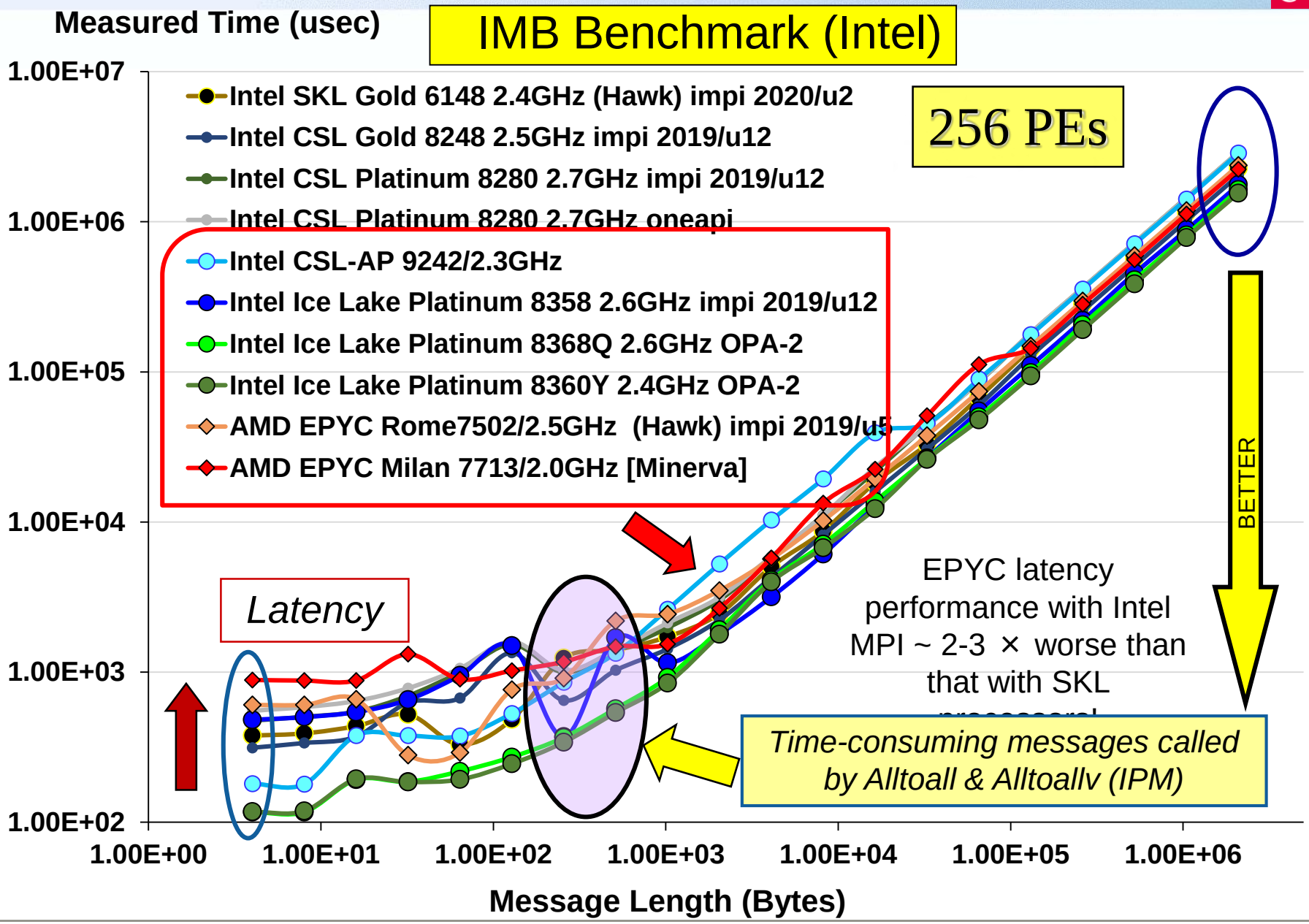
Memory B/W – STREAM / core performance



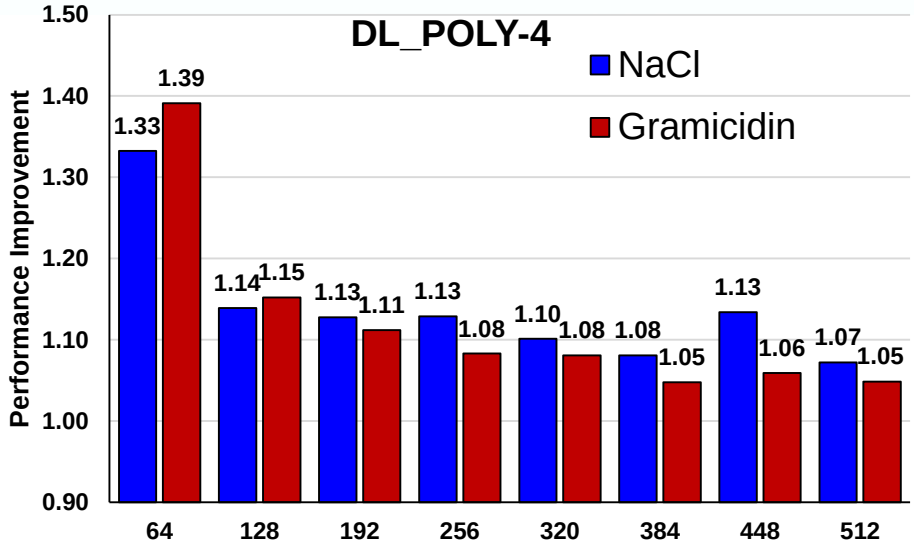
MPI Performance – PingPong



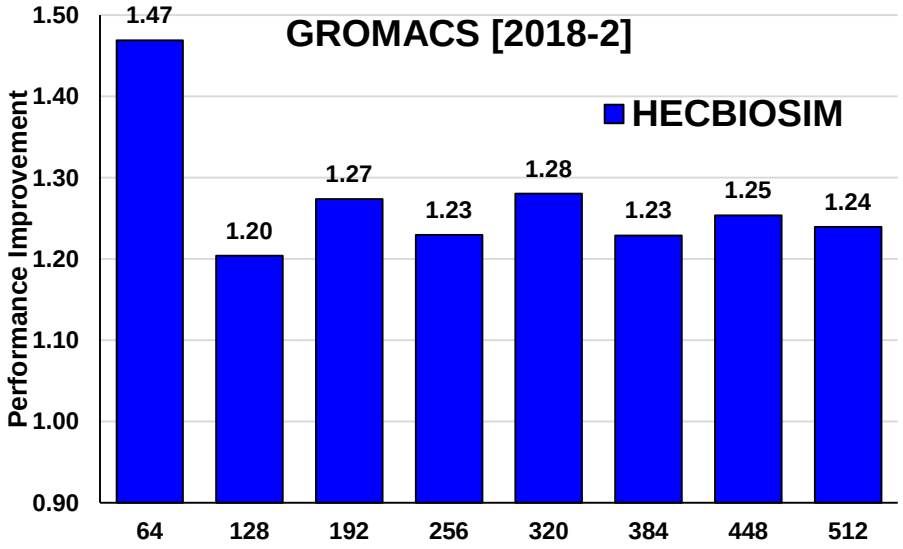
MPI Collectives – Alltoallv (256 PEs)

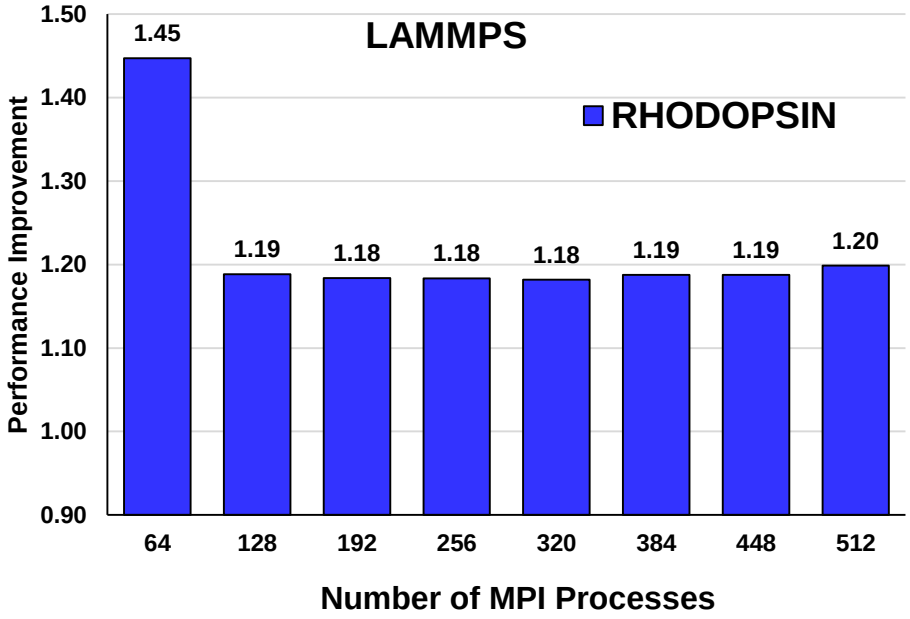


Performance Impact of Turbo Mode



SPARTAN Cluster – AMD
EPYC Rome 7742 / 2.3 GHz





Performance Metrics – “Core to Core” & “Node to Node”

- Analysis of performance Metrics across a variety of data sets
  - ❑ “**Core to core**” and “**node to node**” workload comparisons
    - ***Core to core*** comparison i.e. performance for jobs with a fixed number of cores
    - ***Node to Node*** comparison typical of the performance when running a workload (real life production). Expected to reveal the major benefits of increasing core count per socket
  - ❑ Focus on a variety of “**node to node**” and “**core-to-core**” comparisons e.g., :

1	***Hawk - Dell |EMC Skylake Gold 6148 2.4GHz (T) EDR with 40 cores / node***	***AMD EPYC Milan 7713 nodes with128 cores per node. [1-7 nodes]***
2	***Hawk - Dell |EMC Skylake Gold 6148 2.4GHz (T) EDR with 40 cores / node***	***Intel Xeon Platinum Ice Lake 8358 nodes with 64 cores per node. [1-7 nodes]***

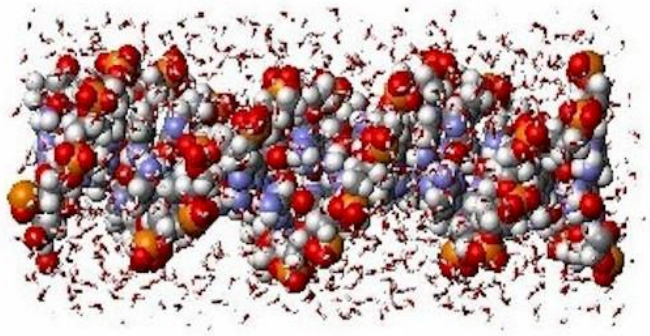
Performance of Computational Chemistry Codes

Molecular Simulation;

1. DL_POLY

Molecular Simulation I. DL_POLY

Molecular Dynamics Codes: AMBER, DL_POLY, CHARMM, NAMD, LAMMPS, GROMACS etc



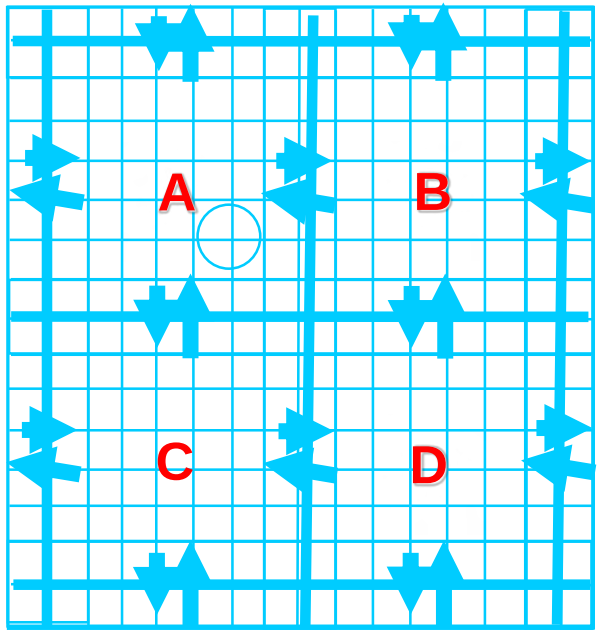
DL_POLY

- Developed as CCP5 parallel MD code by W. Smith, T.R. Forester and I. Todorov
  - UK CCP5 + International user community
  - DLPOLY_classic (replicated data) and DLPOLY_3 & _4 (distributed data – domain decomposition)
- Areas of application:
  - liquids, solutions, spectroscopy, ionic solids, molecular crystals, polymers, glasses, membranes, proteins, metals, solid and liquid interfaces, catalysis, clathrates, liquid crystals, biopolymers, polymer electrolytes.

DL_POLY 4 – Distributed data

Domain Decomposition - Distributed data:

- Distribute atoms, forces across the nodes
  - More memory efficient, can address much larger cases (10^5-10^7)
- Shake and short-ranges forces require only neighbour communication
  - communications scale linearly with number of nodes
- Coulombic energy remains global
  - Adopt **Smooth Particle Mesh Ewald** scheme
    - includes Fourier transform smoothed charge density (reciprocal space grid typically 64x64x64 - 128x128x128)



W. Smith and I. Todorov

Benchmarks

1. NaCl Simulation; 216,000 ions, 200 time steps, Cutoff=12Å

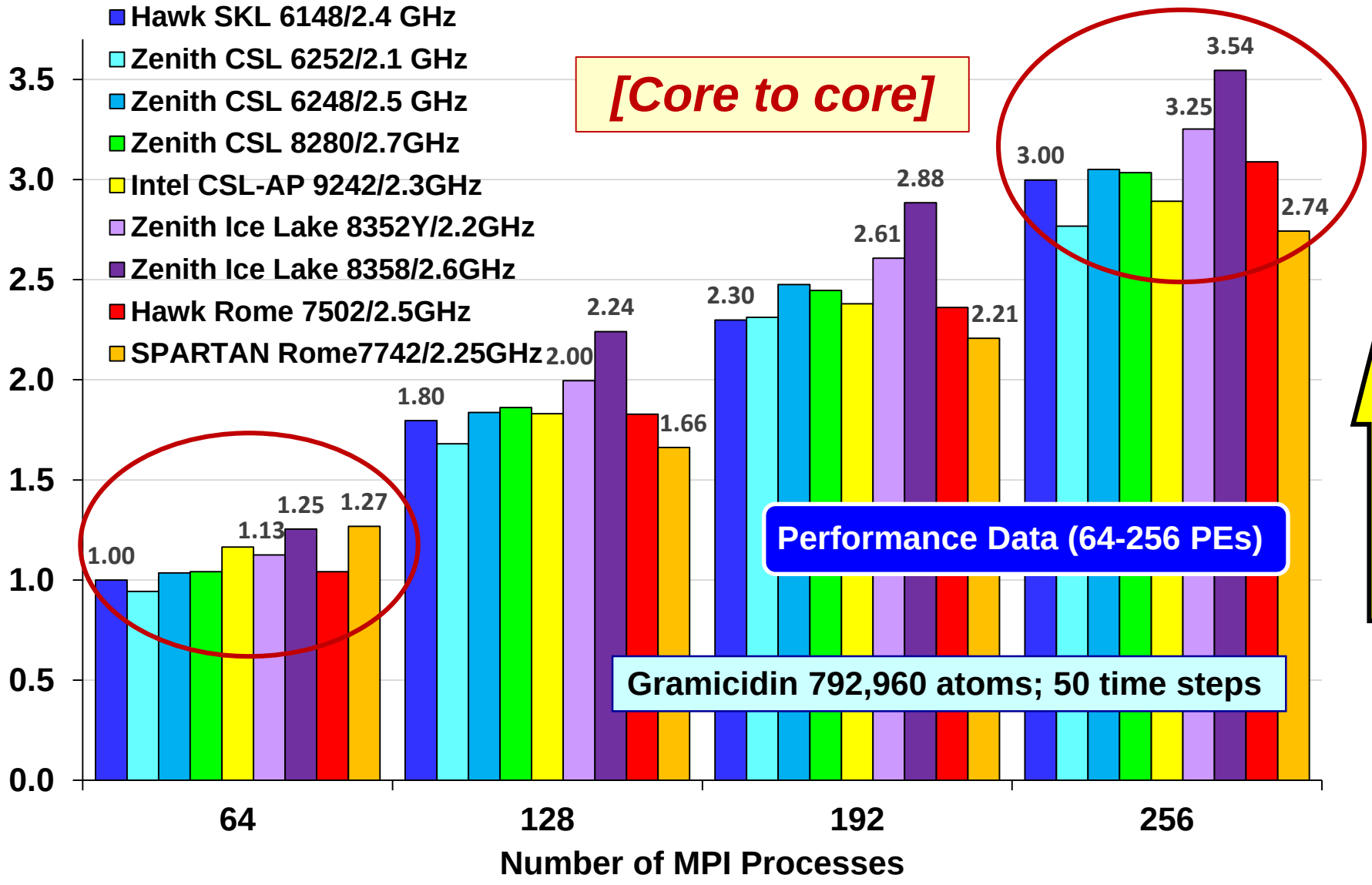
2. Gramicidin in water; rigid bonds + SHAKE: 792,960 ions, 50 time steps

http://www.scd.stfc.ac.uk//research/app/ccg/software/DL_POLY/44516.aspx

DL_POLY 4 – Gramicidin Simulation

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

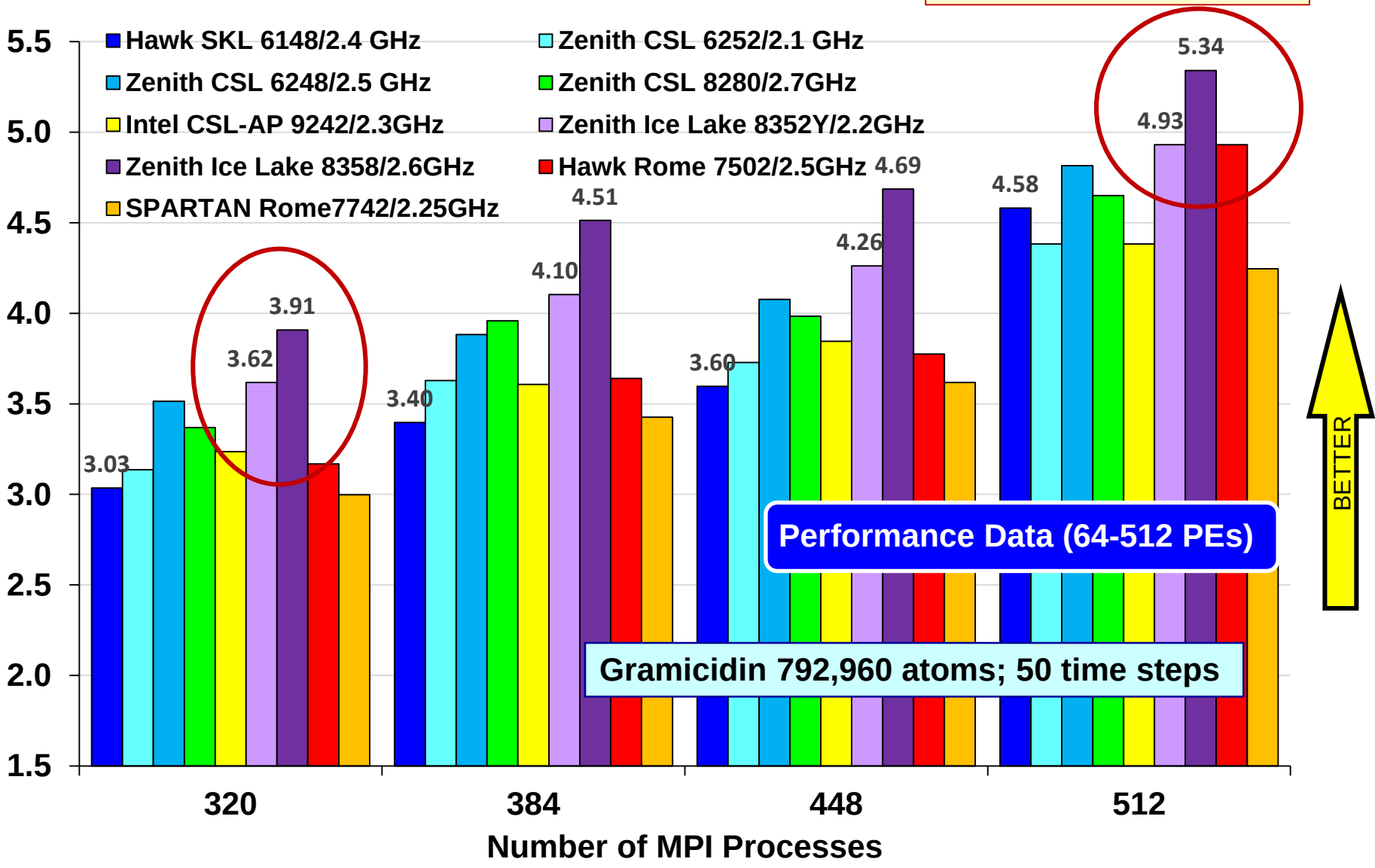
Performance *Relative to the Hawk SKL 6148 2.4 GHz (64 PEs)*



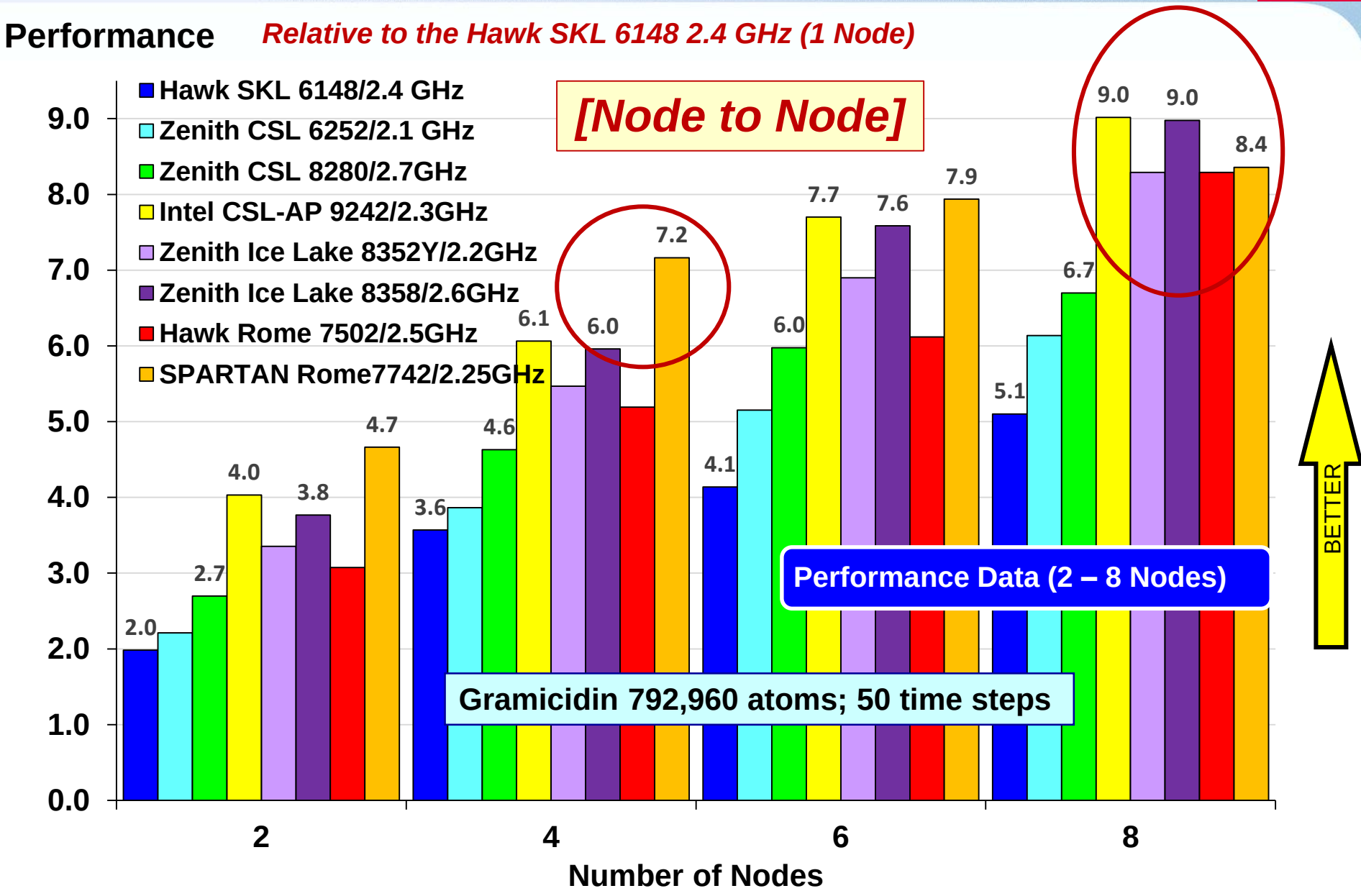
DL_POLY 4 – Gramicidin Simulation

Performance *Relative to the Hawk SKL 6148 2.4 GHz (64 PEs)*

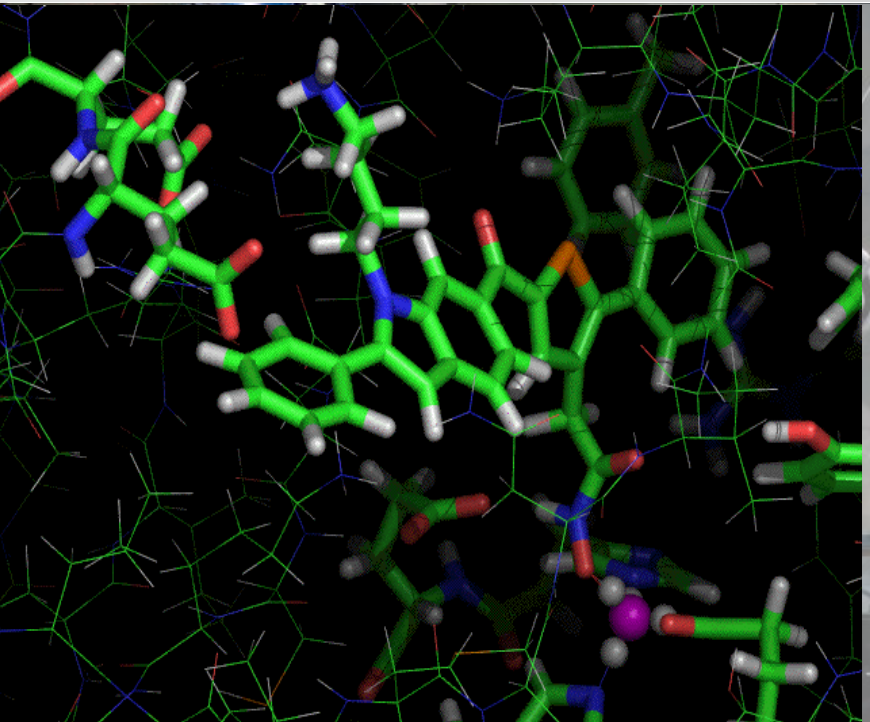
[Core to core]



DL_POLY 4 – Gramicidin Simulation



Performance of Computational Chemistry Codes



Molecular Simulation: 2. Gromacs

GROMACS (GROningen MAchine for Chemical Simulations) is a molecular dynamics package designed for simulations of proteins, lipids and nucleic acids [University of Groningen] .

Versions under Test:

Version 4.6.1 – 5 March 2013

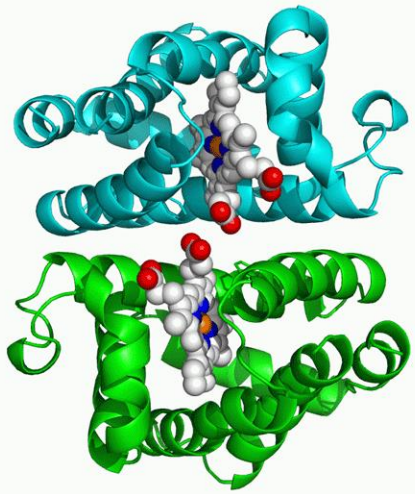
Version 5.0.7 – 14 October 2015

Version 2016.3 – 14 March 2017

Version 2018.2 – 14 June 2018

Version 2019.6 – 28 February 2020

Version 2020.1 – 3 March 2020



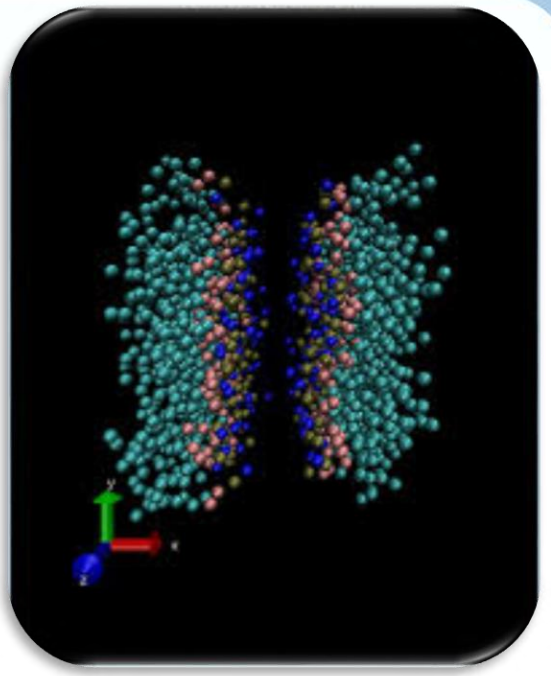
- **Berk Hess et al. "*GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation*". *Journal of Chemical Theory and Computation* 4 (3): 435–447.**

http://manual.gromacs.org/documentation/

GROMACS Benchmark Cases

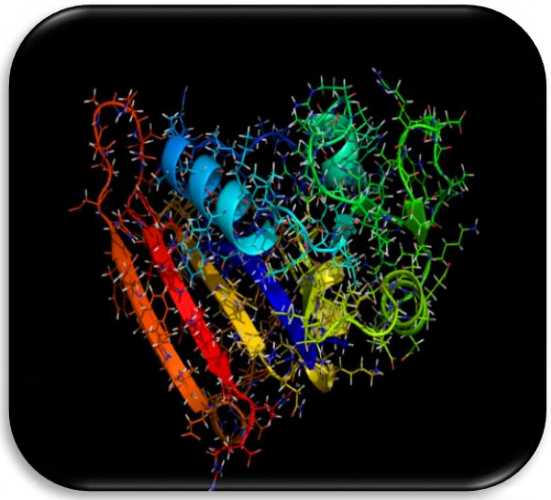
Ion channel system

- The 142k particle ion channel system is the membrane protein GluCl - a pentameric chloride channel embedded in a DOPC membrane and solvated in TIP3P water, using the Amber ff99SB-ILDN force field. This system is a **challenging** parallelization case due to the small size, but is one of the **wanted target sizes** for biomolecular simulations



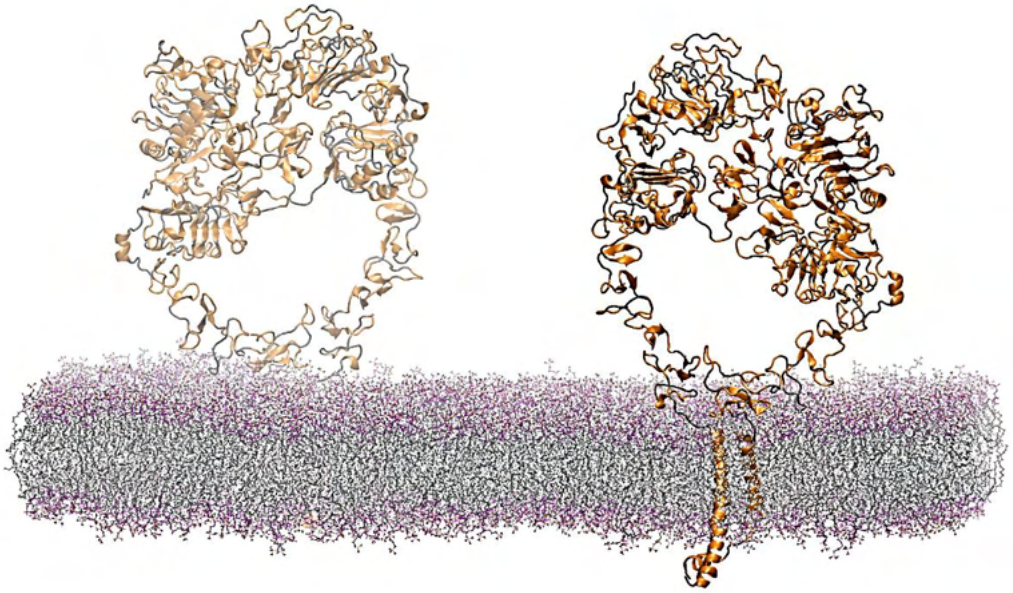
Lignocellulose

- Gromacs Test Case B from the UEA Benchmark Suite. A model of cellulose and lignocellulosic biomass in an aqueous solution. This system of 3.3M atoms is inhomogeneous, and uses **reaction-field electrostatics** instead of PME and therefore should scale well.



GROMACS Benchmark Case III

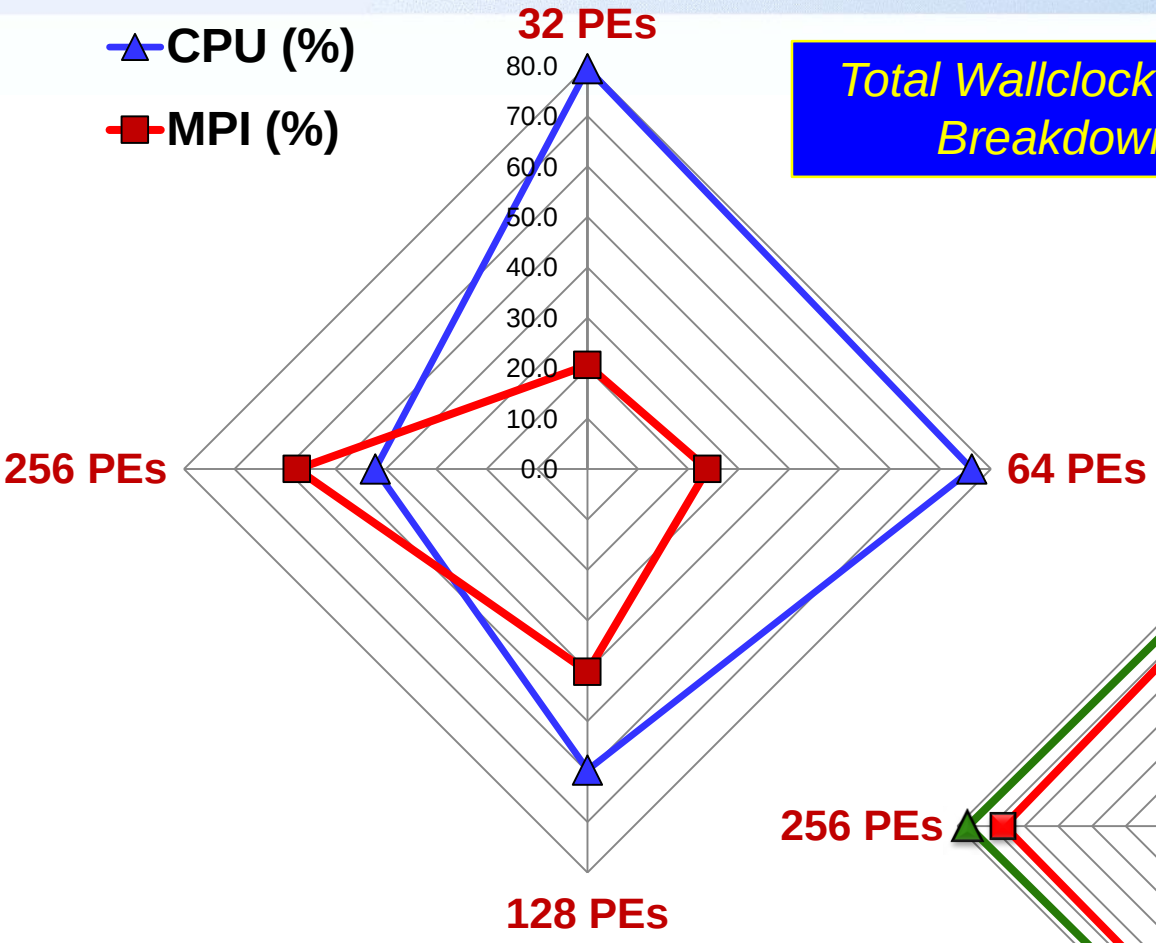
The HECBioSim Benchmarks



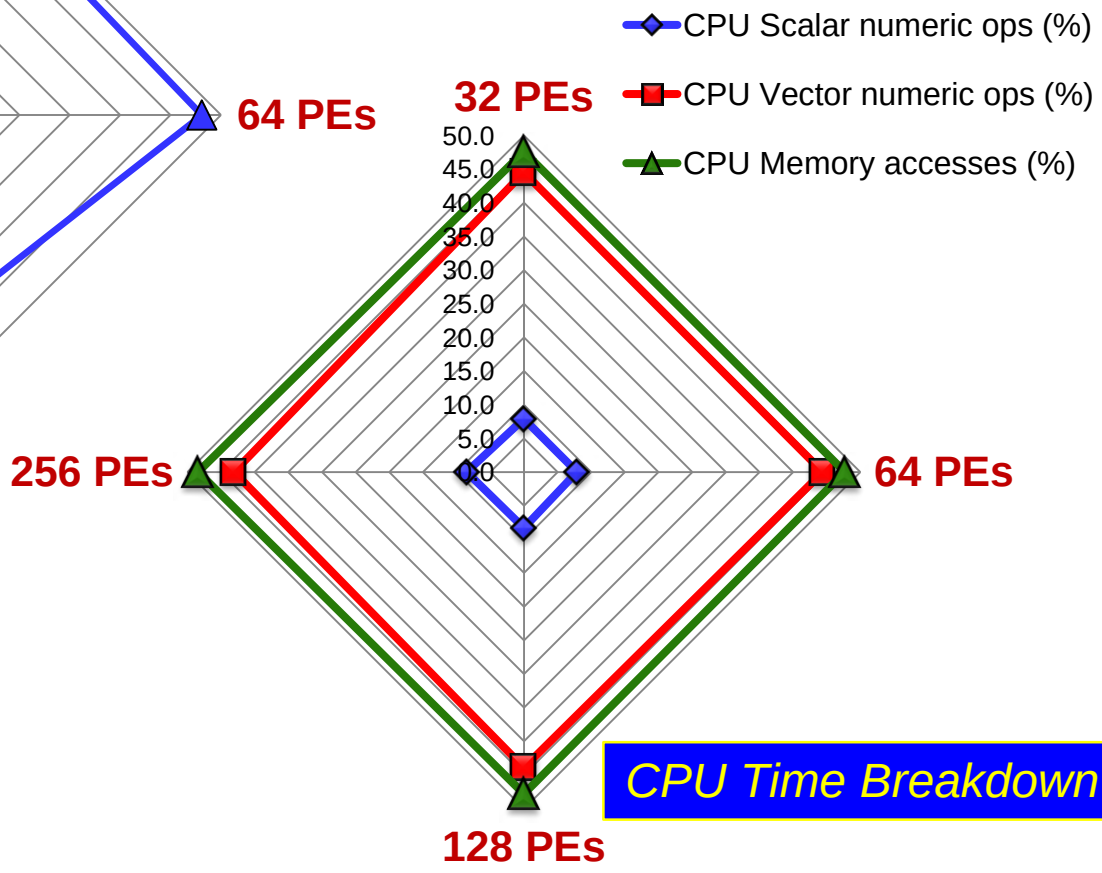
- PME simulation for 1.4M atom system - A Pair of Human Epidermal Growth Factor Receptor (hEGFR) Dimers of 1IVO and 1NQL
- Total number of atoms = **1,403,182**
- Protein atoms = 43,498 Lipid atoms = 235,304 Water atoms = 1,123,392 Ions = 986

https://www.hecbiosim.ac.uk/benchmarks

GROMACS – Ion-channel Performance Report



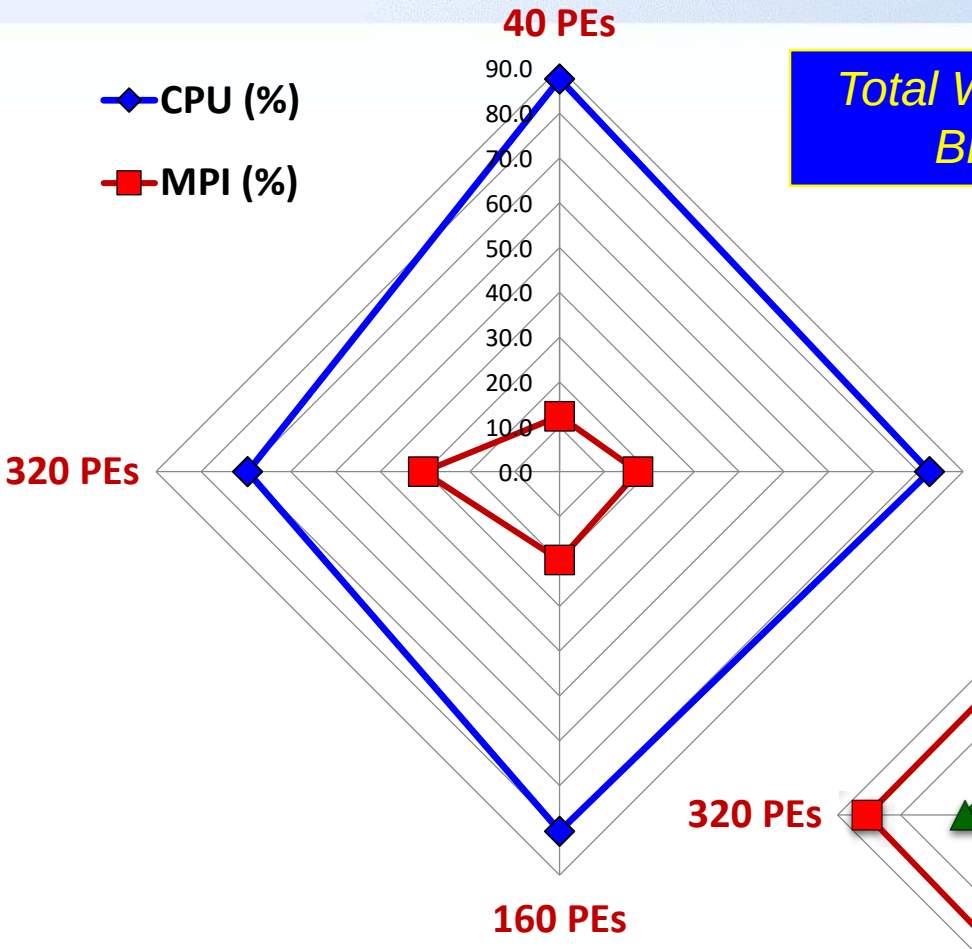
Total Wallclock Time Breakdown



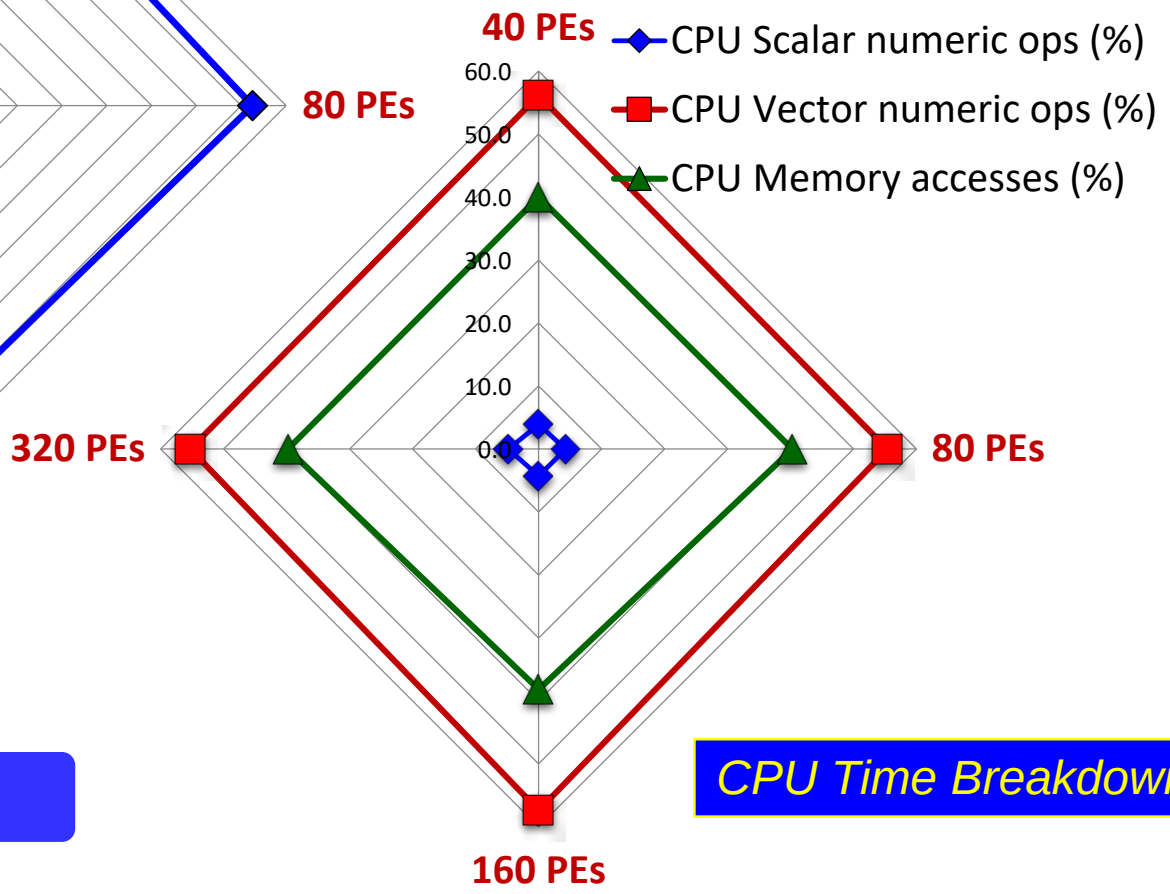
CPU Time Breakdown

Performance Data (32-256 PEs)

GROMACS – Ligocellulose Performance Report



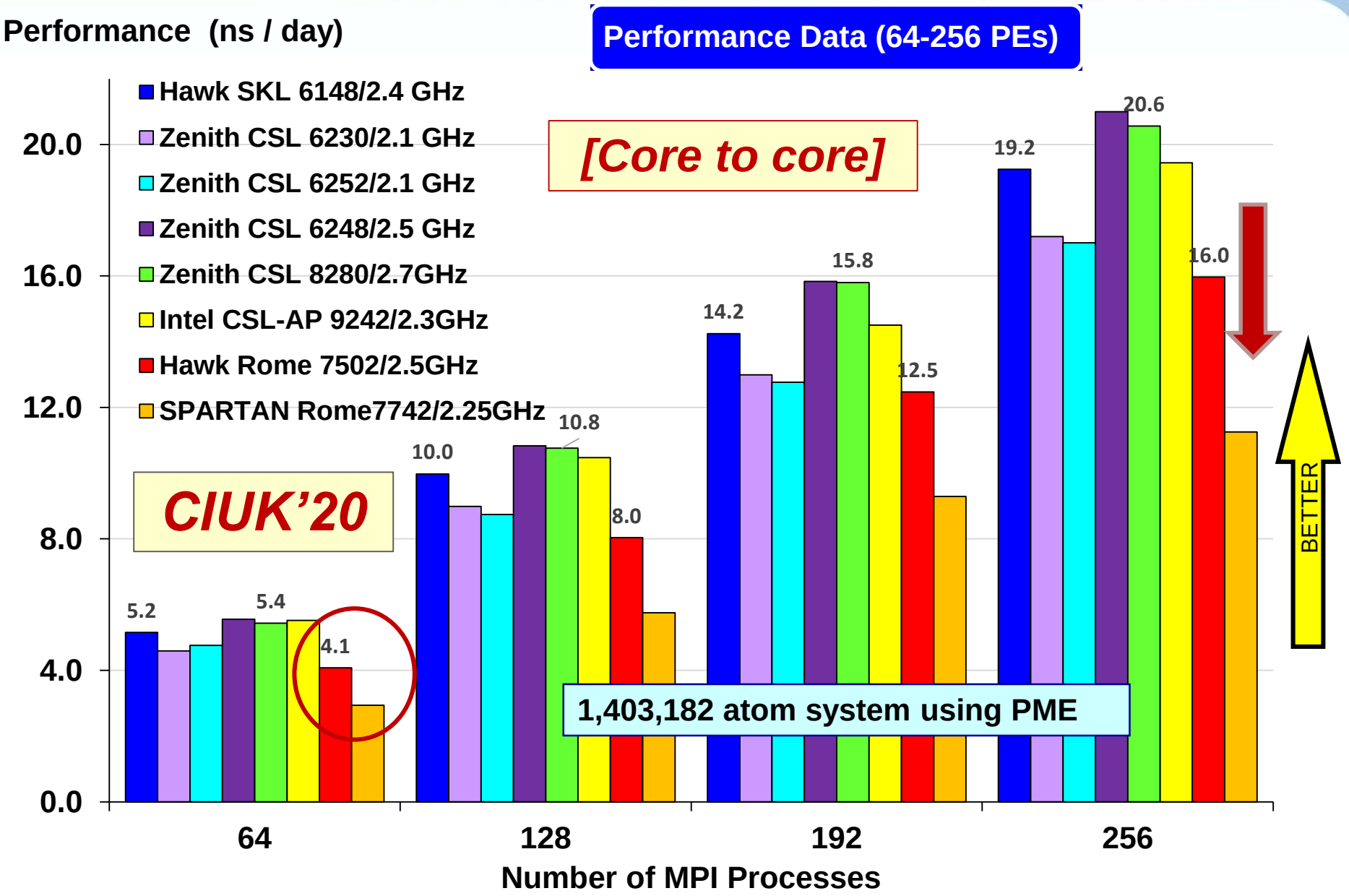
Total Wallclock Time Breakdown



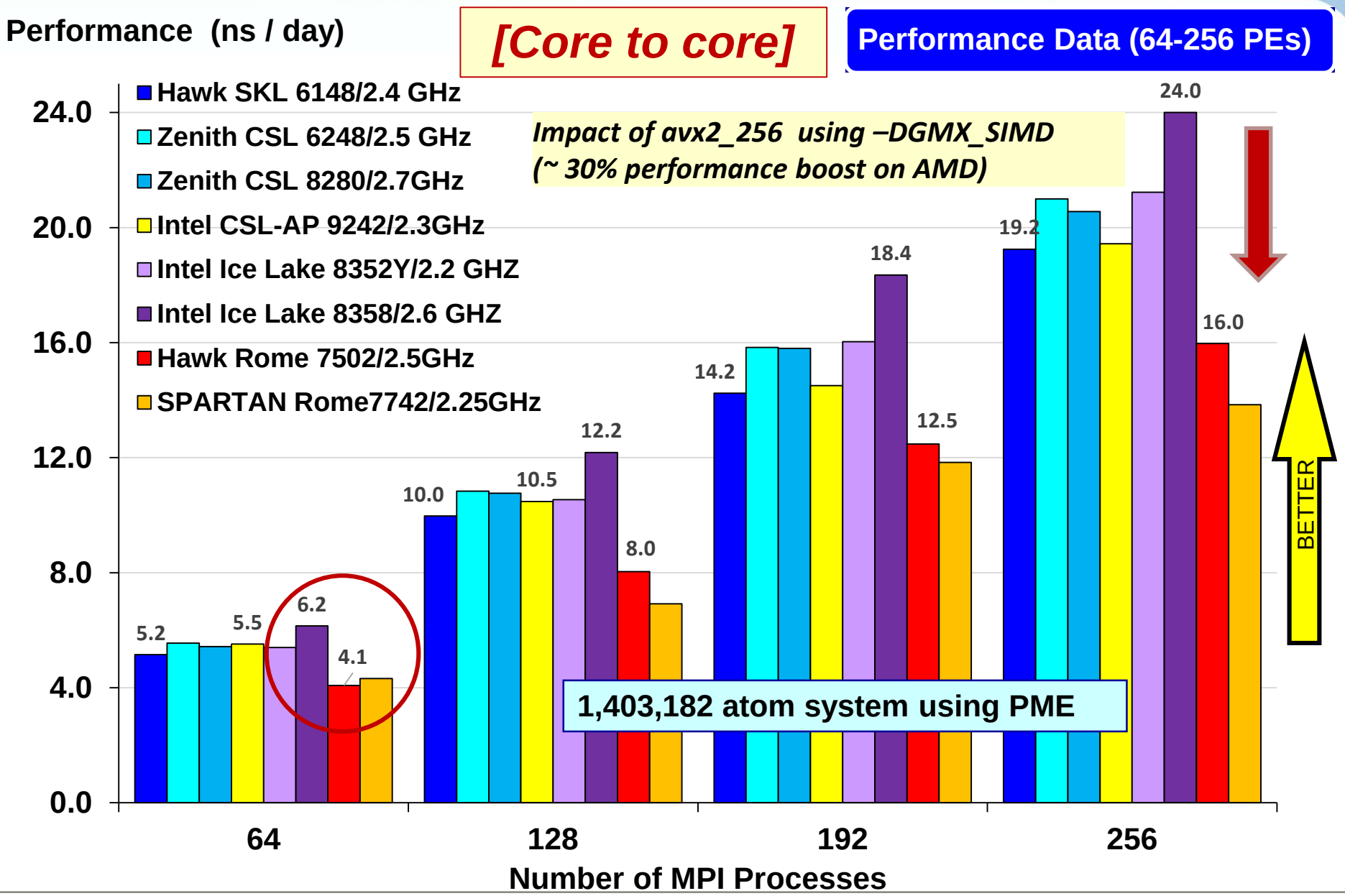
Performance Data (32-256 PEs)

CPU Time Breakdown

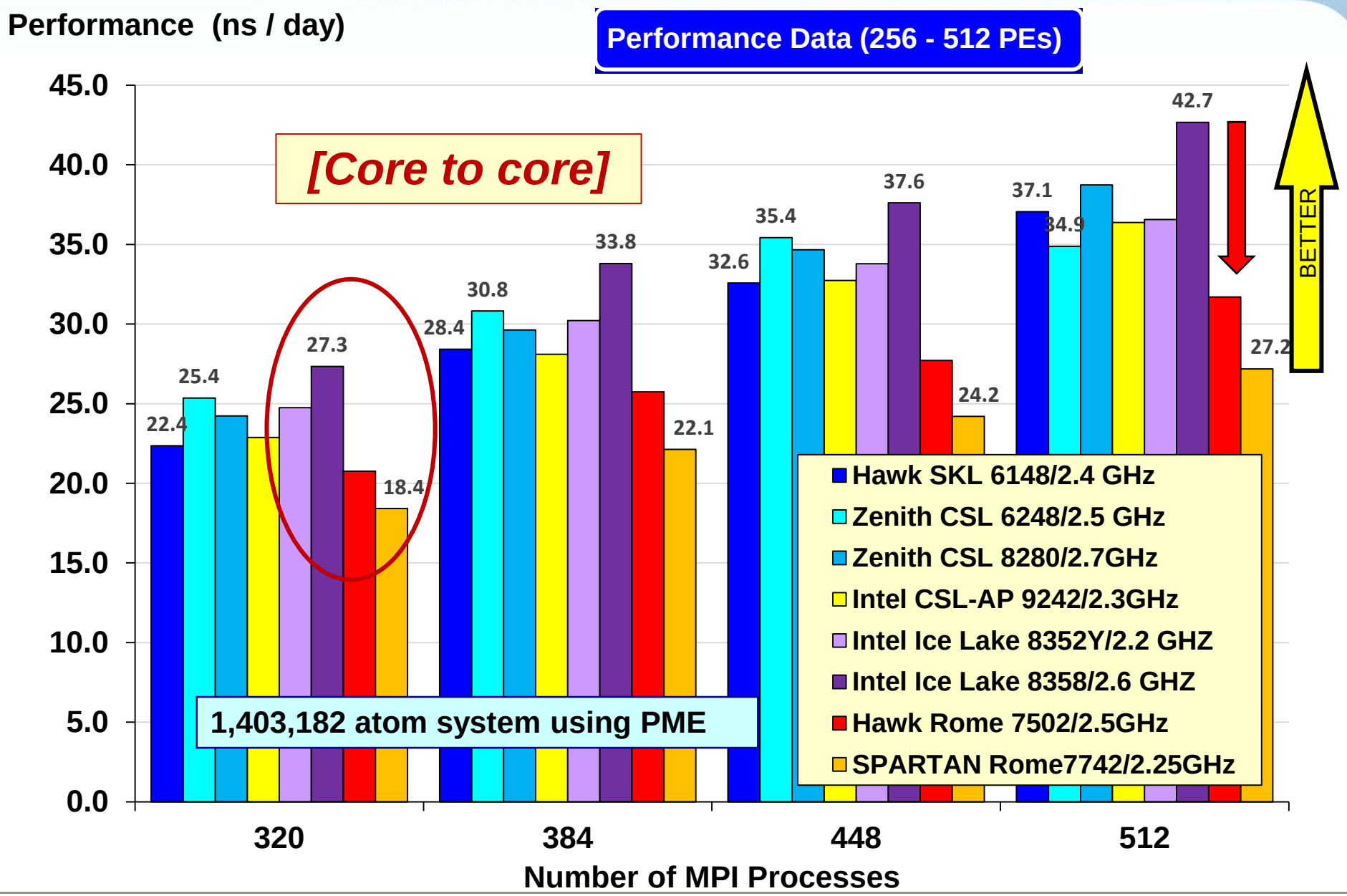
GROMACS – HECBioSim 1.4M Atom System



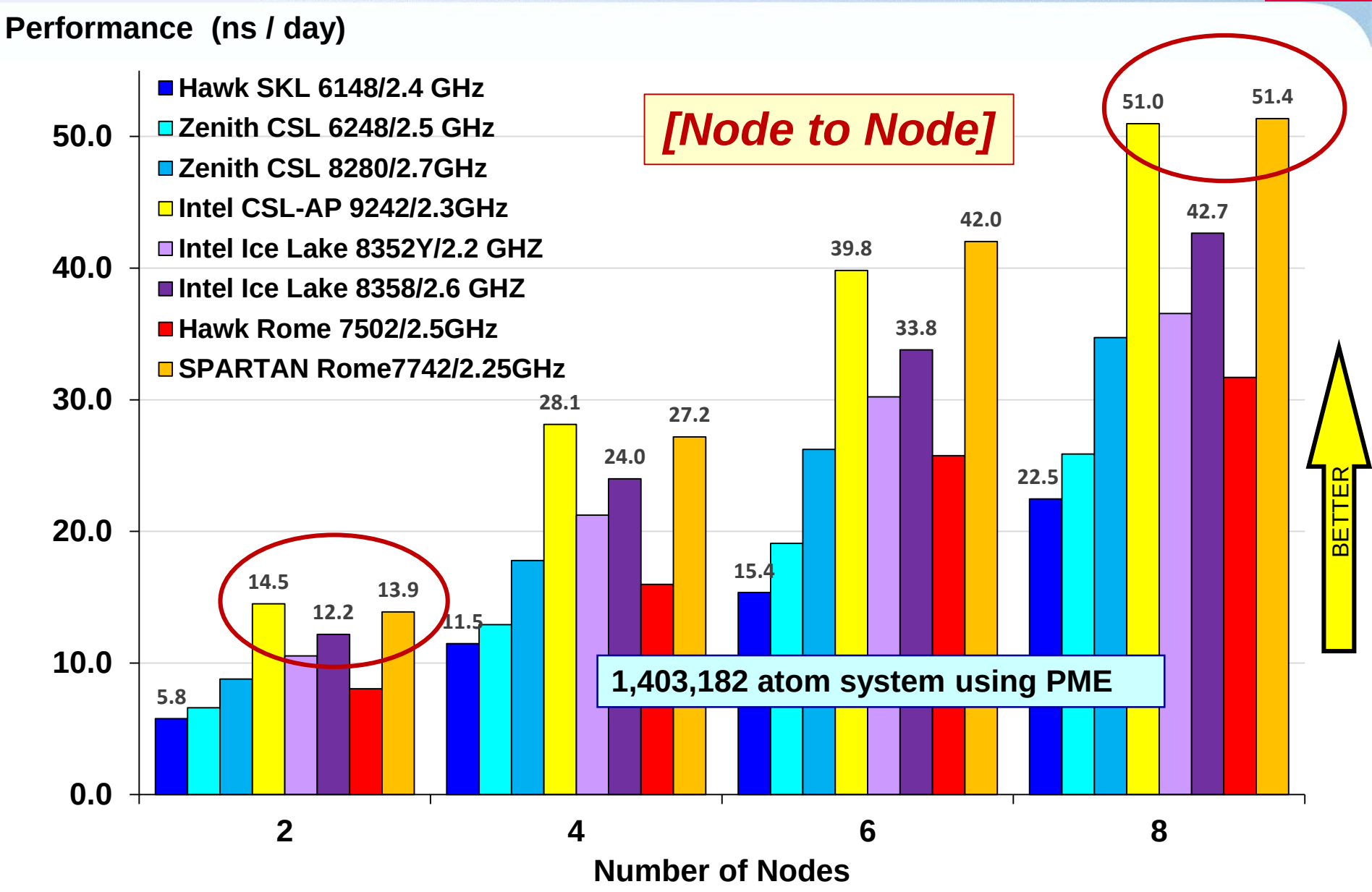
GROMACS – HECBioSim 1.4M Atom System



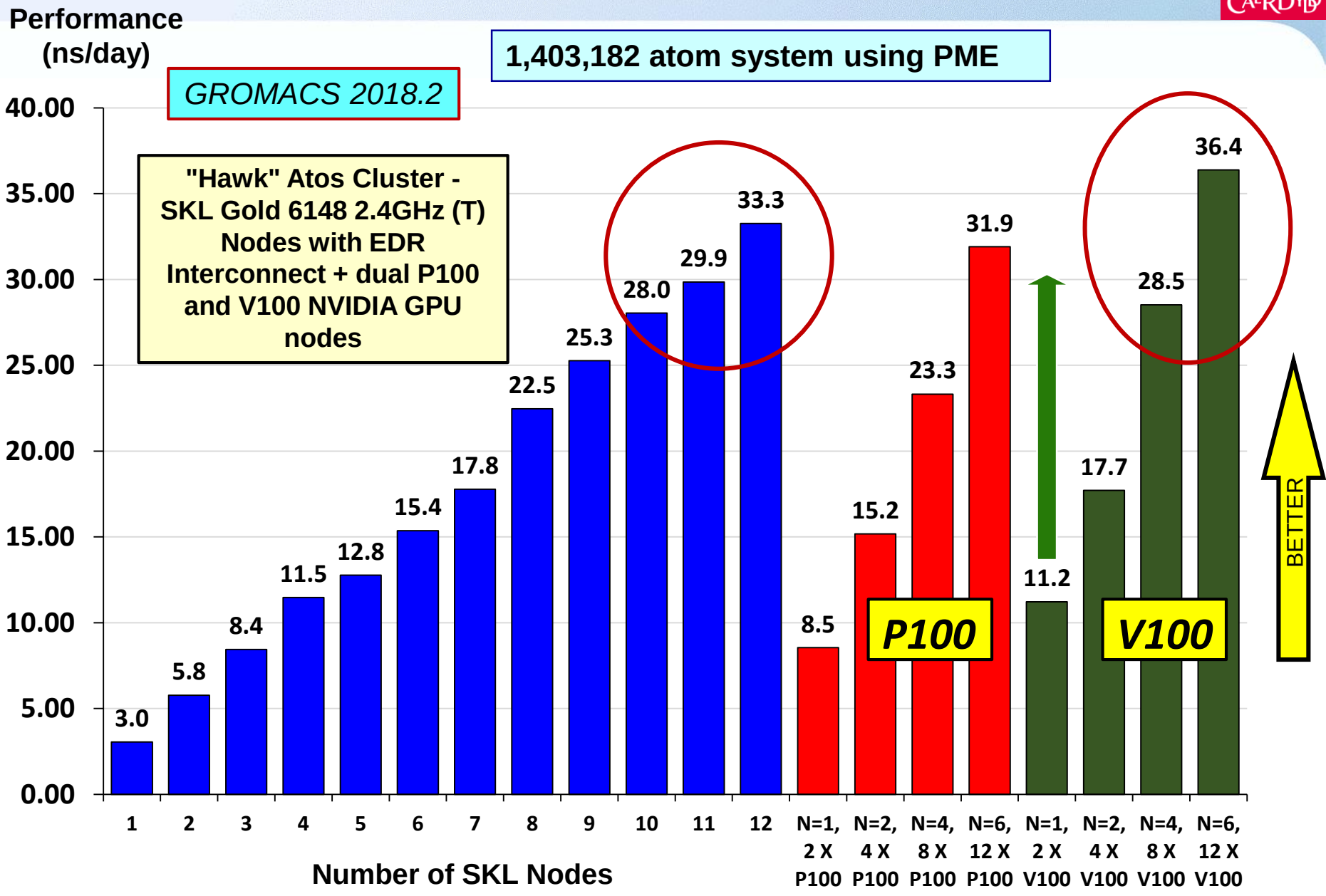
GROMACS – HECBioSim 1.4M Atom System



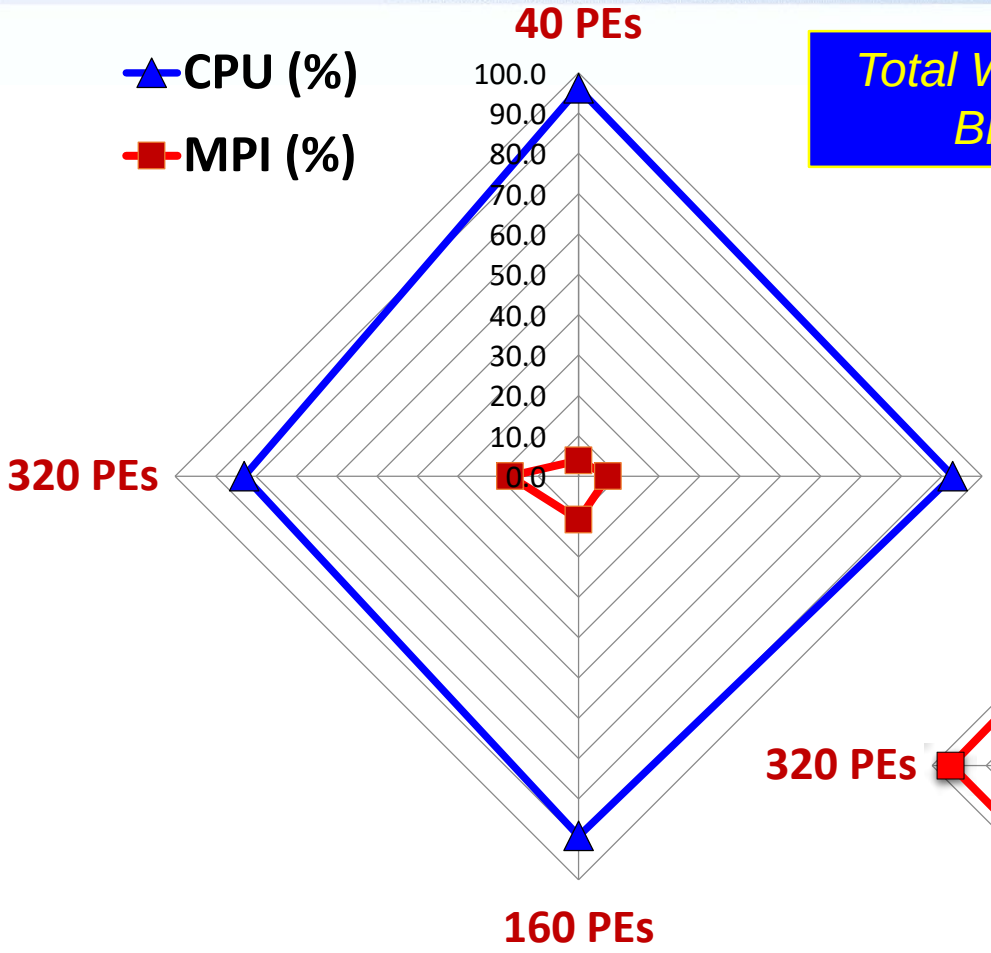
GROMACS – HECBioSim 1.4M Atom System



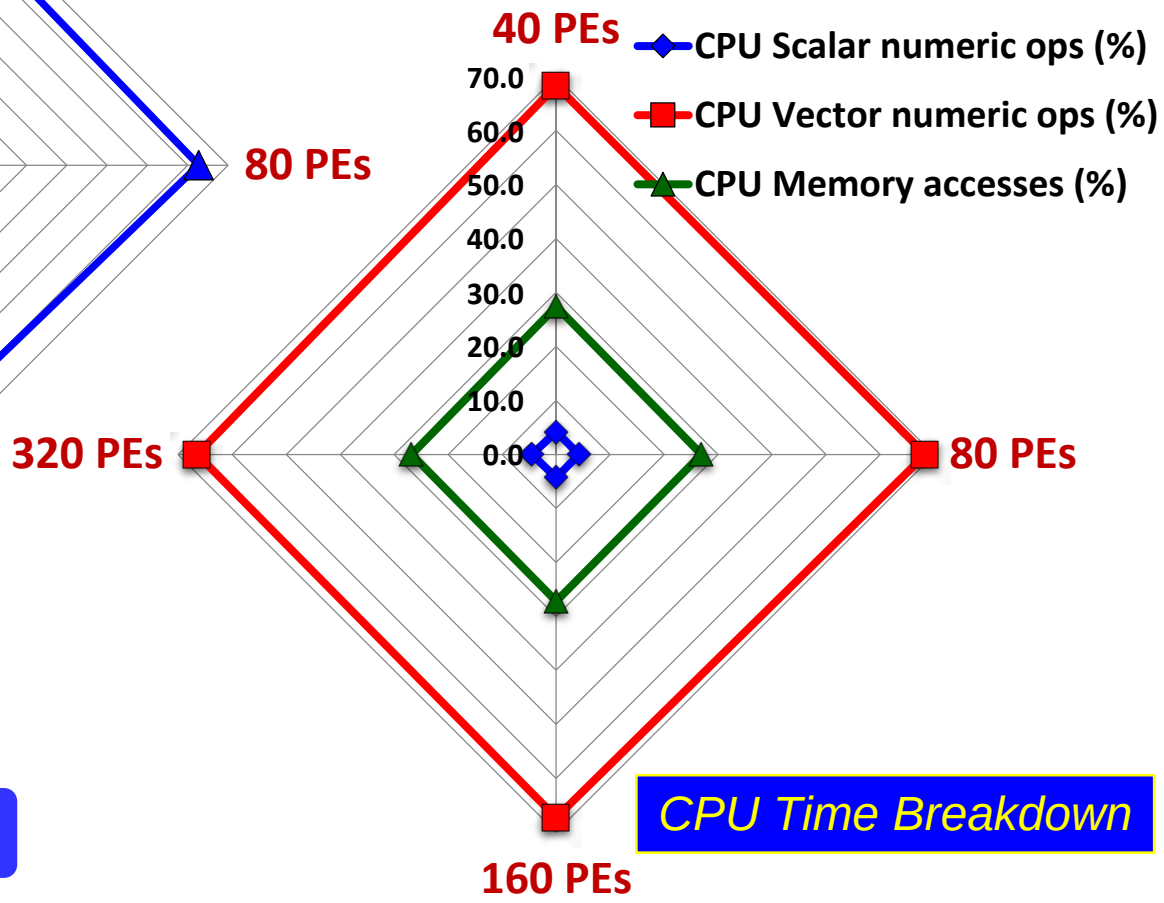
GROMACS – GPU Performance: HECBioSim Simulation



GROMACS – HECBioSim Performance Report



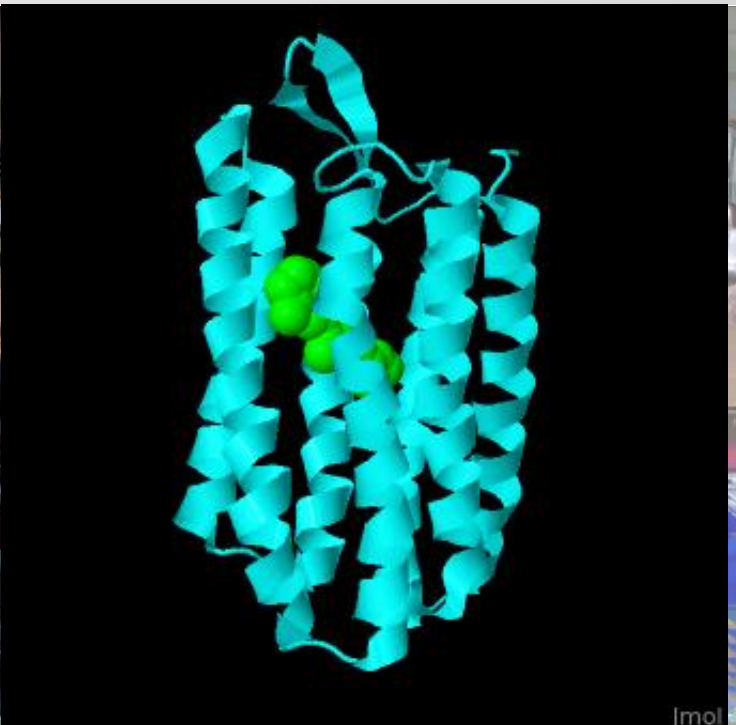
Total Wallclock Time Breakdown



CPU Time Breakdown

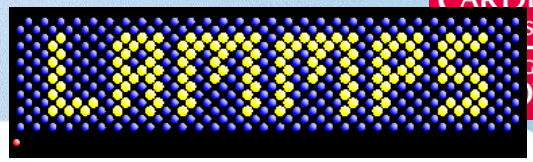
Performance Data (32-256 PEs)

Performance of Computational Chemistry Codes



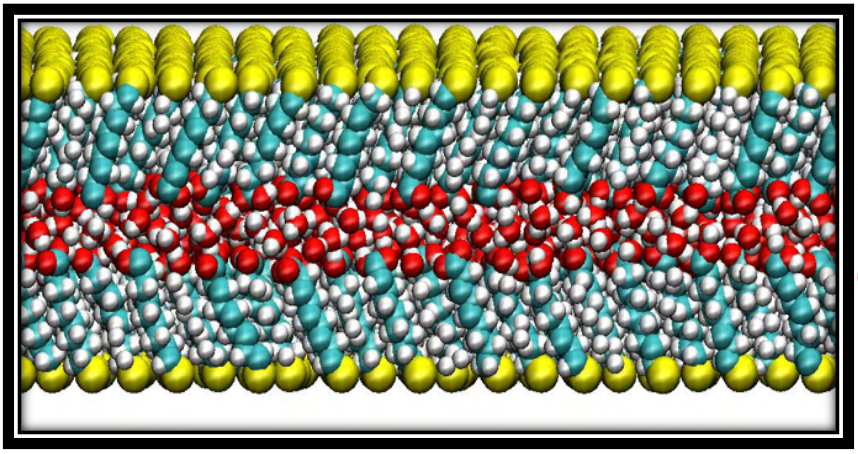
Molecular Simulation: 3. LAMMPS

LAMMPS

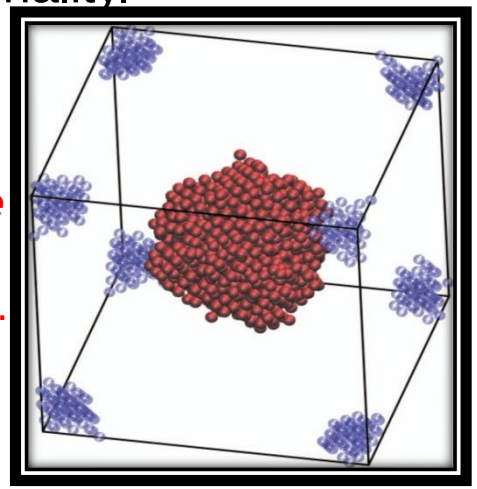


http://lammps.sandia.gov/index.html

- LAMMPS is a **classical molecular dynamics code**, and an acronym for Large-scale Atomic/Molecular Massively Parallel Simulator.
- LAMMPS has potentials for **soft materials** (biomolecules, polymers) and **solid-state materials** (metals, semiconductors) and **coarse-grained or mesoscopic systems**. It can be used to model atoms or, more generically, as a parallel particle simulator at the atomic, meso, or continuum scale.
- LAMMPS runs on single processors or in parallel using message-passing techniques and a spatial-decomposition of the simulation domain. The code is designed to be easy to modify or extend with new functionality.



S. Plimpton, ***Fast Parallel Algorithms for Short-Range Molecular Dynamics***, J Comp Phys, 117, 1-19 (1995).



LAMMPS Benchmark cases

- Two versions of the code used in these studies, 20180822 and 20190605. The GPU version of the latter release also deployed.
- Two of the standard LAMMPS Benchmark cases:
  1. The first is a **standard short-range Lennard-Jones (LJ) potential**;
  2. the second represents a **full protein, Rhodopsin**, including bonds, angles,, dihedrals, constraints etc, along with long-range force computes evaluation using the LAMMPS default implementation of a particle mesh approach.

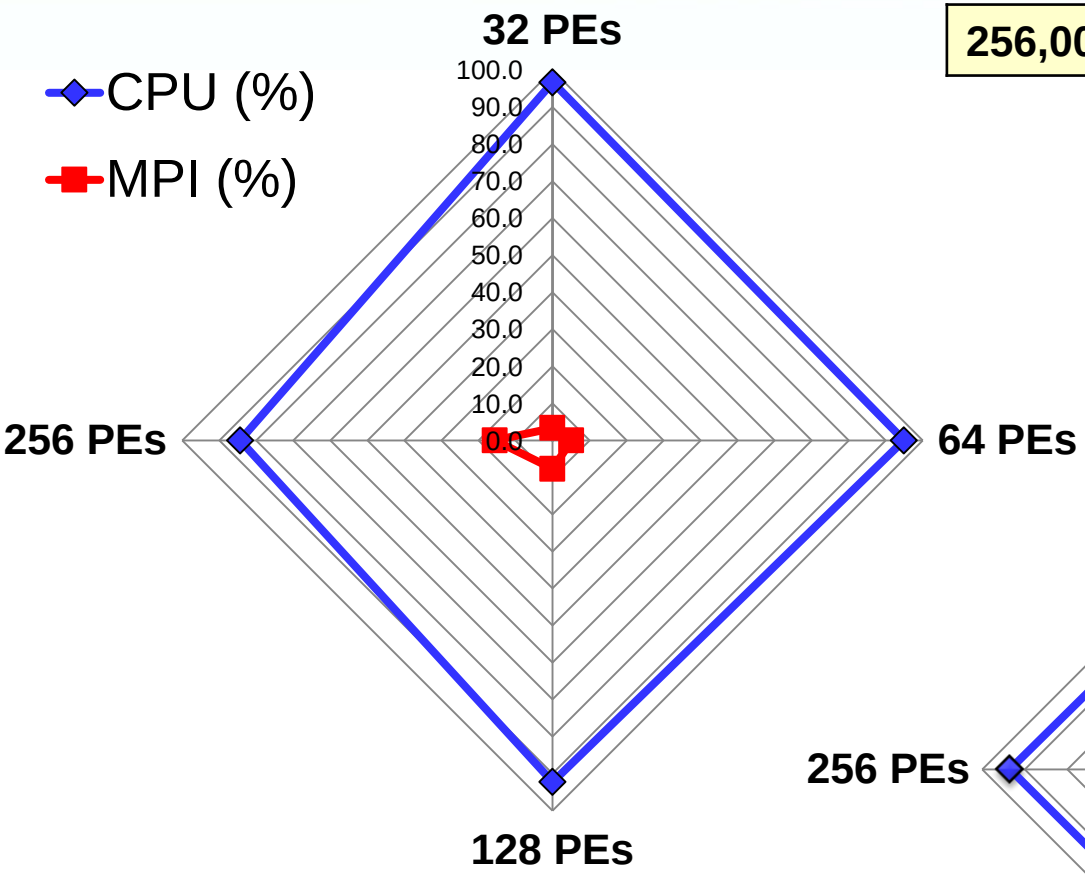
     Standard LAMMPS input files where unaltered other than to increase the number of time steps from 100 to 10,000 in all cases. The number of atoms/particles was set to 512,000 in the rhodopsin calculation via:

     ```
     ./lmp_exe -var x 4 -var y 2 -var z 2 <
     in.rhodo.scaled
     ```

LAMMPS performance on ARCHER/ARCHER-KNL Authors, EPCC Version 0.1, March 17, 2017

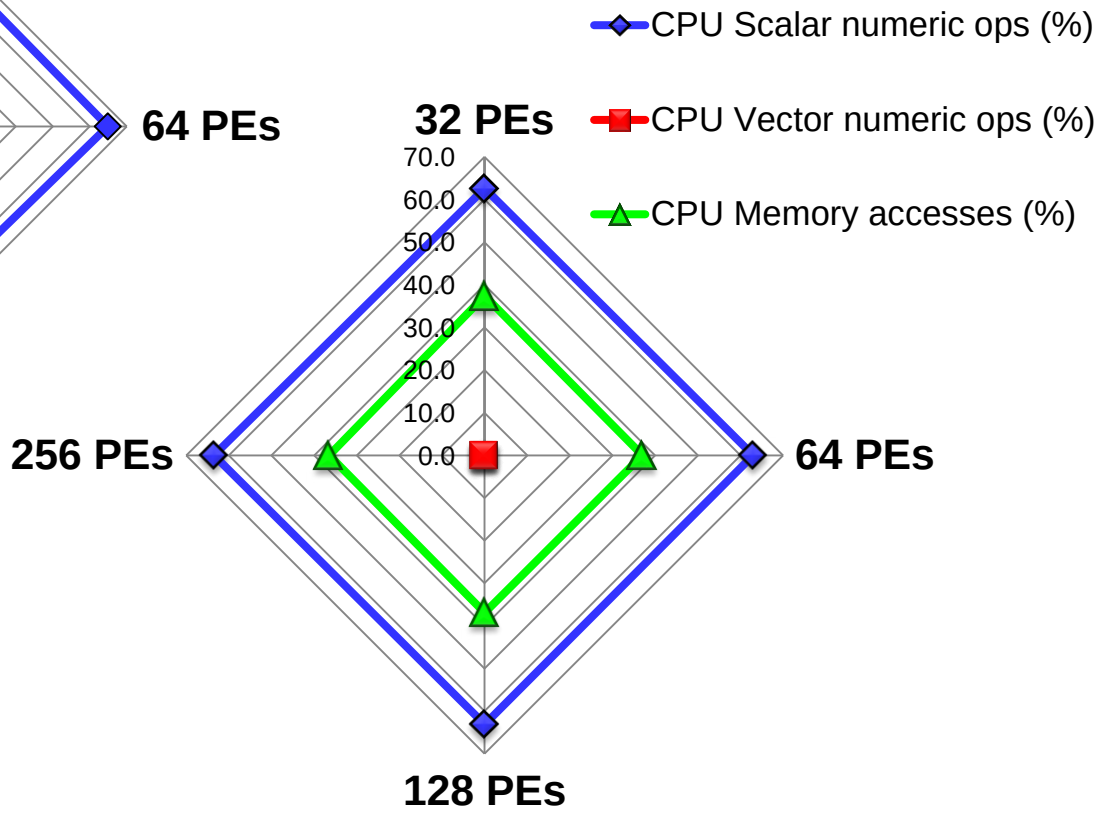
LAMMPS –Lennard-Jones Fluid - Performance Report

256,000 atoms; 5,000 time steps



Total Wallclock Time Breakdown

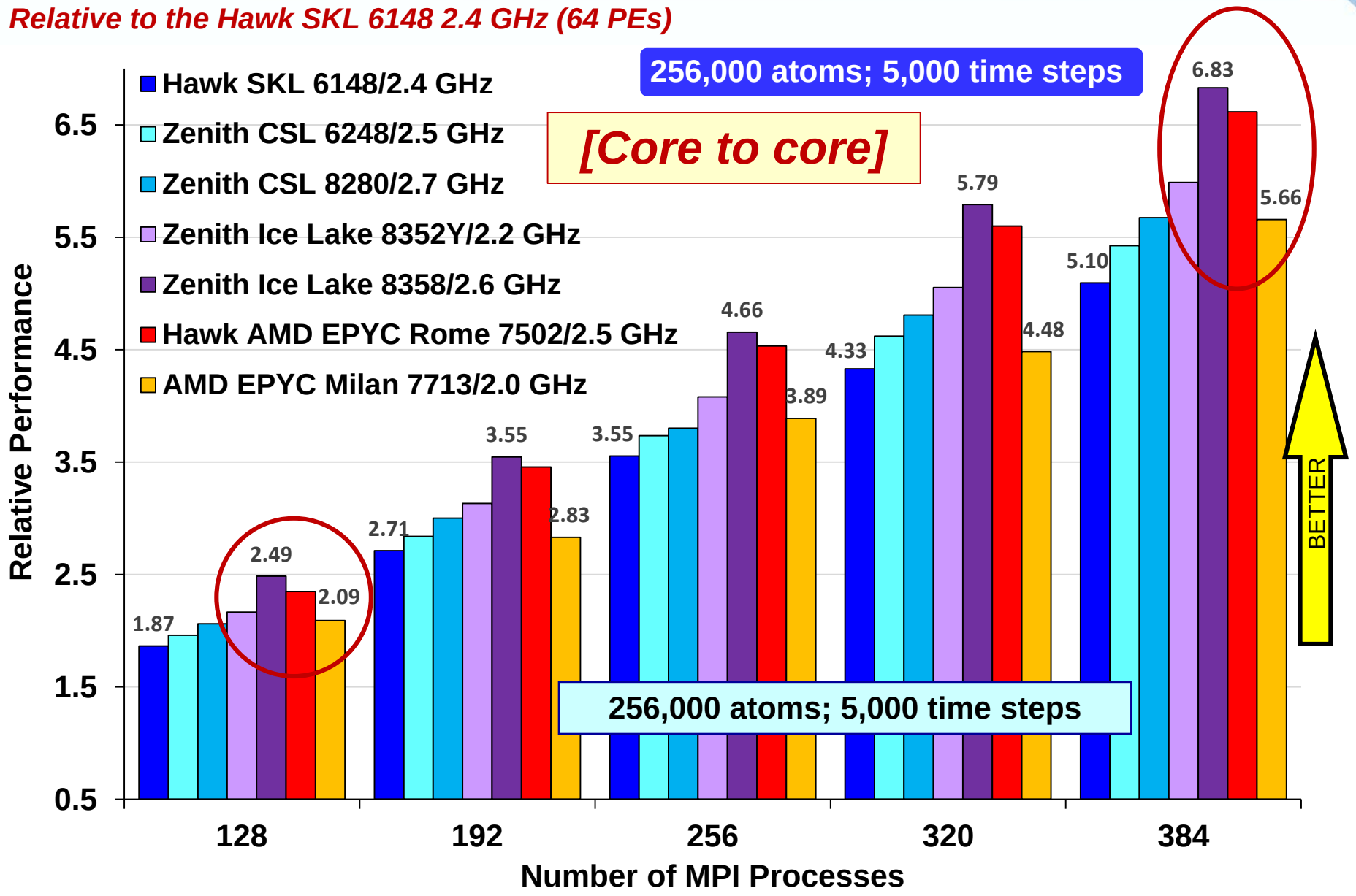
Performance Data (32-256 PEs)



CPU Time Breakdown

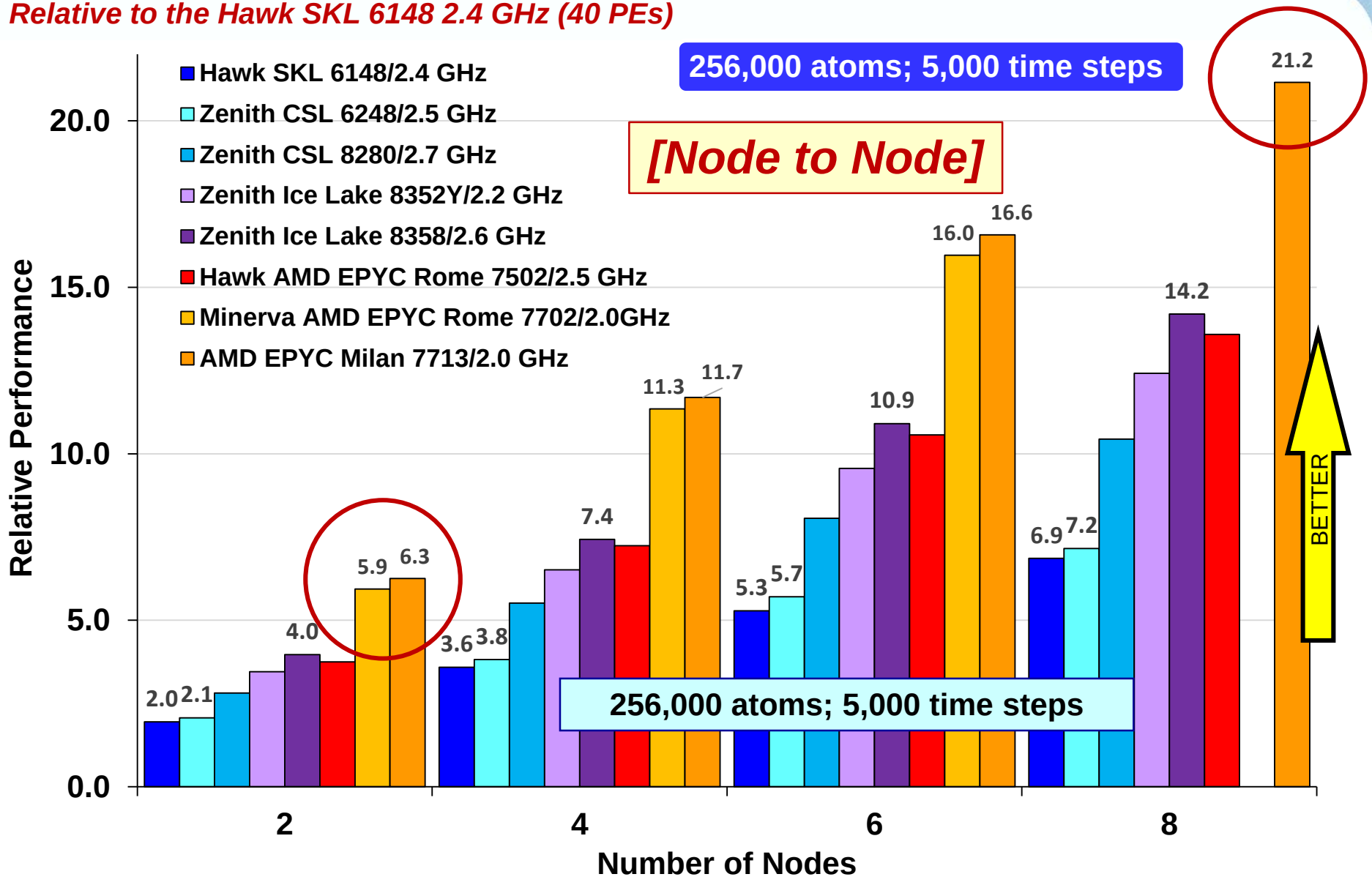
LAMMPS – Atomic fluid with Lennard-Jones Potential

Relative to the Hawk SKL 6148 2.4 GHz (64 PEs)



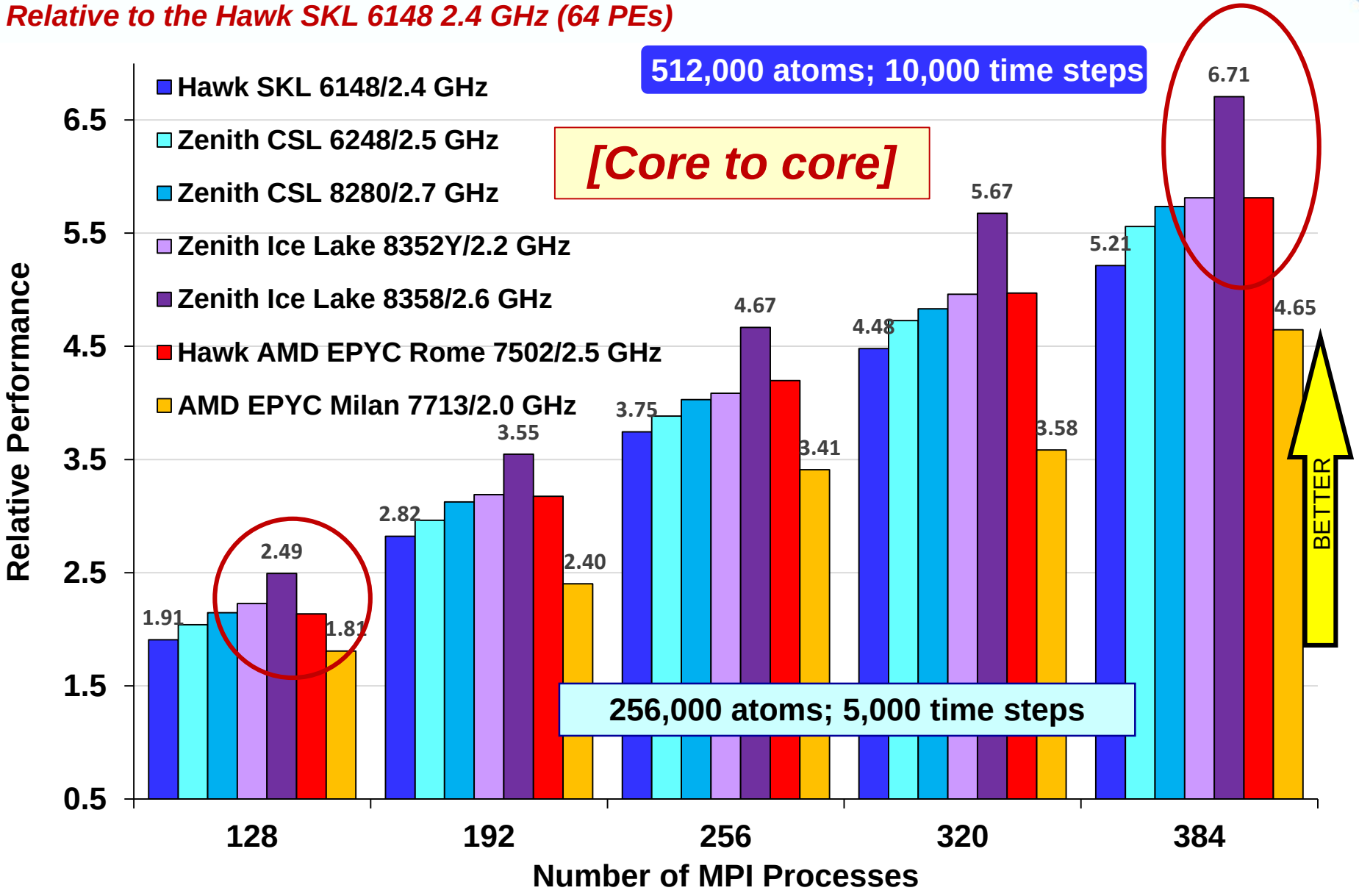
LAMMPS – Atomic fluid with Lennard-Jones Potential

Relative to the Hawk SKL 6148 2.4 GHz (40 PEs)

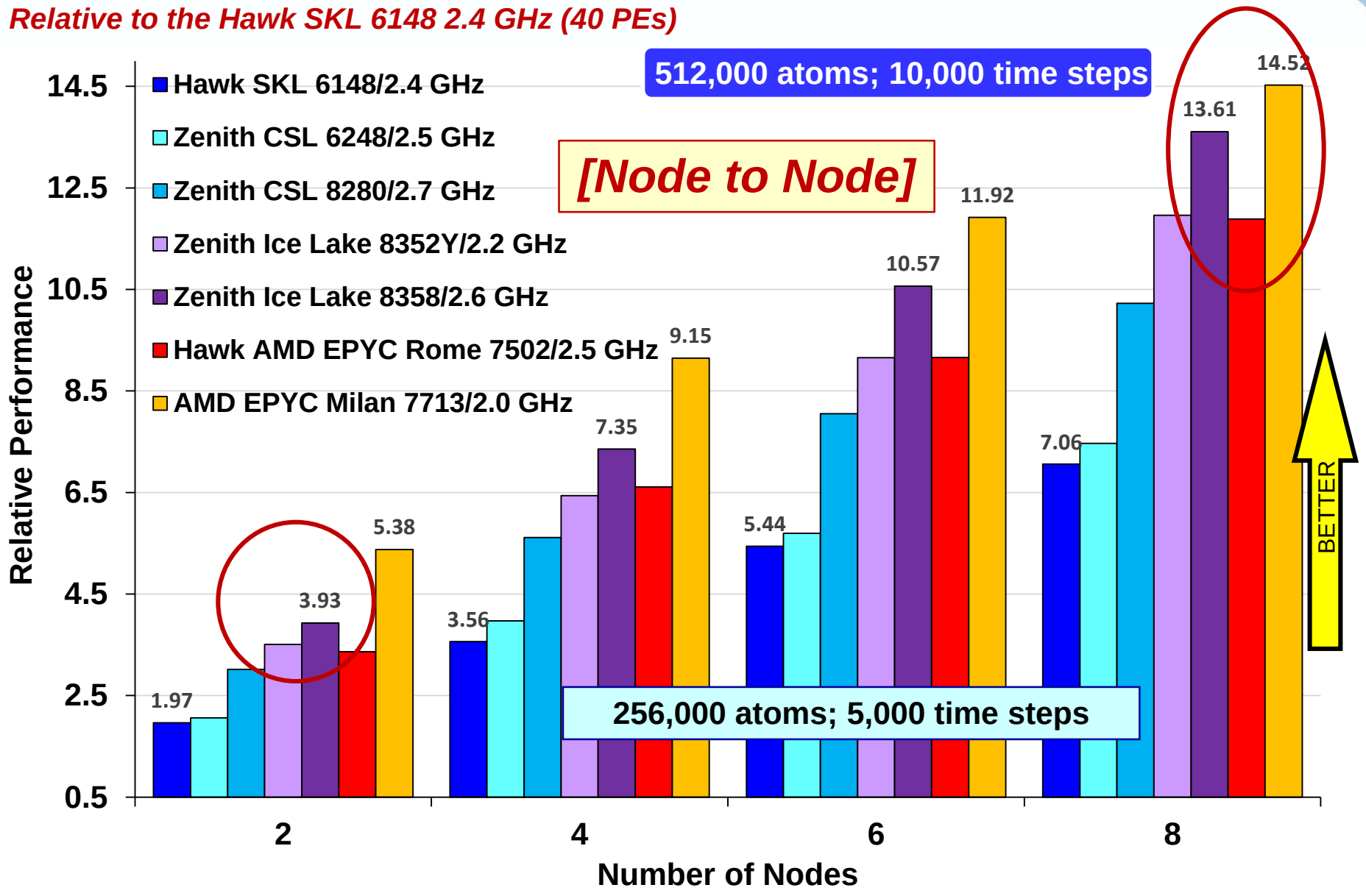


Rhodopsin Scaled Benchmark

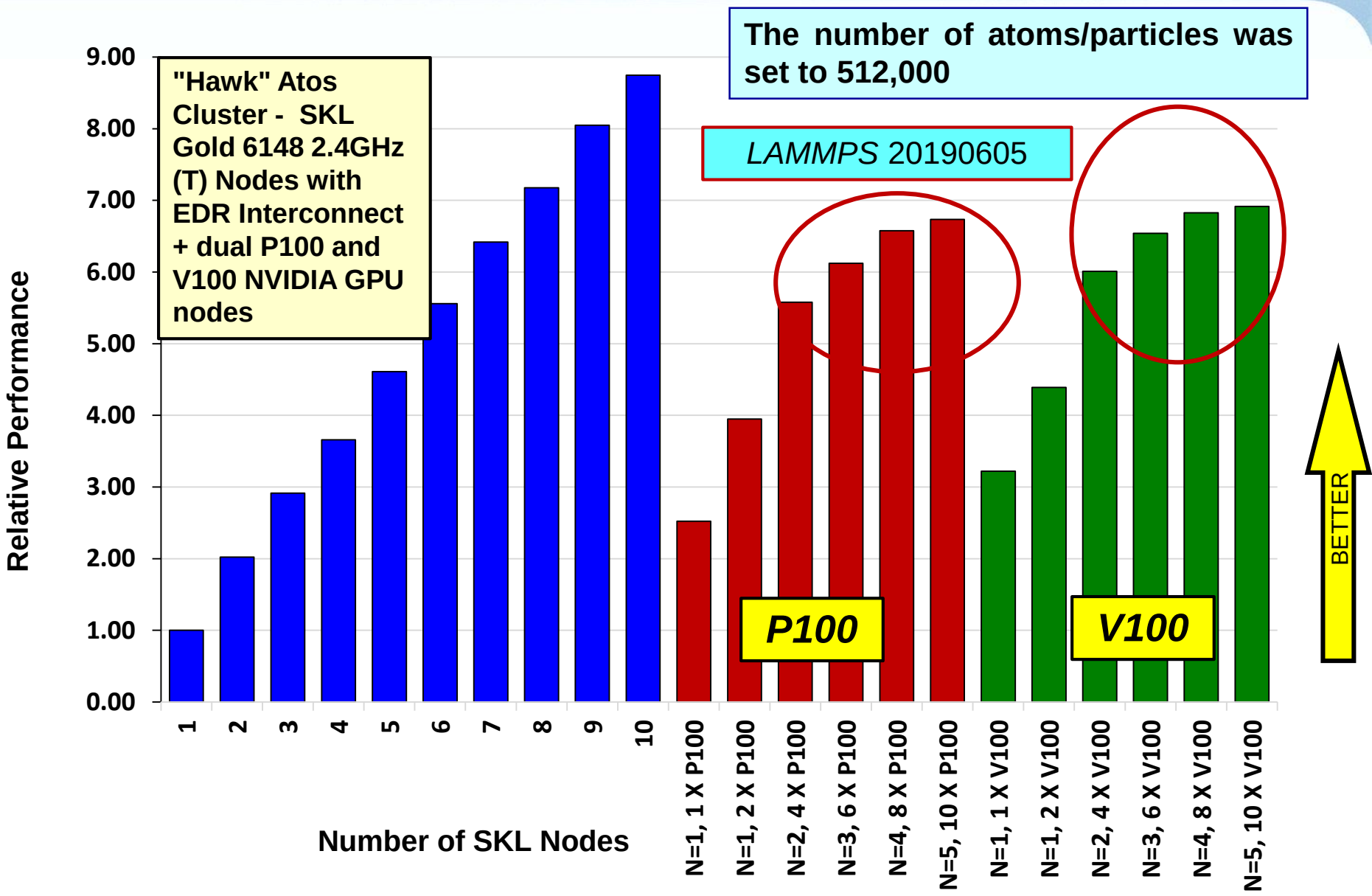
Relative to the Hawk SKL 6148 2.4 GHz (64 PEs)



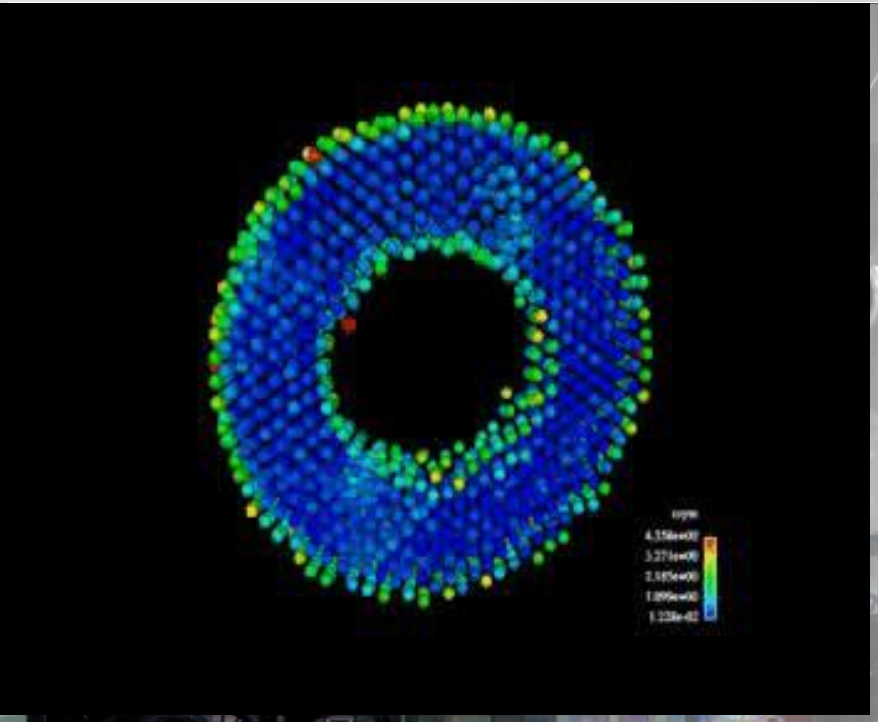
Rhodopsin Scaled Benchmark



LAMMPS – GPU Performance in Rhodopsin Simulation



Performance of Computational Chemistry Codes



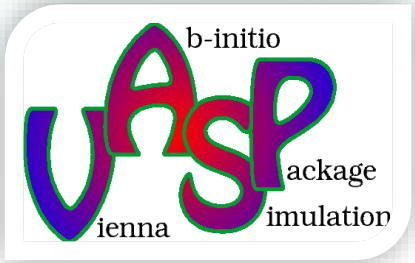
2. Advanced Materials Software: VASP

Advanced Materials Software

Computational Materials

- **VASP** – performs ab-initio QM molecular dynamics (MD) simulations using **pseudopotentials** or the projector-augmented wave method and a plane wave basis set.
- **Quantum Espresso** – an integrated suite of Open-Source computer codes for electronic-structure calculations and materials modelling at the nanoscale. It is based on density-functional theory (**DFT**), plane waves, and **pseudopotentials**
- **SIESTA** - an *O*(N) **DFT** code for electronic structure calculations and *ab initio* molecular dynamics simulations for molecules and solids. It uses norm-conserving **pseudopotentials** and linear combination of numerical atomic orbitals (LCAO) basis set.
- **CP2K** is a program to perform atomistic and molecular simulations of solid state, liquid, molecular, and biological systems. It provides a framework for different methods such as e.g., **DFT** using a mixed Gaussian & plane waves approach (GPW) and classical pair and many-body potentials.
- **ONETEP** (Order-N Electronic Total Energy Package) is a linear-scaling code for quantum-mechanical calculations based on **DFT**.

VASP – Vienna *Ab-initio* Simulation Package



Archer Rank: 1

VASP (**5.4.4**) performs ab-initio QM molecular dynamics (MD) simulations using pseudopotentials or the projector-augmented wave method and a plane wave basis set.

Benchmark	Details
MFI Zeolite	Zeolite ($Si_{96}O_{192}$), 2 k-points, FFT grid: (65, 65, 43); 181,675 points
Pd-O complex	Palladium-Oxygen complex ($Pd_{75}O_{12}$), 10 k-points, FFT grid: (31, 49, 45), 68,355 points

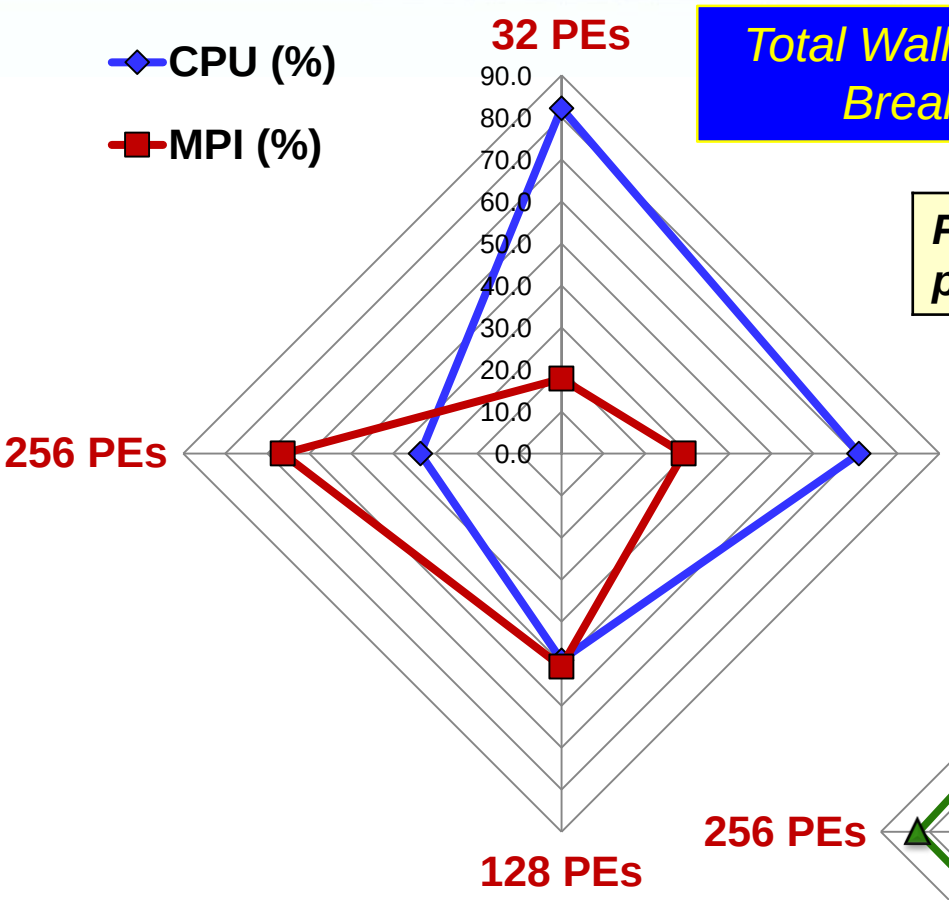
Pd-O Benchmark

- Pd-O complex – $Pd_{75}O_{12}$, 5X4 3-layer supercell running a single point calculation and a planewave cut off of 400eV. Uses the RMM-DIIS algorithm for the SCF and is calculated in real space.
- 10 k-points; maximum number of plane-waves: 34,470
- FFT grid; NGX=31, NGY=49, NGZ=45, giving a total of 68,355 points

Zeolite Benchmark

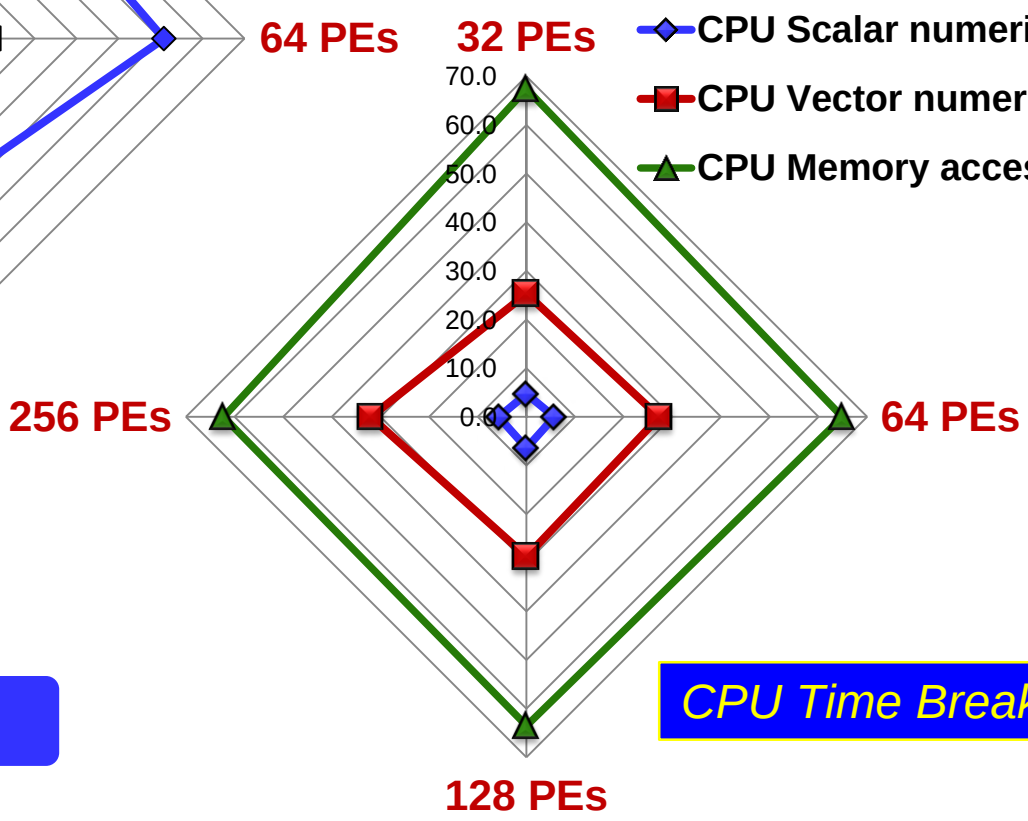
- Zeolite with the MFI structure unit cell running a single point calculation and a planewave cut off of 400eV using the PBE functional
- 2 k-points; maximum number of plane-waves: 96,834
- FFT grid; NGX=65, NGY=65, NGZ=43, giving a total of 181,675 points

VASP – Pd-O Benchmark Performance Report



Total Wallclock Time Breakdown

Palladium-Oxygen complex ($Pd_{75}O_{12}$), 10 k-points, FFT grid: (31, 49, 45), 68,355 points



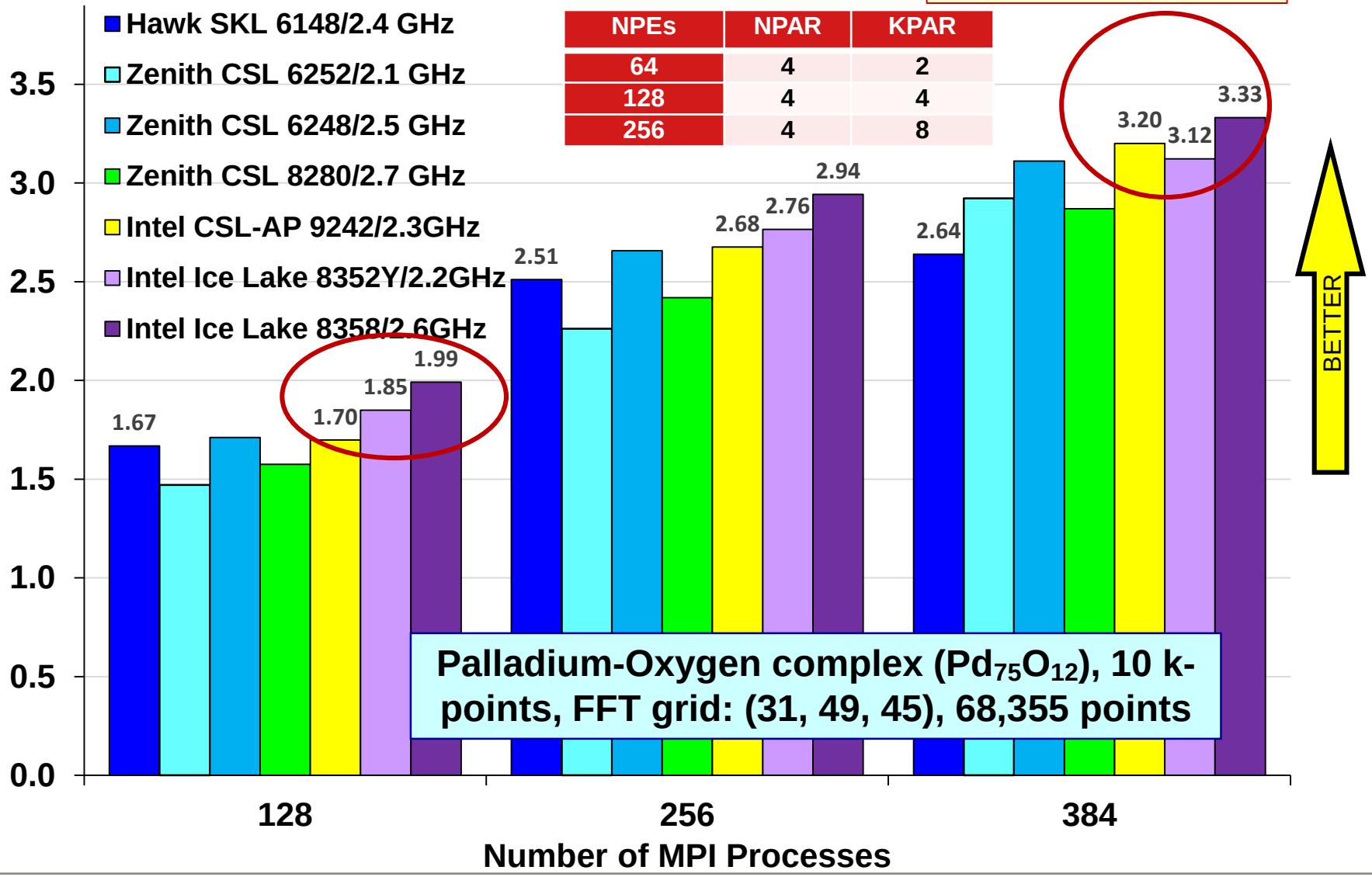
CPU Time Breakdown

Performance Data (32-256 PEs)

VASP 5.4.4 – Pd-O Benchmark - Parallelisation on k-points

Performance *Relative to the Hawk SKL 6148 2.4 GHz (64 PEs)*

[Core to core]

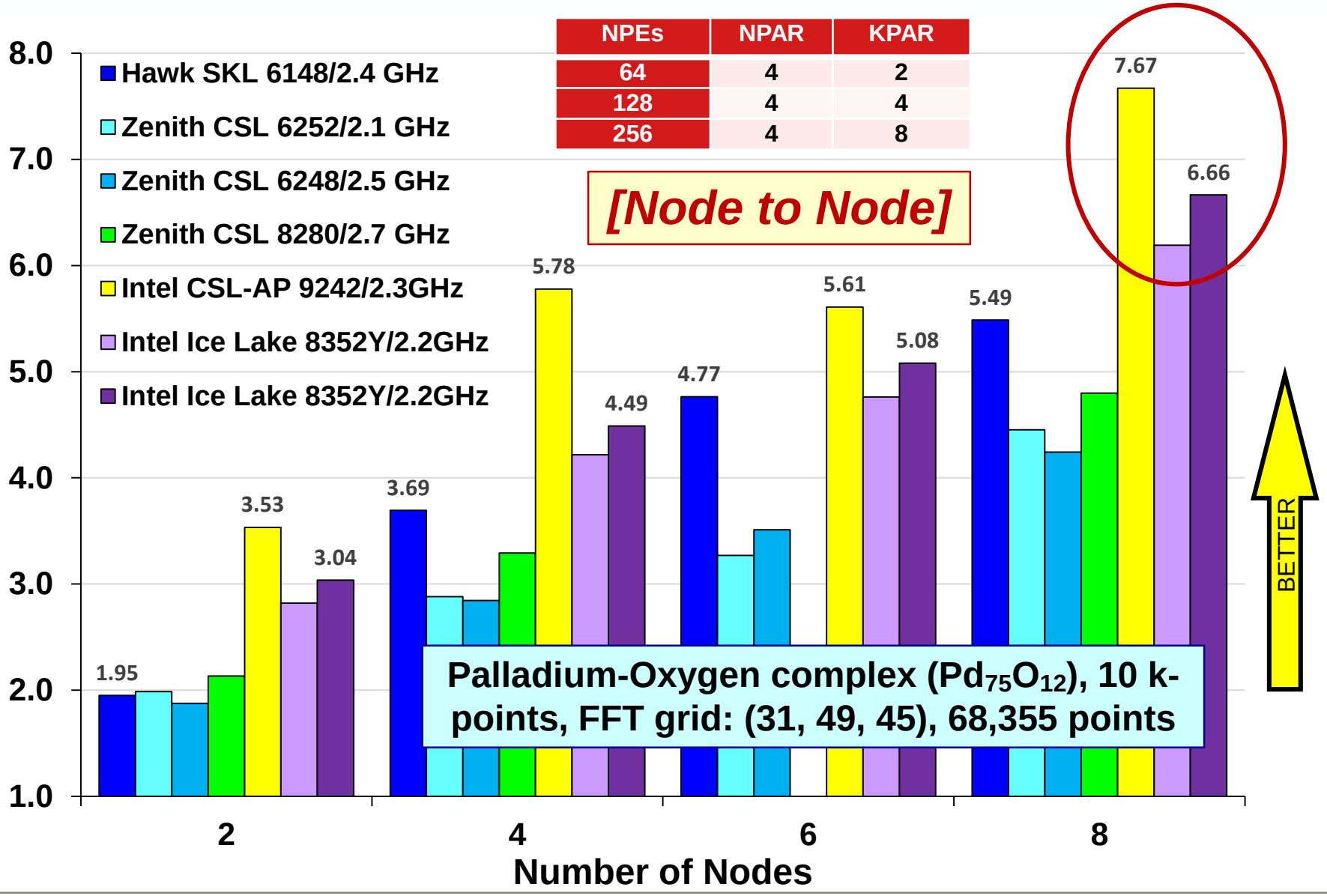


VASP 5.4.4 – Pd-O Benchmark - Parallelisation on k-points

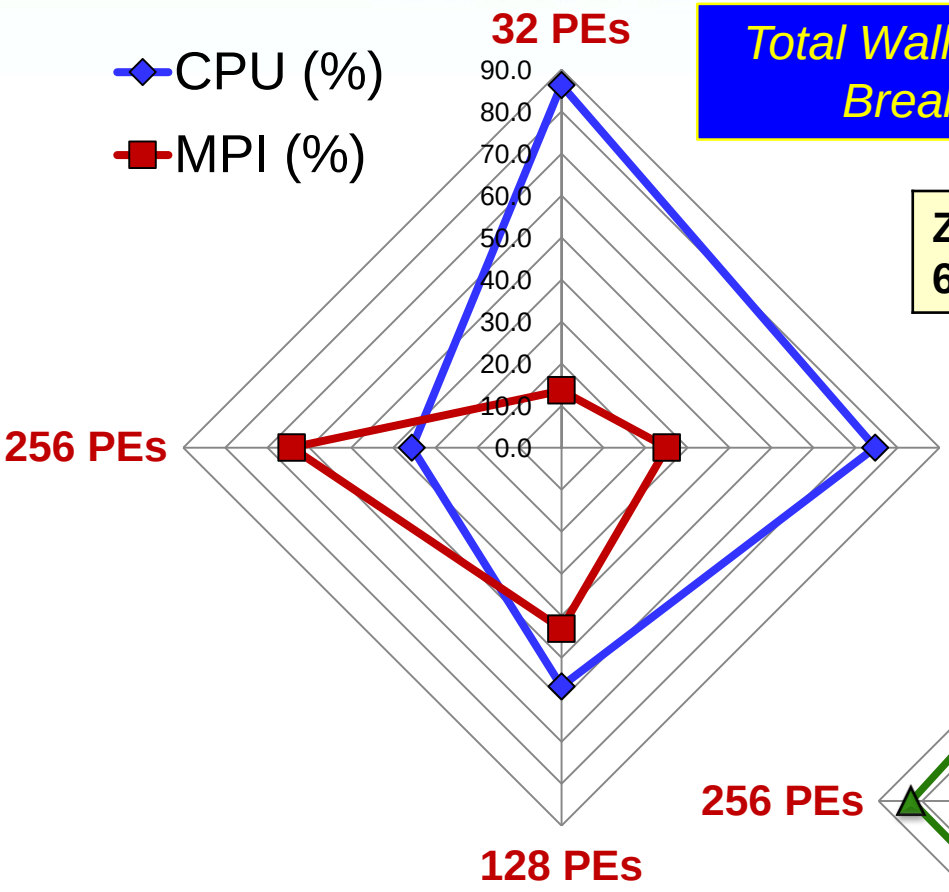
Performance *Relative to the Hawk SKL 6148 2.4 GHz (1 Node)*

NPEs	NPAR	KPAR
64	4	2
128	4	4
256	4	8

[Node to Node]

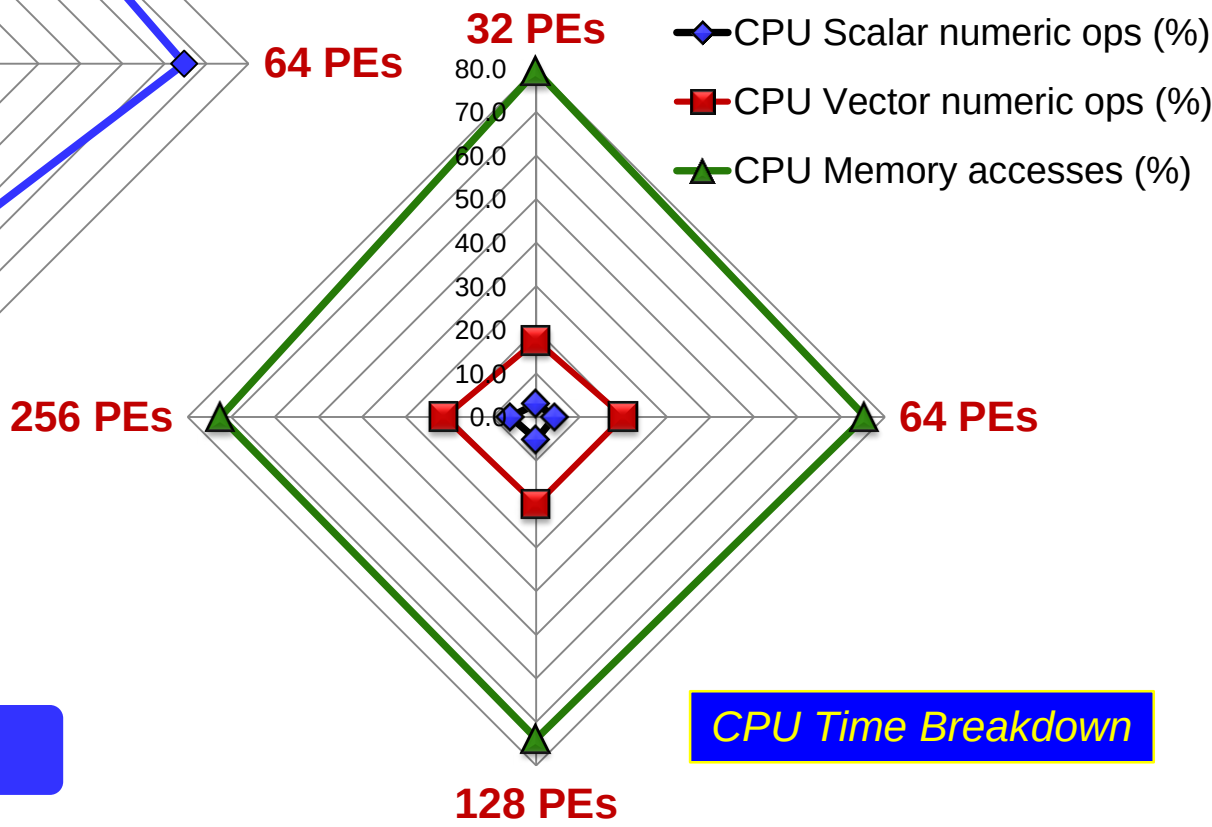


VASP – Zeolite Cluster Performance Report



Total Wallclock Time Breakdown

Zeolite ($Si_{96}O_{192}$), 2 k-points, FFT grid: (65, 65, 43); 181,675 points

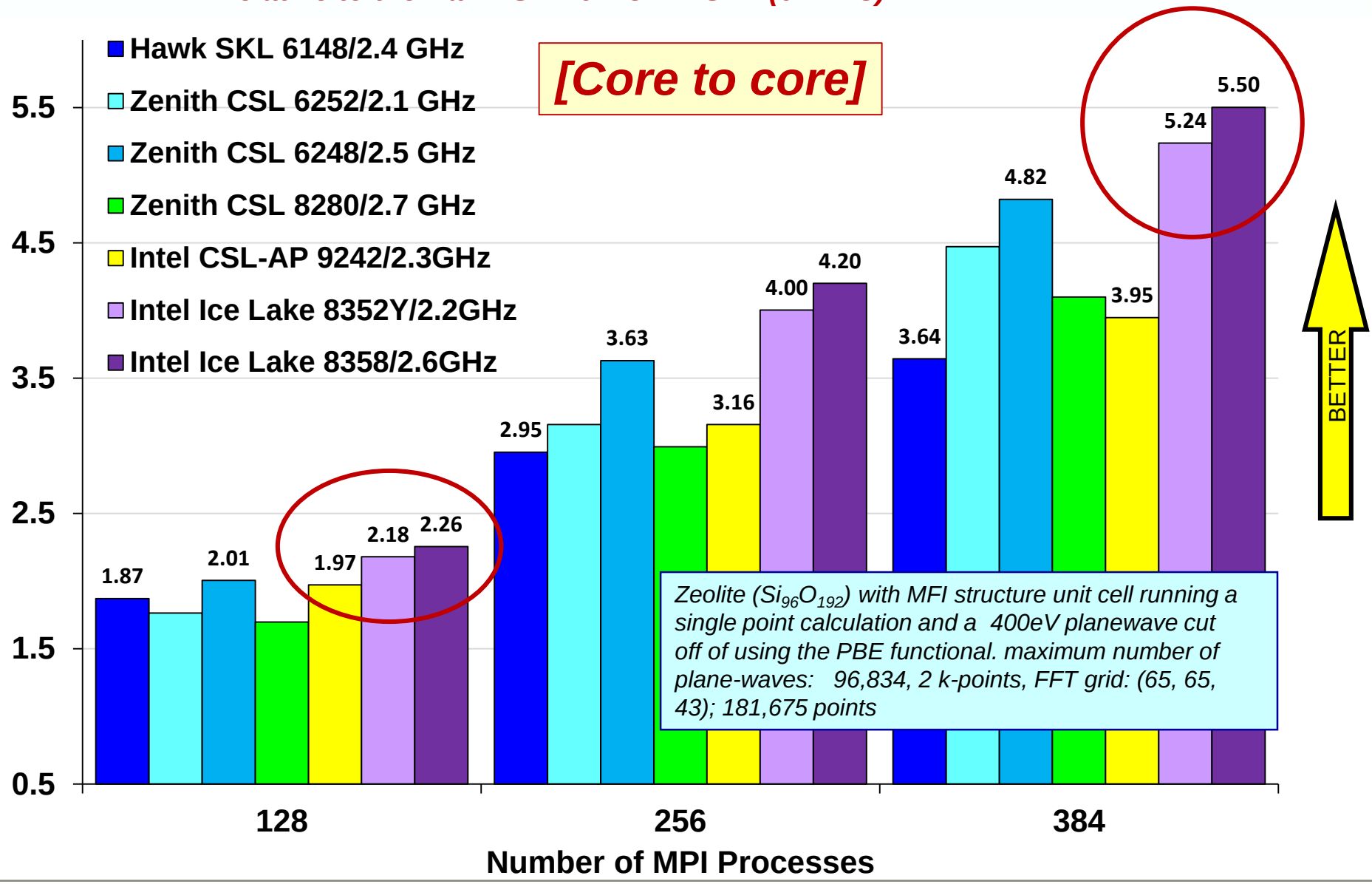


CPU Time Breakdown

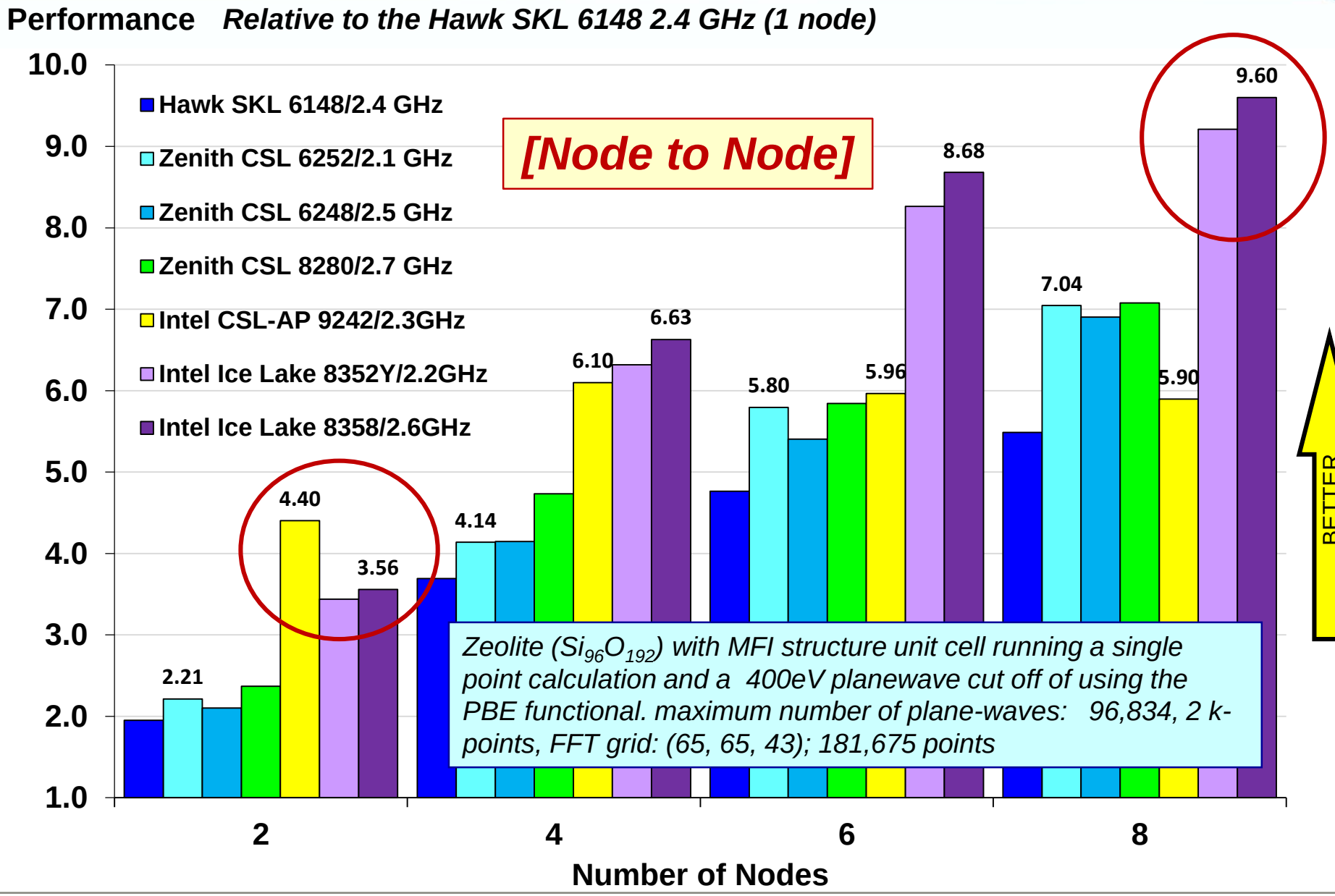
Performance Data (32-256 PEs)

VASP 5.4.4 – Zeolite Benchmark - Parallelisation on k-points

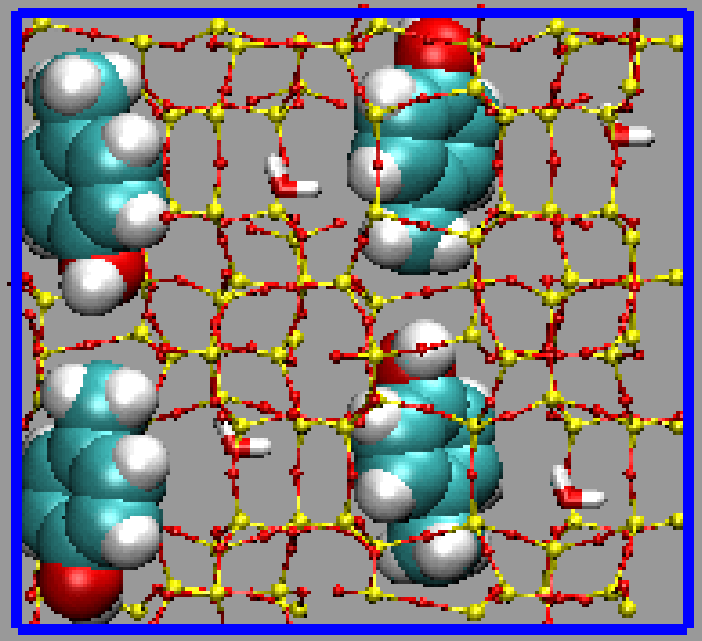
Performance *Relative to the Hawk SKL 6148 2.4 GHz (64 PEs)*



VASP 5.4.4 – Zeolite Benchmark - Parallelisation on k-points

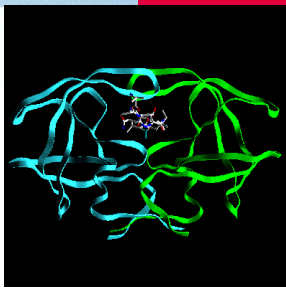


Performance of Computational Chemistry Codes



3. Electronic Structure – GAMESS-UK

The MPI/ScaLAPACK Implementation of the GAMESS-UK SCF/DFT module

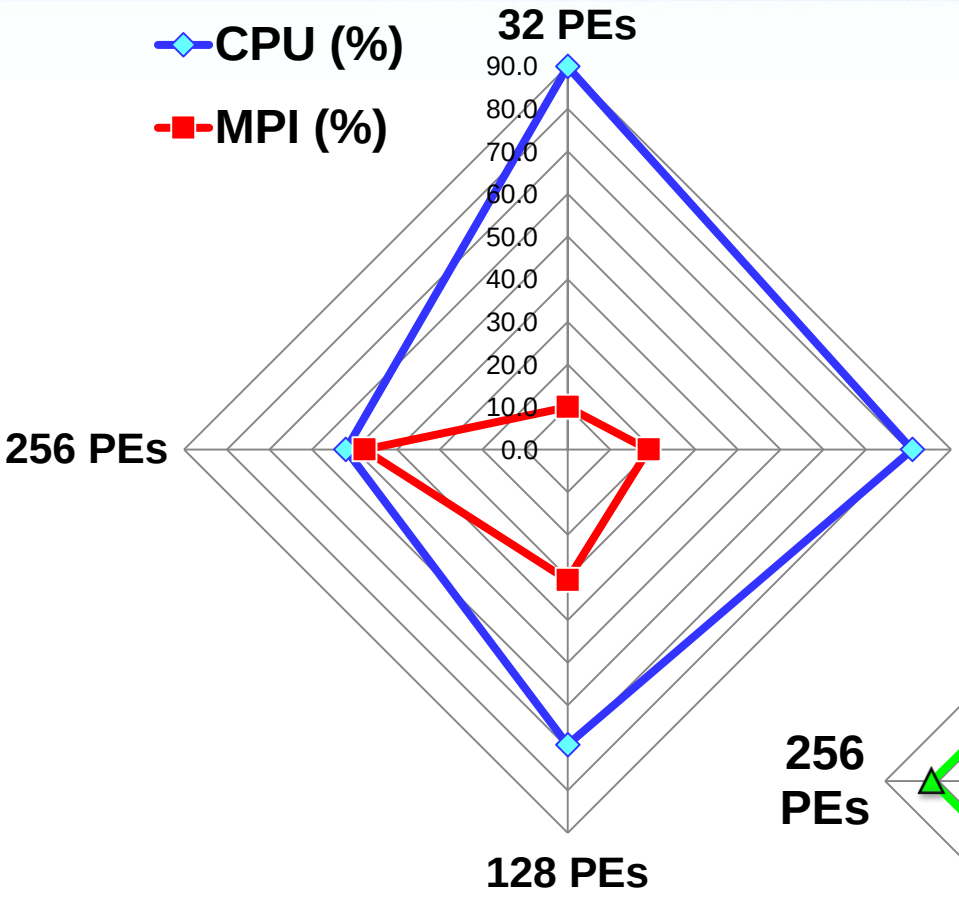


- Pragmatic approach to the replicated data constraints:
- MPI-based tools (such as ScaLAPACK) used in place of Global Arrays
- All data structures except those required for the Fock matrix build are fully distributed (F, P)

- Partially distributed model chosen because, in the absence of efficient one-sided communications it is difficult to efficiently load balance a distributed Fock matrix build.
- Obvious drawback - some large replicated data structures are required.
  - These are kept to a minimum. For a closed shell HF or DFT calculation only **2 replicated matrices** are required, 1 × Fock and 1 × Density (doubled for UHF).

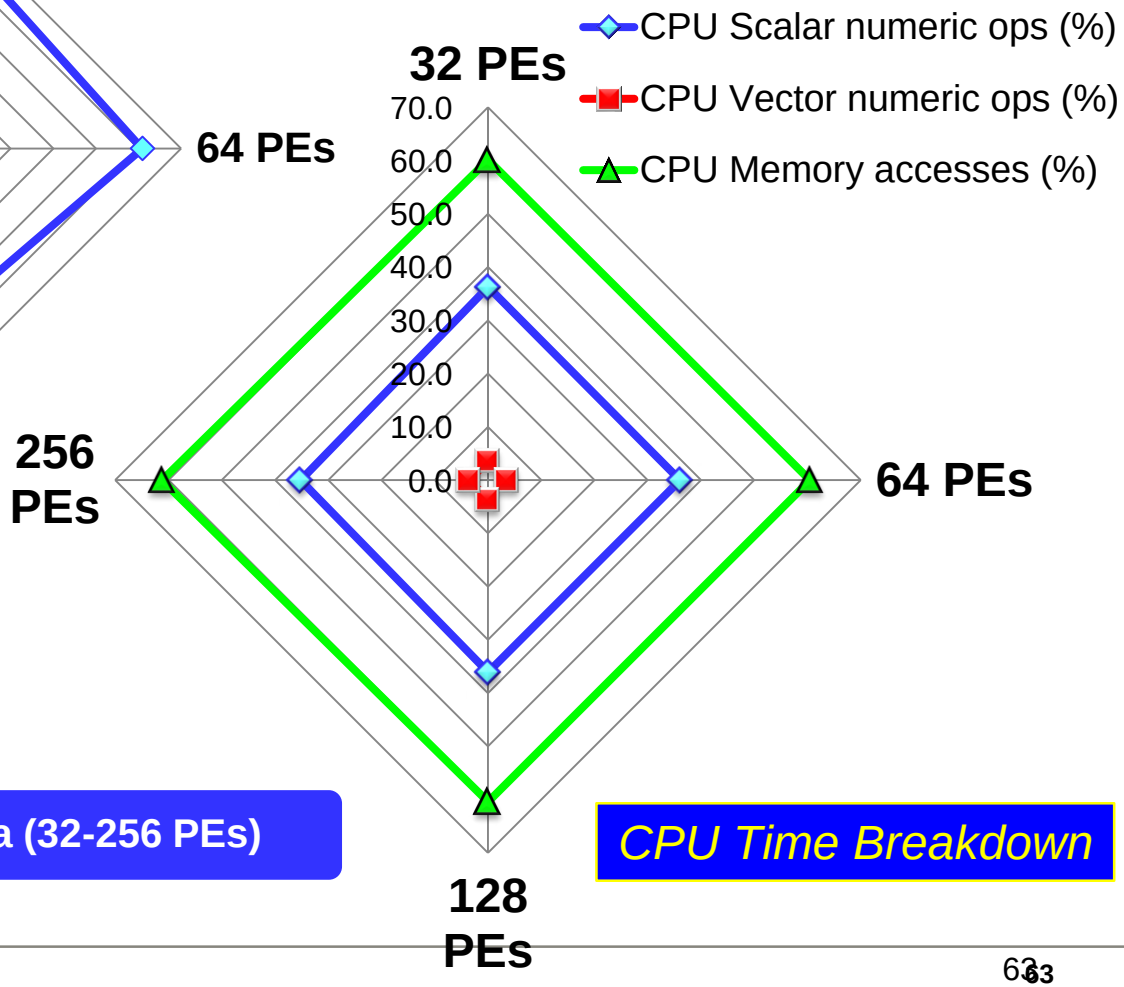
"The GAMESS-UK electronic structure package: algorithms, developments and applications"
M.F. Guest, I. J. Bush, H.J.J. van Dam, P. Sherwood, J.M.H. Thomas, J.H. van Lenthe, R.W.A Havenith, J. Kendrick, Mol. Phys. 103, No. 6-8, 2005, 719-747.

GAMESS-UK.MPI DFT – DFT Performance Report



Cyclosporin 6-31G** basis (1855 GTOs); DFT B3LYP

Total Wallclock Time Breakdown



Performance Data (32-256 PEs)

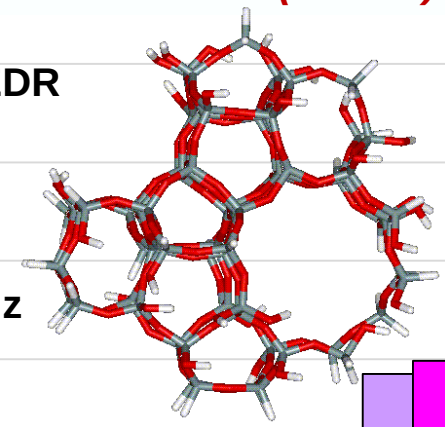
CPU Time Breakdown

GAMESS-UK Performance - Zeolite Y cluster

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

[Core to core]

Performance *Relative to the Hawk SKL 6148 2.4 GHz (64 PEs)*



- Hawk SKL 6148 2.4 GHz (T) EDR
- Zenith CSL 6248/2.5 GHz
- Intel CSL-AP 9242/2.3 GHz
- Zenith Ice Lake 8352Y/2.2 GHz
- Intel Ice Lake 8360Y/2.4 GHz
- Zenith Ice Lake 8358/2.6 GHz
- Intel Ice Lake 8368Q/2.6 GHz

1.00 1.21 1.48 1.80 1.94 2.71 2.58 2.70 3.89 3.29 3.49 4.97

Performance Data (64-256 PEs)

BETTER

Zeolite Y cluster SioSi7 DZVP (Si,O), DZVP2 (H) B3LYP(3975 GTOs)

5.0 4.5 4.0 3.5 3.0 2.5 2.0 1.5 1.0 0.5 0.0

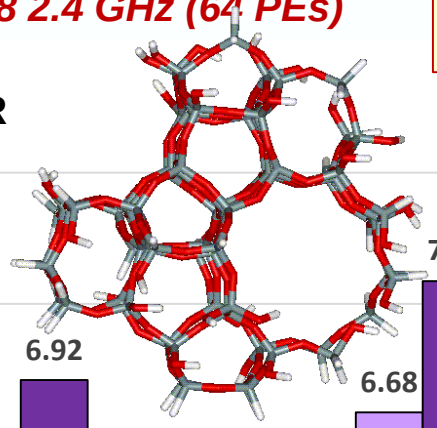
64 128 192 256

Number of MPI Processes

GAMESS-UK Performance - Zeolite Y cluster

Performance *Relative to the Hawk SKL 6148 2.4 GHz (64 PEs)*

[Core to core]



Hawk SKL 6148 2.4 GHz (T) EDR
Zenith CSL 6248/2.5 GHz
Intel CSL-AP 9242/2.3 GHz
Zenith Ice Lake 8352Y/2.2 GHz
Zenith Ice Lake 8358/2.6 GHz

8.5
7.5
6.5
5.5
4.5
3.5
2.5
1.5

3.88 4.39 4.86 5.53
5.22 5.99 6.68 7.52
6.00 6.92 7.68 8.56

320 384 448 512

Number of MPI Processes

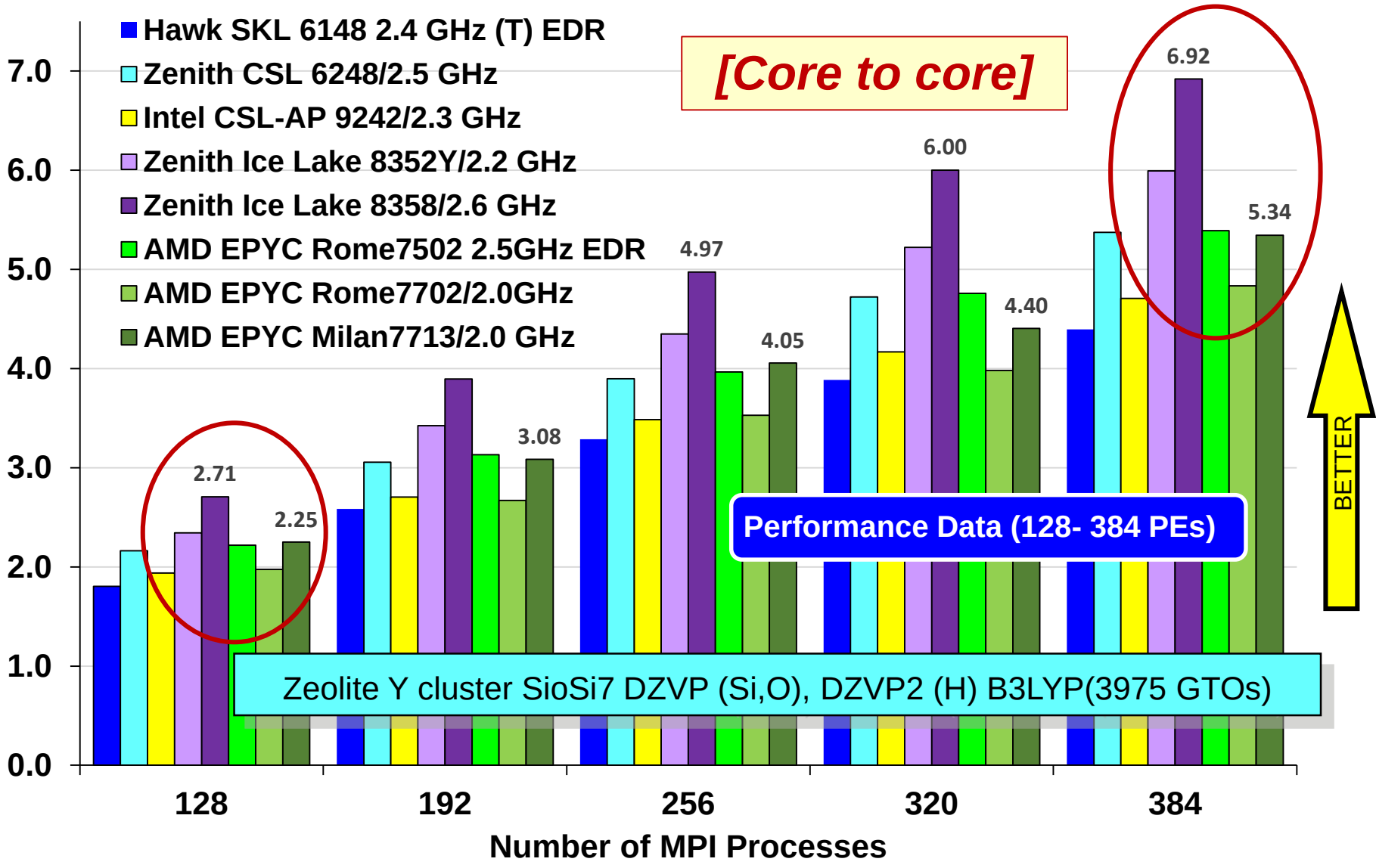
BETTER

Performance Data (320- 512 PEs)

Zeolite Y cluster SioSi7 DZVP (Si,O), DZVP2 (H) B3LYP(3975 GTOs)

GAMESS-UK Performance - Zeolite Y cluster

Performance *Relative to the Hawk SKL 6148 2.4 GHz (64 PEs)*

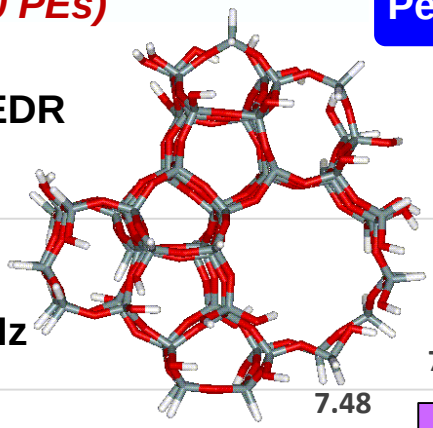


GAMESS-UK Performance - Zeolite Y cluster

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Relative to the Hawk SKL 6148 2.4 GHz (40 PEs)

Performance Data (1 – 6 Nodes)



Hawk SKL 6148 2.4 GHz (T) EDR
Zenith CSL 6248/2.5 GHz
Intel CSL-AP 9242/2.3 GHz
Zenith Ice Lake 8352Y/2.2 GHz
Intel Ice Lake 8360Y/2.4 GHz
Zenith Ice Lake 8358/2.6 GHz
Intel Ice Lake 8368Q/2.6 GHz

[Node to Node]

Performance

10.0
8.0
6.0
4.0
2.0
0.0

1.00 2.42 2.37 2.42
1.88 4.30 4.34 4.64
3.51 7.48 7.97 8.11
4.74 9.91 11.10

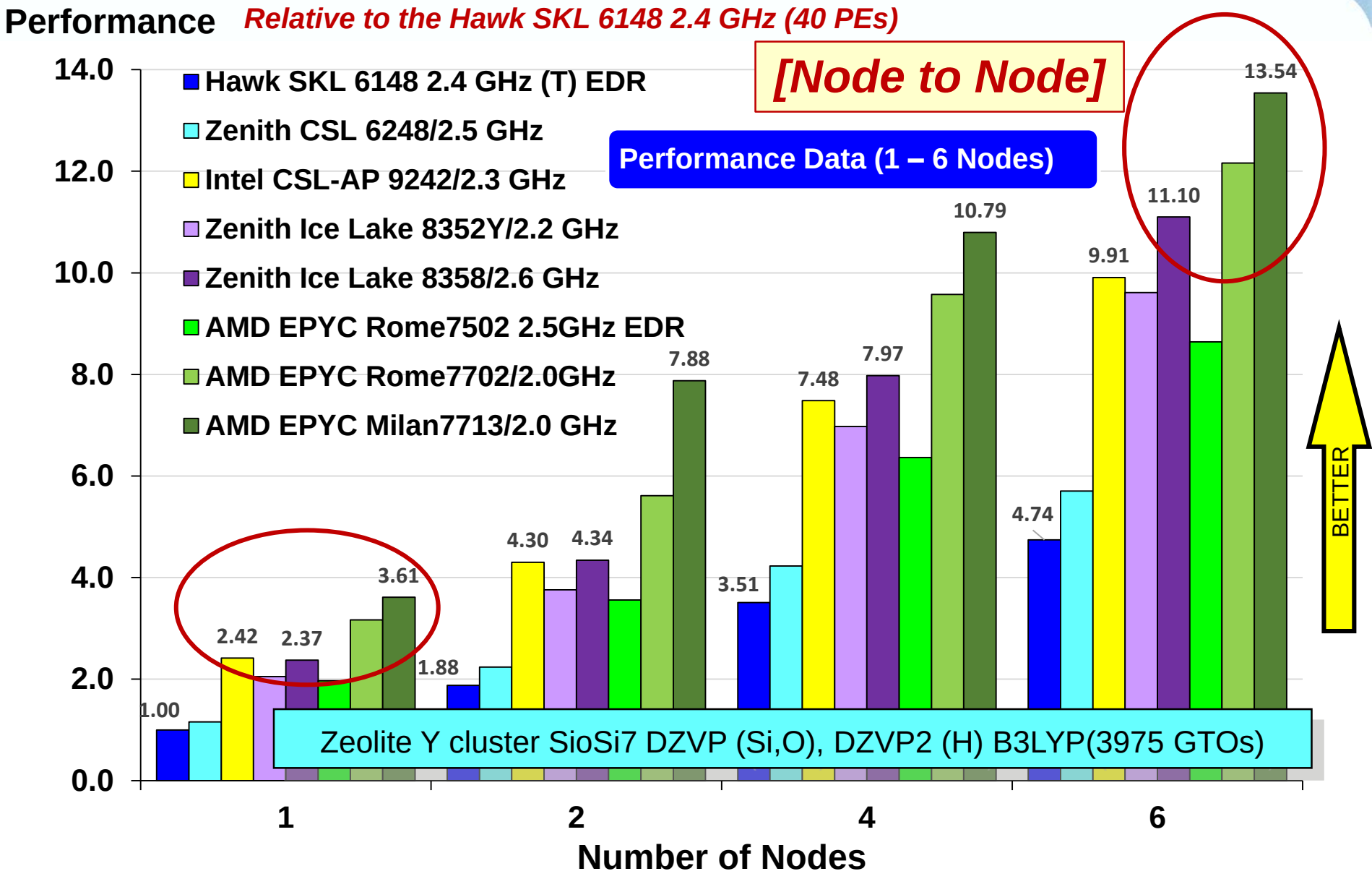
1 2 4 6

Number of Nodes

BETTER

Zeolite Y cluster SioSi7 DZVP (Si,O), DZVP2 (H) B3LYP(3975 GTOs)

GAMESS-UK Performance - Zeolite Y cluster



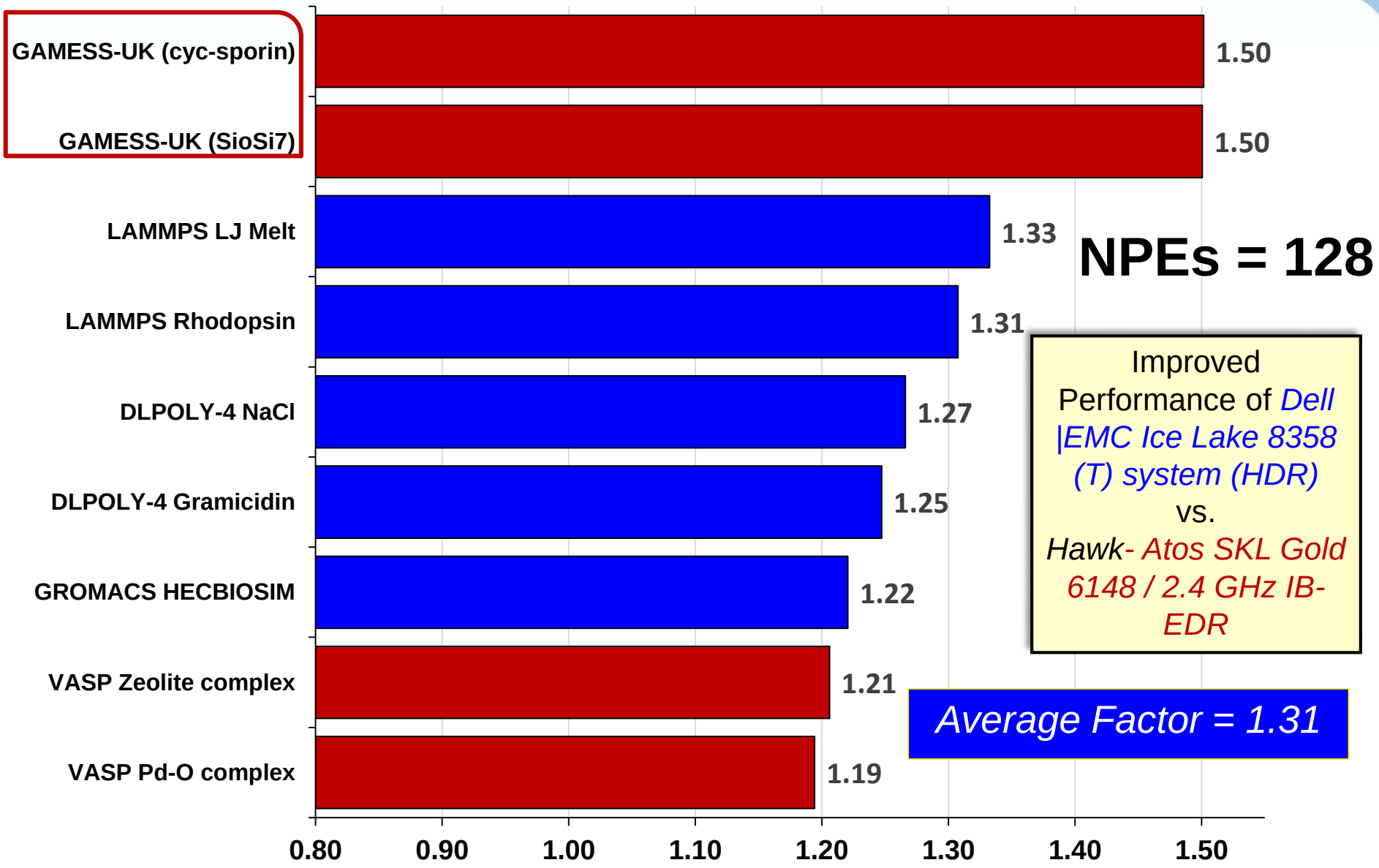
Performance of Computational Chemistry Codes

Relative Performance as a Function of Processor Family

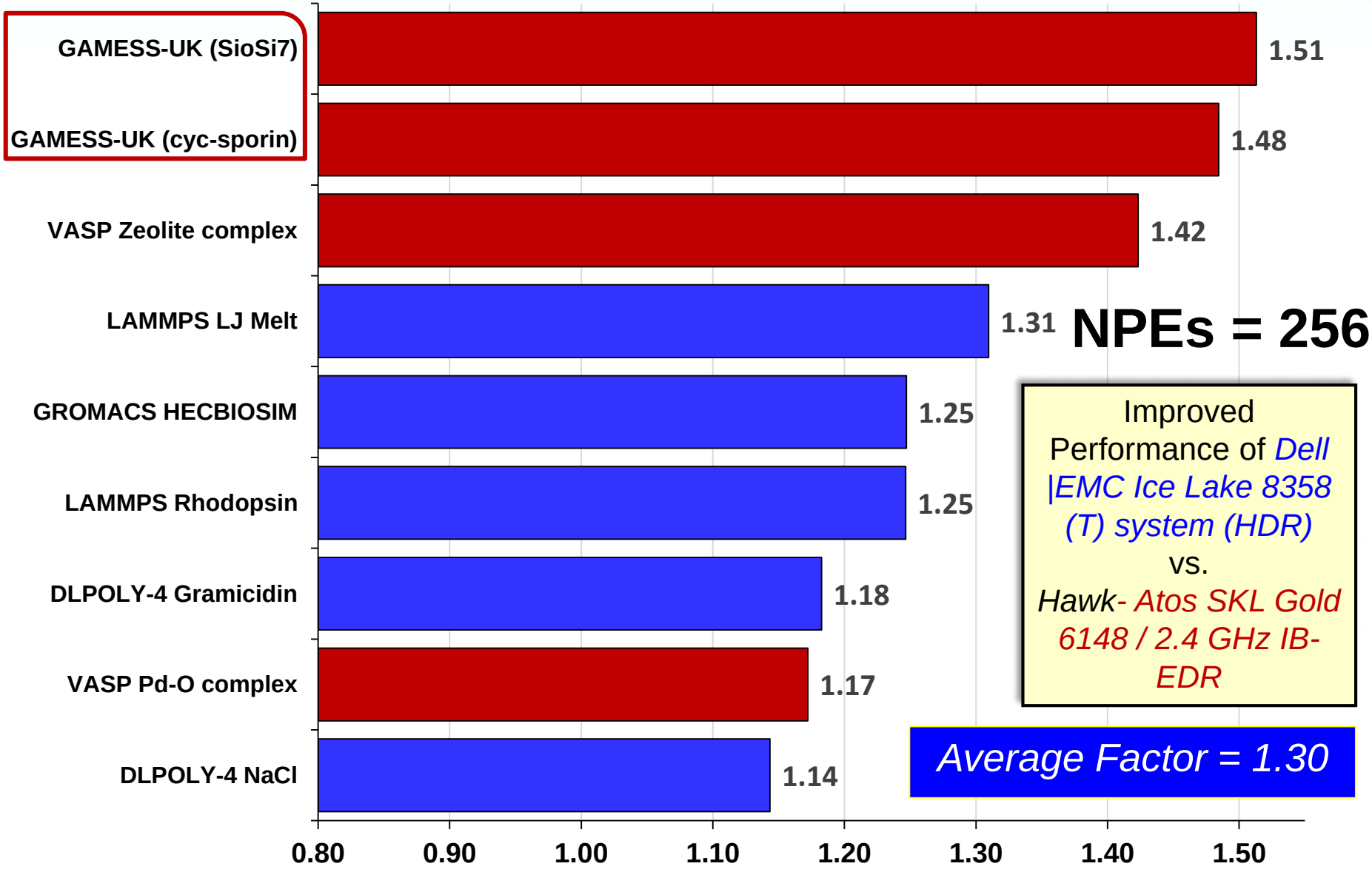
Summary – Core-to-Core Comparisons

- A ***Core-to-Core* comparison** suggests on average that the Intel Ice Lake 8358 2.6 GHz SKU outperforms all others, although relative performance is sensitive to effective use of the AVX instructions.
- Low utilisation of AVX-512 leads to weaker performance of the SKL, CSL and Ice Lake CPUs and **better performance of the Rome & Milan-based clusters** e.g. DLPOLY, LAMMPS
- With significant AVX-512 utilisation, Ice Lake Lake systems outperform the AMD Milan systems in core-to-core comparisons e.g. Gromacs, notwithstanding the use of AVX2-256.
- **Modest improvement** at best on moving from Skylake to Cascade Lake systems – more dramatic improvement moving to Ice Lake.
- **Strong Performance** of the **CSL Gold 6248 system** (2.5Ghz), but surprisingly weak performance from the CSL Platinum 8280 (2.7GHz)
- **Improved CPU** performance of **Spartan 7742 cluster** across all applications compared to CIUK'20 findings (Turbo Mode!)
- Baselined across **P100** and **V100** NVIDIA GPU performance.

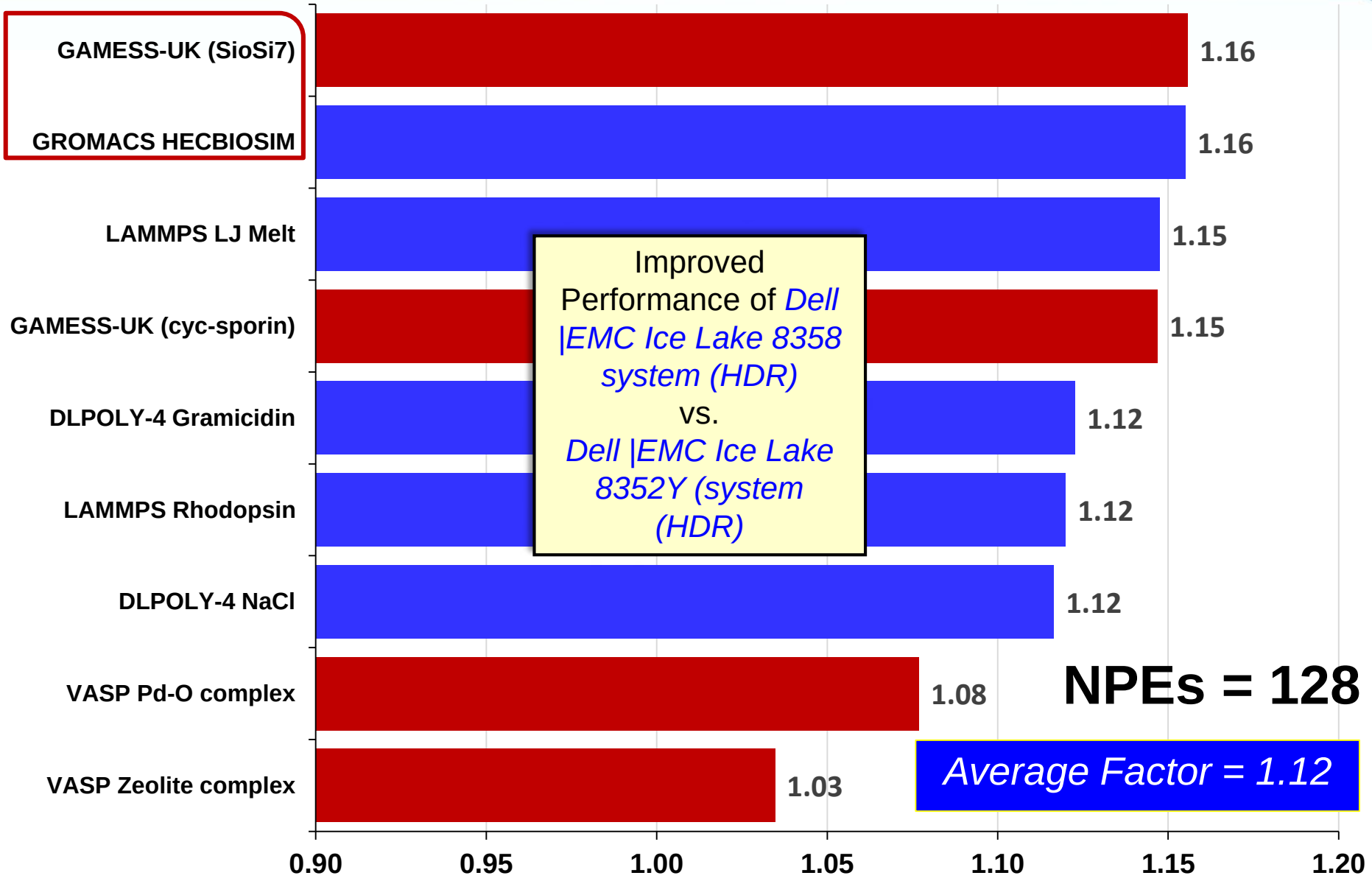
Ice Lake 8358 2.6 GHz HDR vs. SKL 6148 2.4 GHz EDR



Ice Lake 8358 2.6 GHz HDR vs. SKL 6148 2.4 GHz EDR



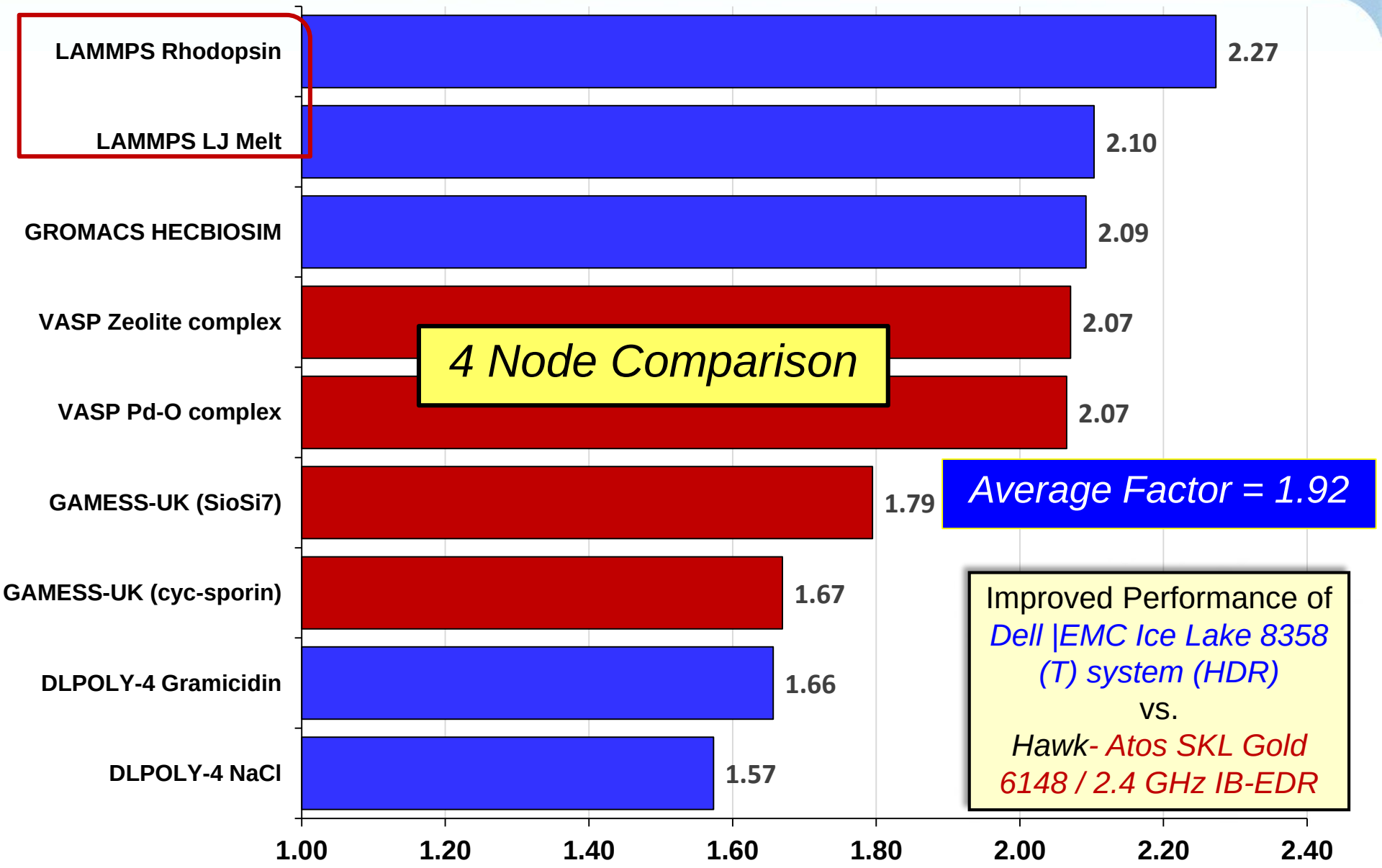
Ice Lake 8358 2.6 GHz HDR vs. Ice Lake 8352Y 2.2 GHz



Summary – Node-to-Node Comparisons

- Given superior core performance, a ***Node-to-Node comparison*** typical of the performance when running a workload shows the Ice Lake 8358 delivering **superior performance** compared to (i) the SKL Gold 6148 (64 cores vs. 40 cores) by a factor of between 1.6 – 2.3 across all applications.
- The **AMD Rome 7702 and Milan 7713 (128 cores)** along with the **Intel CSX-AP 9242** (96 cores) are the dominant systems given the "high" core counts. e,g,. GROMACS and GAMESS-UK.
- The comparison with the SMD Rome-based **Spartan 7742-based cluster** now resolved following the previous reported CIUK performance issues (attributed to Turbo mode)
- **Pricing** – remains of course a key issue, but lies outside the scope of this presentation.

Ice Lake 8358 2.6 GHz HDR vs. SKL 6148 2.4 GHz EDR



Acknowledgements

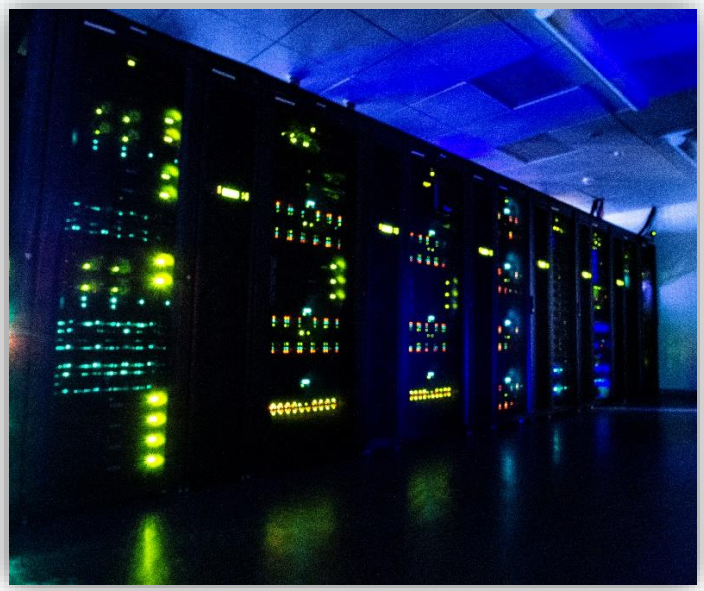
- **Joshua Weage and Joseph Stanfield**, Dave Coughlin, Derek Rattansey, and Christopher Huggins for access to, and assistance with, the variety of EPYC and Ice Lake SKUs at the Dell Benchmarking Centre.
- **Toby Smith, Ian Lloyd and Alexander White** for access to and assistance with the CXL-AP and Ice Lake clusters at the Swindon Benchmarking Lab
- ***Okba Hamitou, Luis Cebamanos and Chrisophe Bertherlot*** and access to the SPARTAN and Ice Lake & Milan systems (Genji) at the Atos HPC, AI & QLM Benchmarking Centre.
- **Alexandros Avdis, Matthew Parfitt and Joshua Short** for access to the Boston HPC centre. This single node benchmarking is to continue following the conference.

Summary

Focus here on systems featuring **current processors from AMD** (EPYC Milan SKUs – the 7713 and 7513 etc.) and **Intel** (Ice Lake & Cascade Lake-AP).

- Baseline cluster: the Skylake (SKL) **Gold 6148/2.4 GHz** and **AMD EPYC Rome 7502 2.5Gz** cluster – "Hawk" – at Cardiff University.
- Two AMD EPYC Milan clusters featuring the 64-core **7713 2.0 GHz**.
- Four Intel Xeon Ice Lake clusters, the 32-core Platinum **8358** (2.6 GHz) and **8352Y** (2.2 GHz), the 38-core **8368Q** (2.6 GHz), the 36-core **8360Y** (2.4GHz) plus other Cascade Lake and Cascade Lake-AP systems.
- Consider performance of both synthetic and **end-user applications**. Latter include molecular simulation (**DL_POLY, Gromacs & Lammps**), materials modelling (**VASP**) and electronic structure (**GAMESS-UK**).
- Scalability analysis by **processing elements (cores)** and by **nodes** (guided by ARM Performance Reports). Baselined across **P100** and **V100** NVIDIA GPU performance.
- Remains work in progress e.g., Single Node benchmarks with Boston. Please reference these in the CIUK proceedings.

Any Questions?



Martyn Guest 029-208-79319

Christine Kitchen 029-208-70455

Jose Munoz 029-208-70626

CIUK 2021 Keynote Presentation

Professor Simon McIntosh-Smith
Professor of High Performance Computing
PI for the Isambard GW4 Tier 2 HPC service
Head of the HPC Research Group
Department of Computer Science
University of Bristol

Heterogeneous Computing: past, present and future

ABSTRACT: Heterogeneous computing has been a part of HPC almost since the field began, but the current wave, which began 15 years ago, is the most significant revolution in the field since the commodity cluster. In this talk we will explore where this trend has come from, discuss its implications, and consider how our community needs to act in order to be ready for the era of heterogeneous Exascale supercomputers. We will also discuss related initiatives in the UK, such as ExCALIBUR and the UK's Exascale programme.

BIO: Simon McIntosh-Smith is Professor of HPC at the University of Bristol, UK. He began his career in industry as a microprocessor architect, first at Inmos and STMicro in the 1990s, before co-designing the world's first fully programmable GPU at Pixelfusion in 1999. In 2002 he co-founded ClearSpeed Technology where, as Director of Architecture and Applications, he co-developed the first modern many-core HPC accelerators, which lead to the creation of the first modern heterogeneous supercomputers, such as Tsubame 1.0 at Tokyo Tech in 2006. He now leads the HPC Research Group in Bristol, where his research focuses on advanced computer architectures and performance portability. He leads the Isambard supercomputer service which combines Arm-based CPUs with a diverse range of CPUs and GPUs from all the main vendors.

Prof. Simon McIntosh-Smith
Head of the HPC research group
University of Bristol, UK

Twitter: @simonmcs
Email: simonm@cs.bris.ac.uk

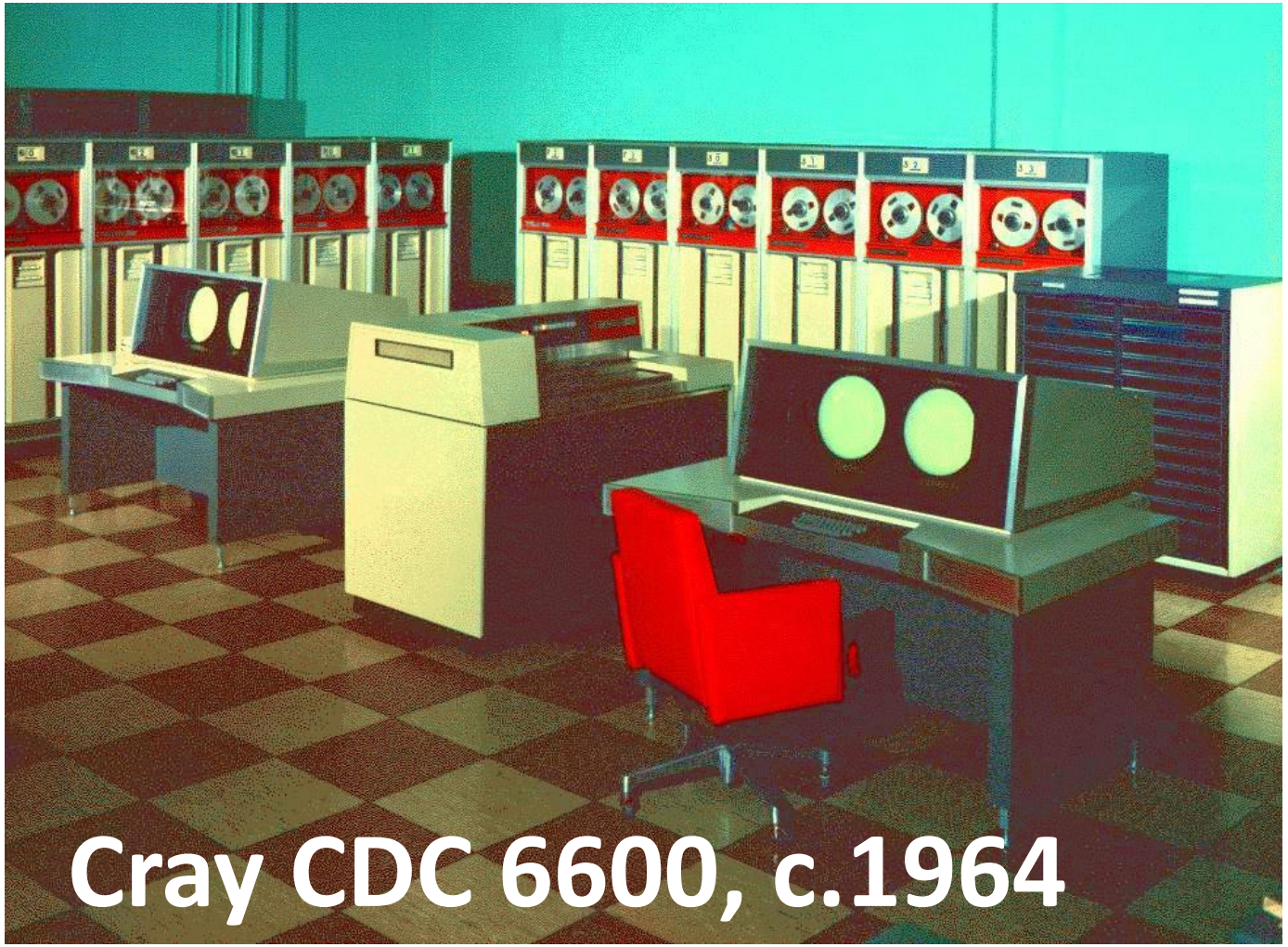
Heterogeneous Computing: past, present and future

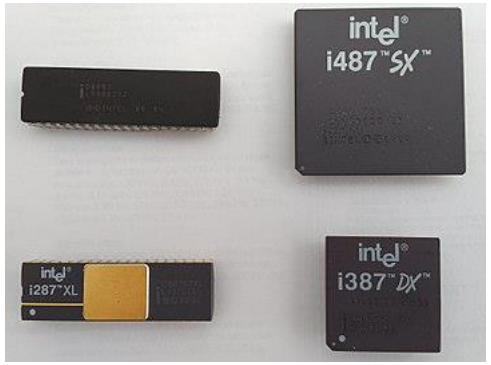
University of BRISTOL

http://uob-hpc.github.io

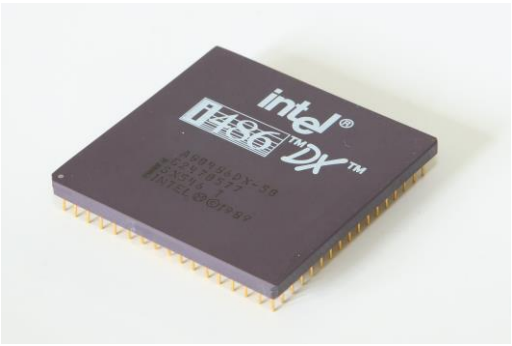
EXCALIBUR

Heterogeneity has been around since the dawn of computing



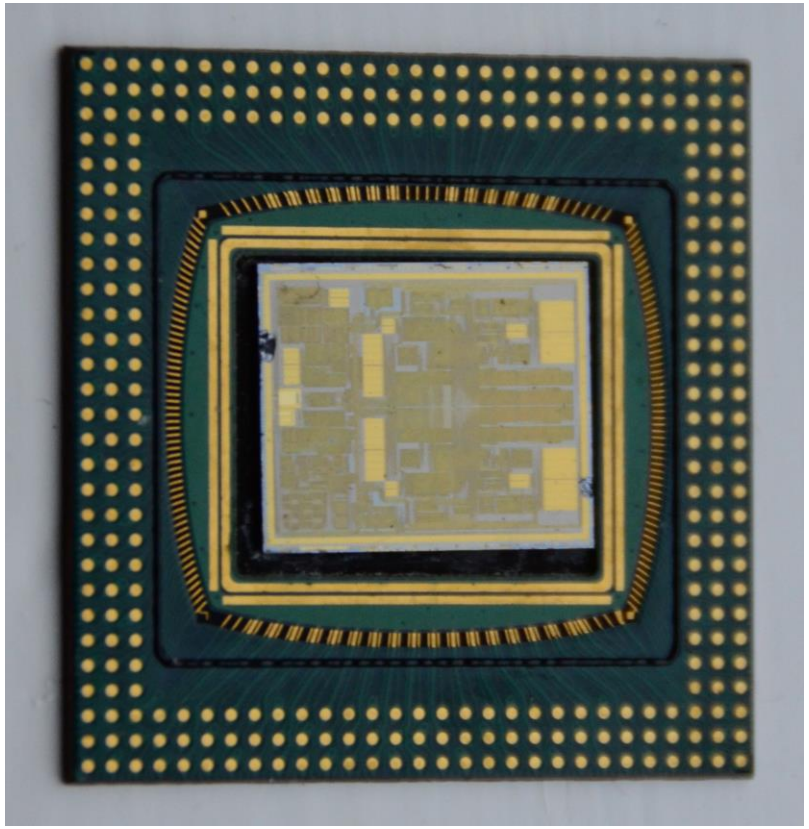


Intel floating point co-processors, 1980s

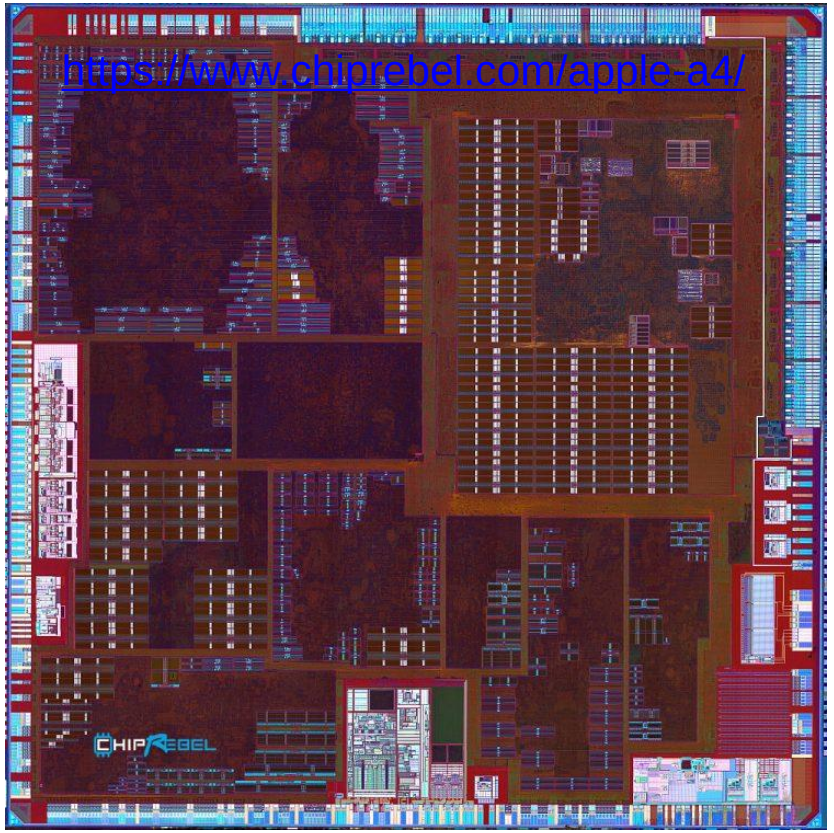


Intel 486 with integrated f.p., c.1989

Other areas of computing had been heterogeneous for years…



Inmos/STMicro “Chameleon” ST40
Dual 64-bit cores, dual issue SIMD, accelerators for video, audio c.1996.

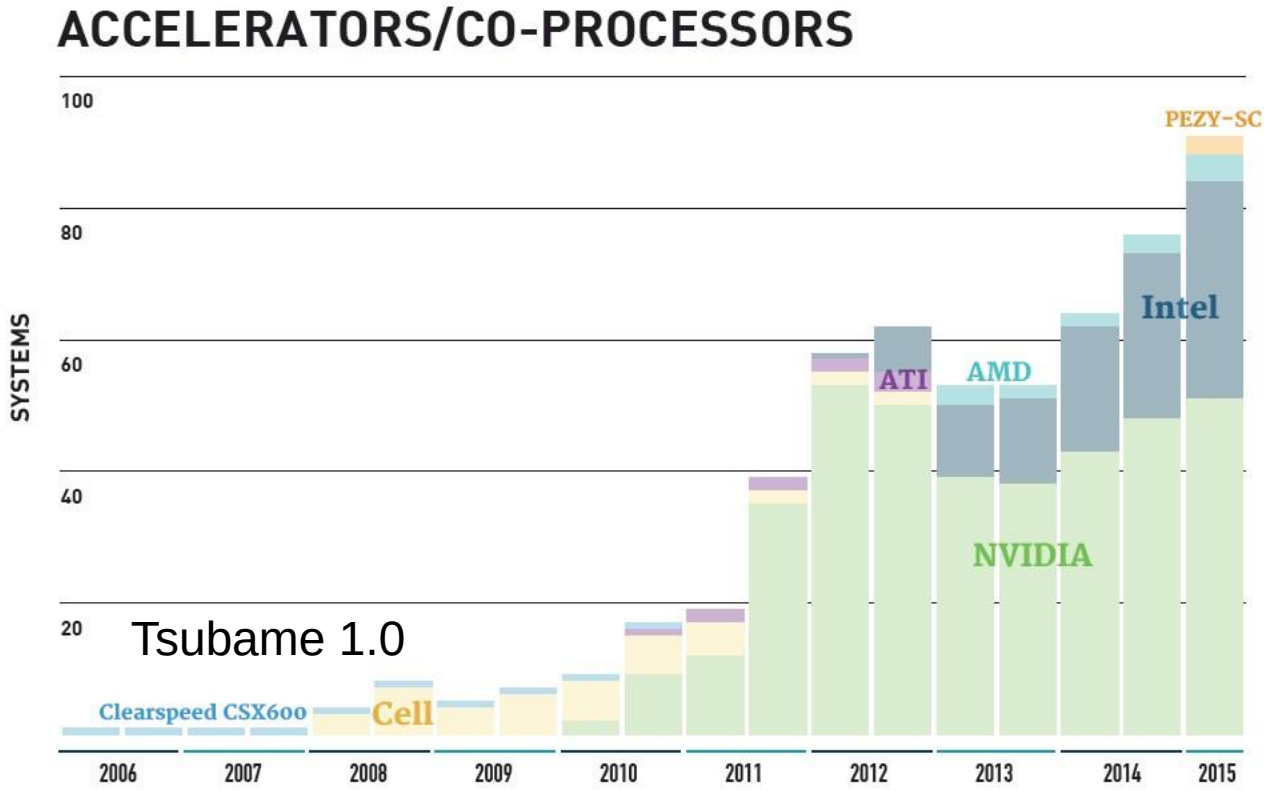


Apple A4 (first in-house design), c.2010.
iPhone 4, integrates CPU, GPU and other accelerators.

The modern era of accelerators dawned 15 years ago…

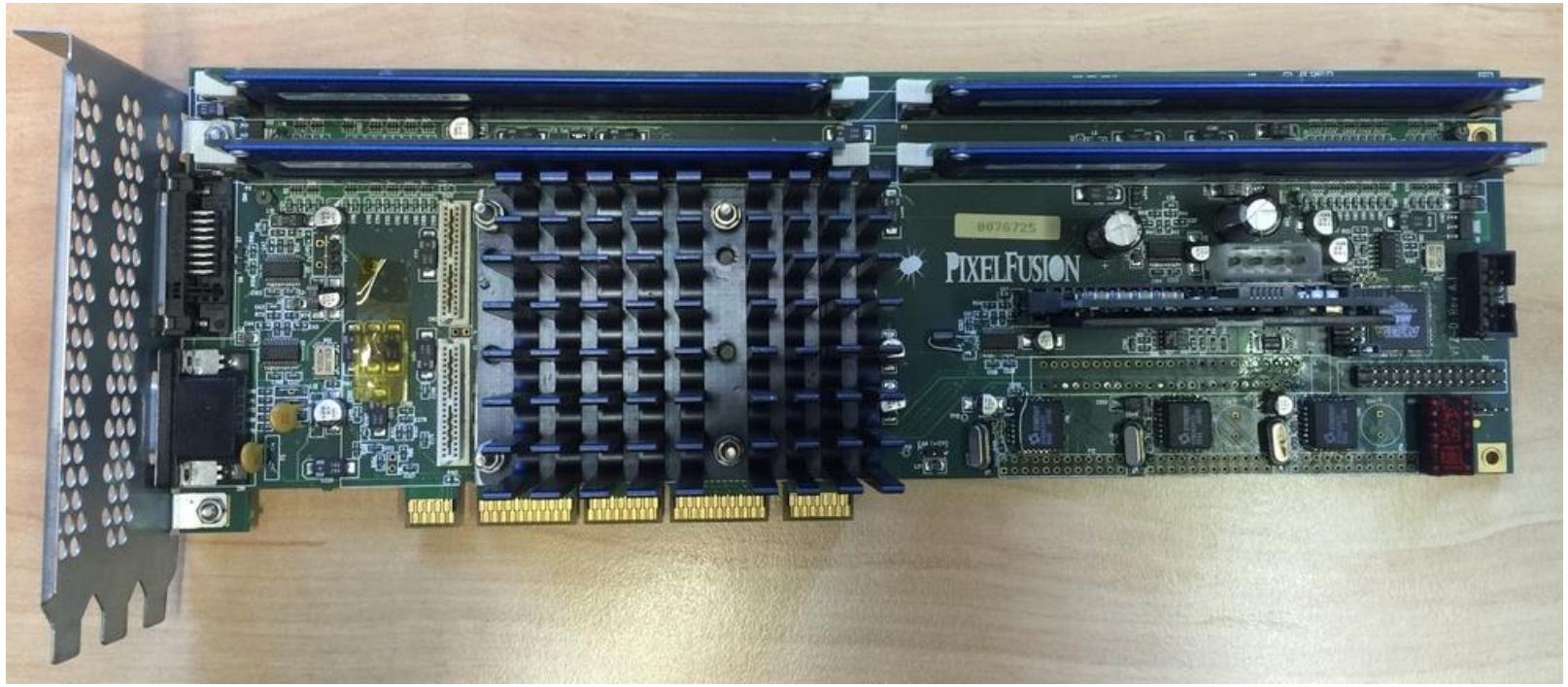
ClearSpeed e620 (2006):

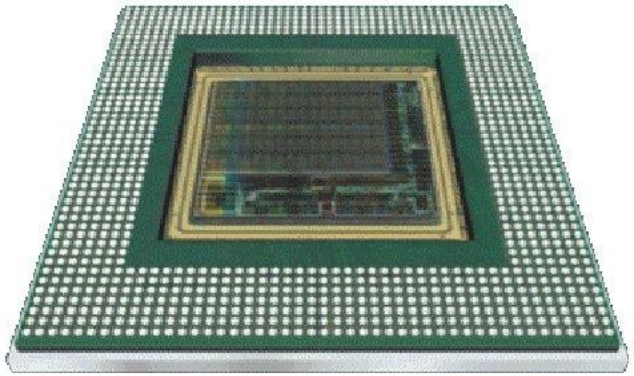
- 80 GFLOP/s 64-bit
- 1GB DRAM with ECC
- 35W



https://www.top500.org

But the seeds had had been sown at the turn of the millennium…





PixelFusion F150, developed in Bristol c.1999
Predecessor to ClearSpeed
The world's first fully programmable GPU, 1,536-way parallel SIMD.

Heterogeneous considerations

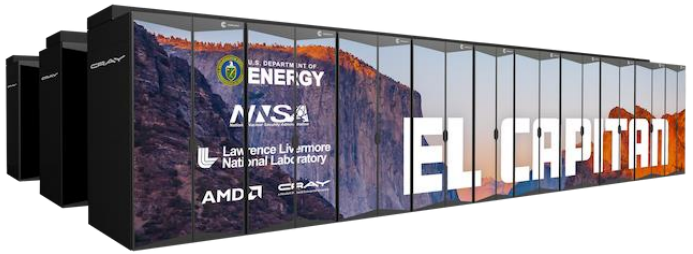
- **Motivation**: ***energy efficiency***
  - Wider, slower → more energy efficient
  - Counteract dark silicon problem

- **Drawback**: ***radically different programming models***
  - From 1 thing to program to 2 or more
  - We were used to this for graphics
  - No universal agreement on the programming model though…

Next-generation supercomputers largely heterogeneous

The coming generation of Exascale systems will include a diverse range of architectures at massive scale, most (but not all) heterogeneous:

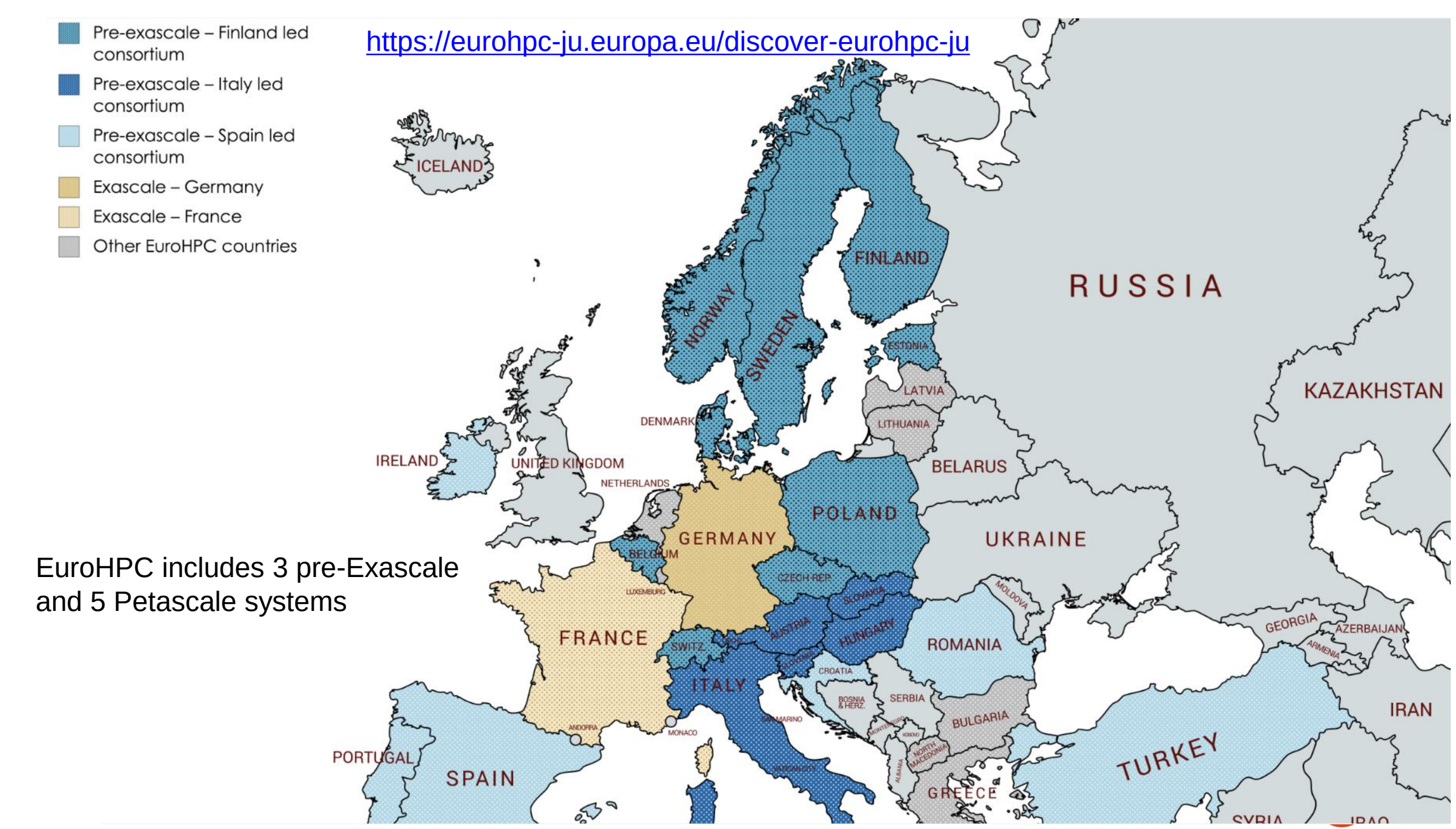
- **Fugaku**: Fujitsu A64FX Arm CPUs
- **Perlmutter**: AMD EYPC CPUs and NVIDIA GPUs
- **Frontier**: AMD EPYC CPUs and Radeon GPUs
- **Aurora**: Intel Xeon CPUs and Xe GPUs
- **El Capitan**: AMD EPYC CPUs and Radeon GPUs



University of BRISTOL

The Next Platform, Jan 13th 2020: "HPC in 2020: compute engine diversity gets real"
https://www.nextplatform.com/2020/01/13/hpc-in-2020-compute-engine-diversity-gets-real/

https://eurohpc-ju.europa.eu/discover-eurohpc-ju



EuroHPC includes 3 pre-Exascale and 5 Petascale systems

The UK's HPC service ecosystem is intentionally diverse

University of BRISTOL

https://epsrc.ukri.org/research/facilities/hpc/tier-2-hpc-centres/

ExCALIBUR

Isambard 2

- **21,504** Armv8 cores (336n x 2s x 32c)
  - **Marvell ThunderX2 32 core @2.5GHz**
- 3,456 core Fujitsu A64fx system
- >600 registered users
- Includes a “Multi Architecture Comparison System (MACS)”
  - Includes interesting CPUs and GPUs:
    - AMD Rome, Intel Cascade Lakes, IBM POWER9
    - NVIDIA V100 and P100 GPUs
    - Currently adding AMD Milan, Intel Icelake, NVIDIA A100 GPUs and AMD Mi100 GPUs

University of BRISTOL

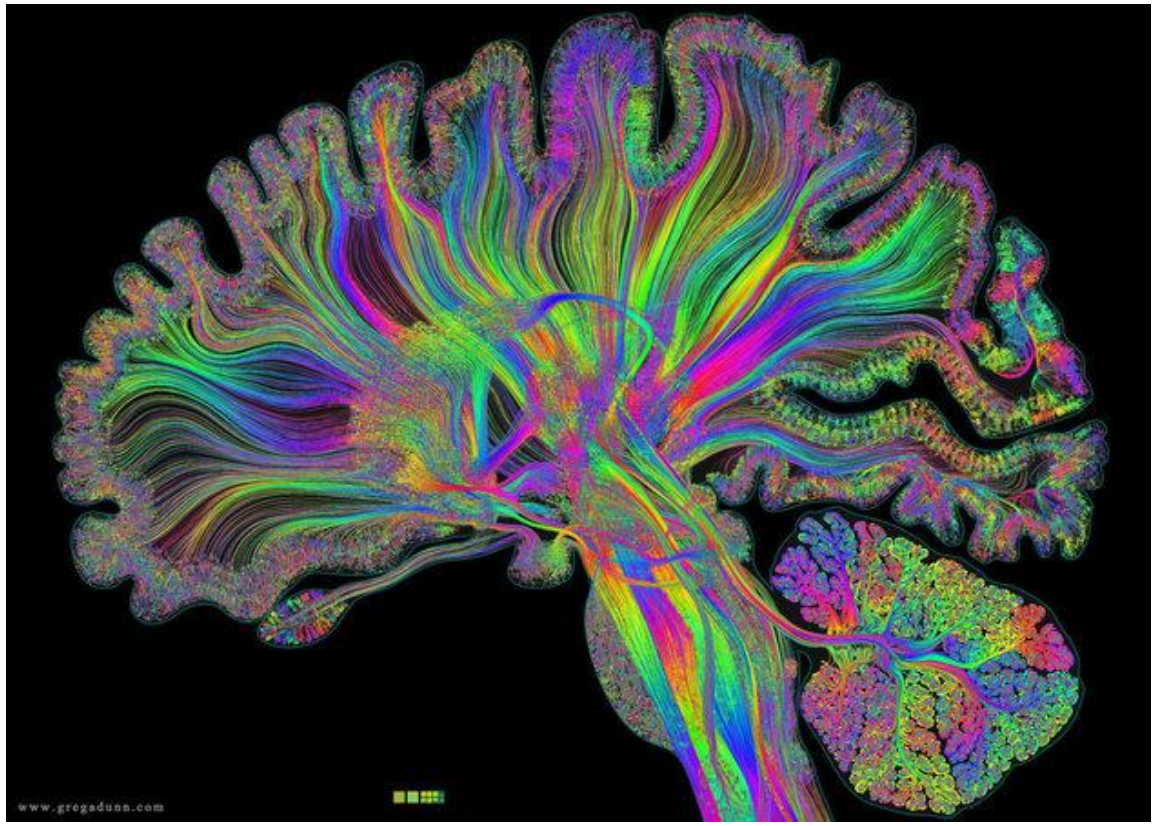
A performance analysis of the first generation of HPC-optimized Arm processors, S. McIntosh-Smith, J. Price, T. Deakin & A. Poenaru, CC:PE, Feb 2019. https://doi.org/10.1002/cpe.5110

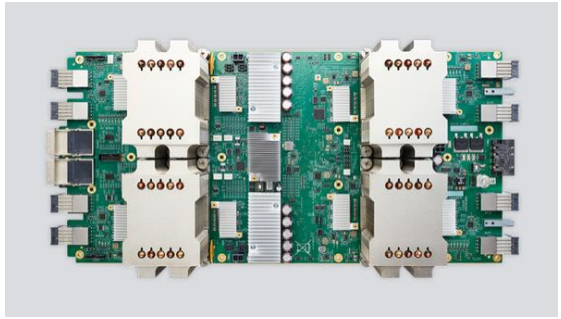
ExCALIBUR programme in the UK tackling the challenge

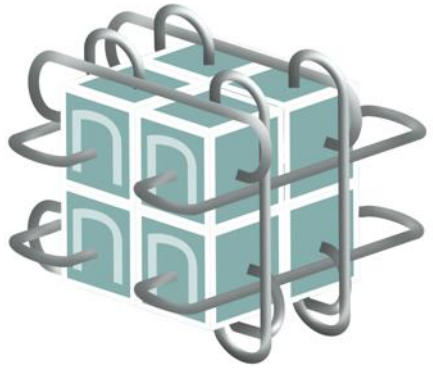
- See talks earlier today from Dr Elizabeth Bent (UKRI), Dr Rob Akers (UKAEA) and Dr Nigel Wood (Met Office)

- £46M over 5 years, focused on getting the UK's science codes Exascale ready
- Addressing heterogeneity by multiple routes:
  - Separation of concerns
  - Domain Specific Languages (DSLs)
  - Performance portable programming languages, etc.

Emerging architectures for AI / Machine Learning

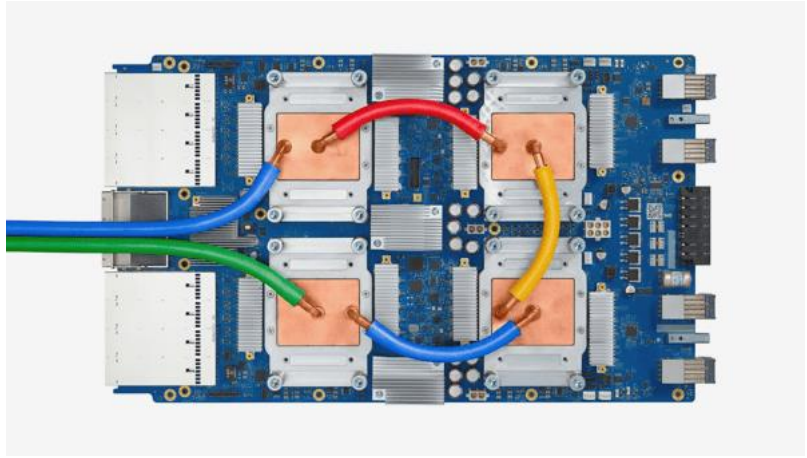






Google’s Tensorflow Processing Unit (TPU), GraphCore, Intel’s Nervana

Google's Tensor Processing Units



Cloud TPU v3:
420 TFLOP/s
128 GB HBM
$2.40 / TPU hour

V4 supposedly improves performance by 2.7x

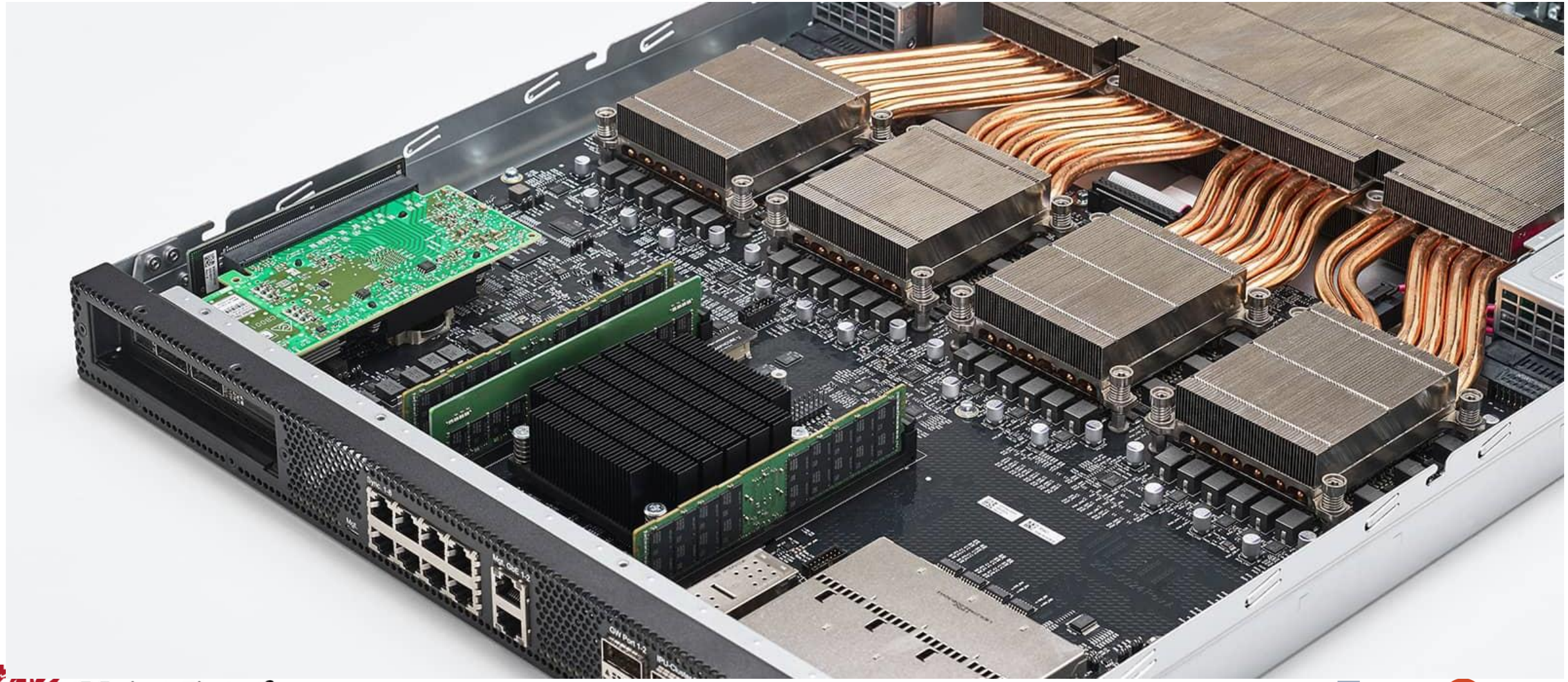
Cloud TPU v3 Pod:
100+ PFLOP/s
32 TB HBM
2-D toroidal mesh network

https://cloud.google.com/tpu

University of BRISTOL

ExCALIBUR

Graphcore is already onto their 2nd generation “IPU”

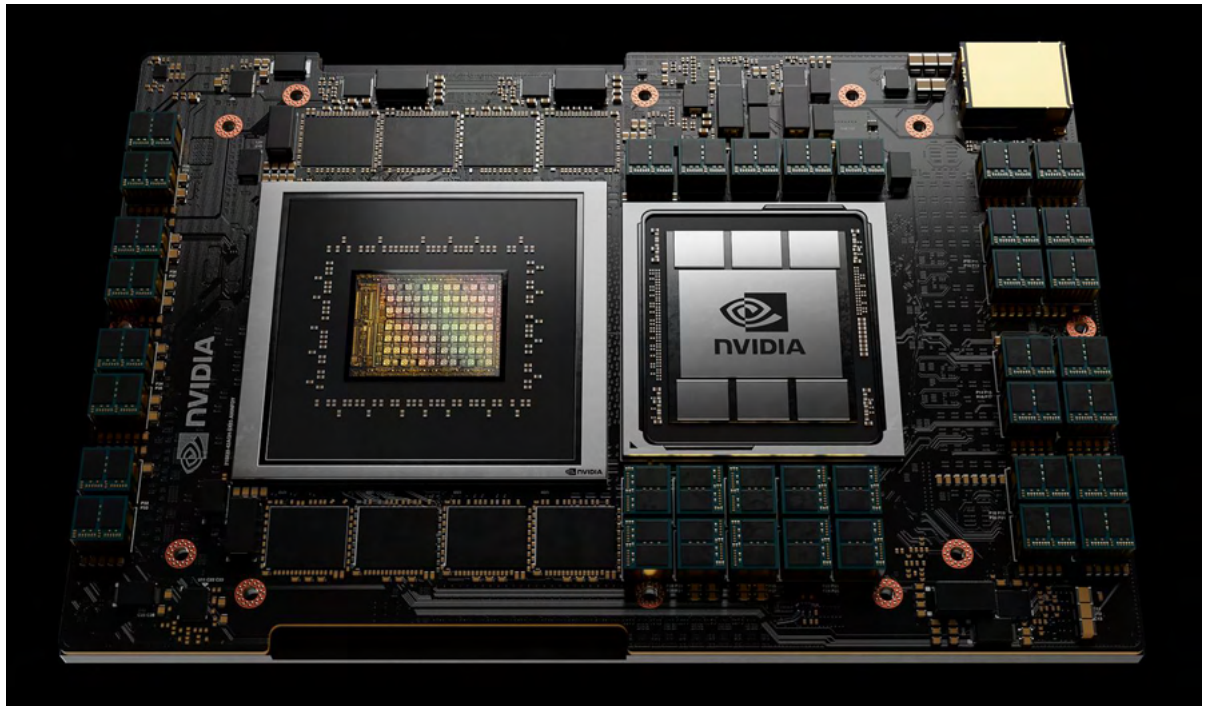


Graphcore IPU-M2000

- 4 x Colossus MK2 GC200 IPUs in a 1U box
- 1 PetaFLOP “AI compute” (16-bit FP)
- 5,888 processor cores, 35,328 independent threads
- Up to 450 GB of exchange memory (off-chip DRAM)

- 59.4B 7nm transistors in 823mm^2
- 900MB of on-chip fast SRAM per IPU (3x first gen.)
- 250 TFLOP/s AI compute per chip, 62.5 TFLOP/s single-precision

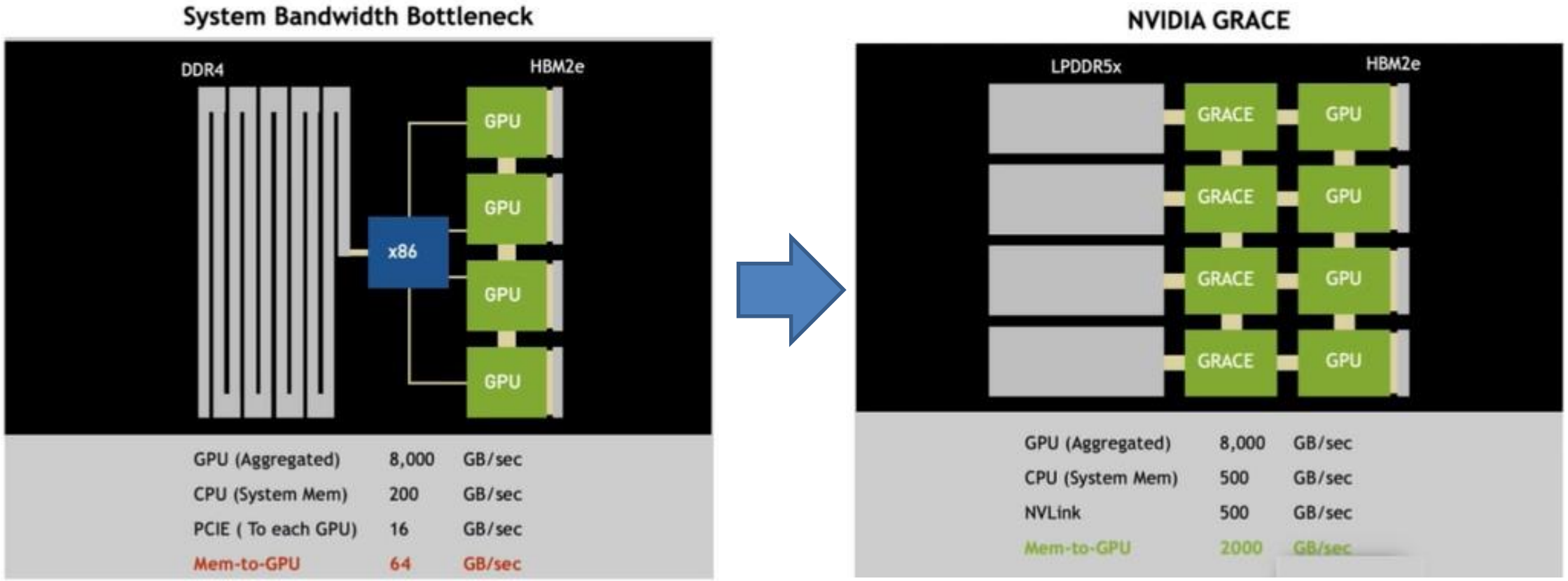
CPUs and GPUs becoming more tightly coupled



- NVIDIA announced their own Arm CPUs, "**Grace**"
- 900GB/s interconnect between the CPU and GPU
  - >10x fastest PCIe
- Very high memory bandwidth for a CPU
  - >500GB/s
- Shipping 2023 with their next-gen GPUs

https://www.nextplatform.com/2021/04/12/nvidia-enters-the-arms-race-with-homegrown-grace-cpus/

More balanced intra-node interconnects



Three of the big issues facing parallel programming

1. **Massive parallelism**
   - Fugaku has over 7.63 **million** cores, each with 2x 512-bit wide vectors

2. **Heterogeneity**
   - CPUs, GPUs and more, from multiple vendors
     - Intel, AMD, NVIDIA, Fujitsu, IBM, Amazon, Google, …
   - Non traditional architectures
     - Graphcore IPUs, Google TPUs, Cerebras, vector engines, FPGAs, …

3. **Managing complex memory hierarchies**

Are heterogeneous systems going to become more diverse?

“Not necessarily, mon amis!

- Hercule Poirot

Why not?

Prof. Satoshi Matsuoka argued strongly for **limiting the diversity of heterogeneity** at SC21 in the panel session "Heterogeneity in Hardware: Opportunities and Challenges for Software and Applications": https://sc21.hubb.me/fe/schedule-builder/sessions/877000. The relevant part of the panel starts 34 minutes in.

Why can't we just have lots of different accelerators for everything?

Because of **software developer productivity**, and because of **scaling and load balance**

1. Developing heterogeneous codes significantly increases the burden on software developers, and thus the cost of developing software.
2. Scaling and load balancing across a system with one type of accelerator already hard enough; with more than one, likely intractable.

Satoshi Matsuoka
@ProfMatsuoka

Although diverse heterogeneity is good for smartphones, during the Tue #SC21 panel "Heterogeneity in Hardware: Opportunities and Challenges for Software and Applications" & my invited talk Thu morning plenary I will controversially present why that is a BAD idea for modern HPC.

10:30 PM · Nov 14, 2021 · Twitter

7 Retweets **1** Quote Tweet **4**

Satoshi Matsuoka @ProfMatsuoka · Nov 14

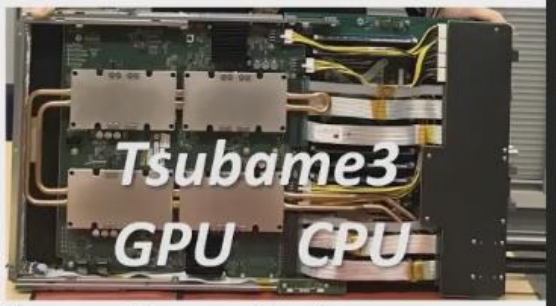
Replying to @ProfMatsuoka

... and you will find out why machines like Tsubame, Fugaku, Summit are successful while those extrapolating that we have multitudes of heterogeneous customized accelerator components will not be.

1 3 13

Accelerators vs. Amdahl's Law & Gustafson's Law (3)

- **It is no accident that, every successful large-scale accelerated supercomputers (esp. GPU machines) are**
  - built with a singular node configuration across the entire machine
  - tight coupling and robust interconnect (& I/O) to sustain maximum bandwidth in/out of accelerator processor
  - dominant processing on the GPU for maximum performance
  - SPMD with very good load balancing (incl. data parallel DNN training)



- Tsubame, Tianhe-2A, Titan/Summit, Piz-Daint, ABCI, Fugaku, Frontier, Lumi, Aurora, …
- … and this is the consequence of physical laws, so will continue to be applicable to future machines (no extreme heterogeneity, asynchrony, …)

Possible alternatives to heterogeneous systems

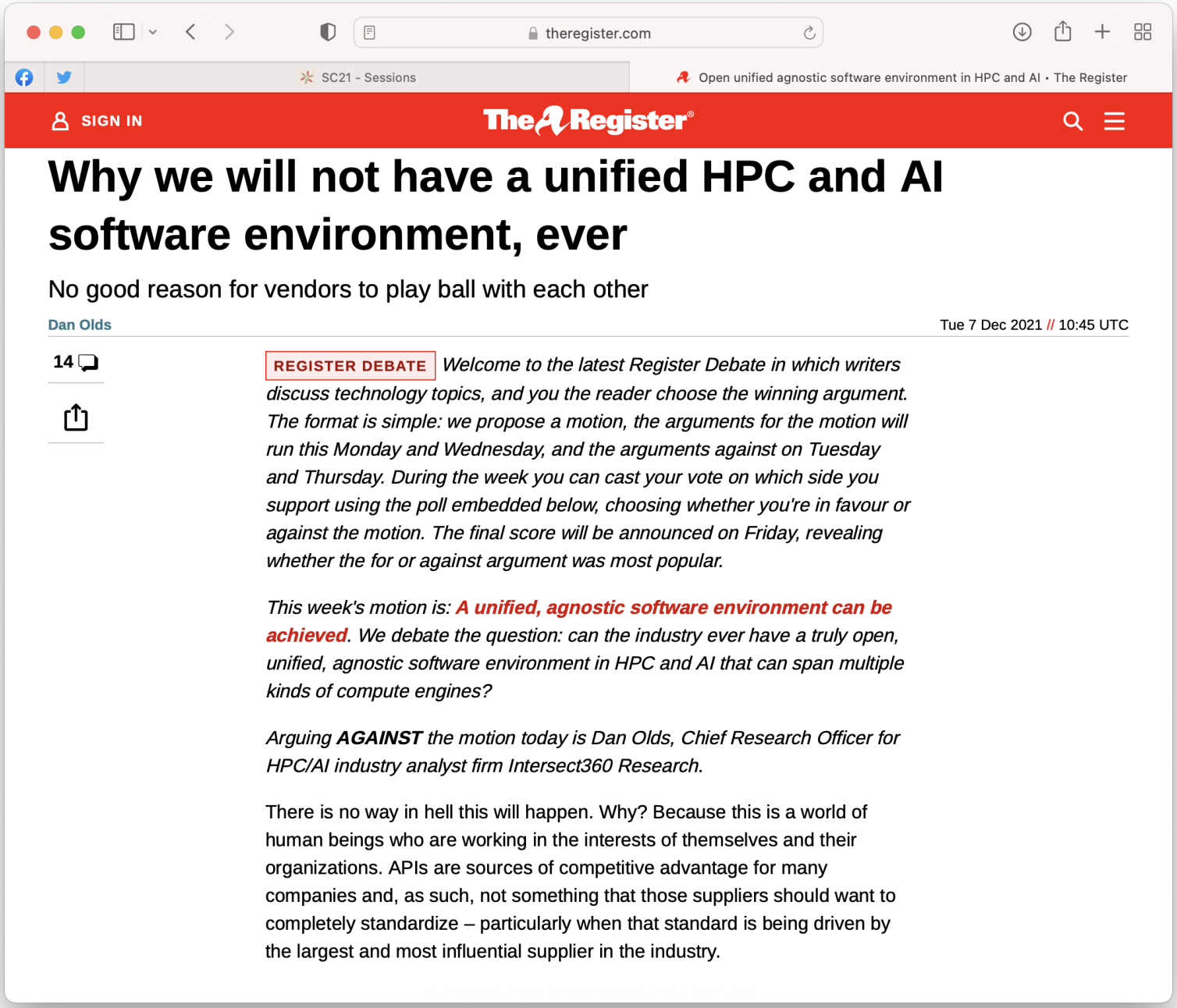
- What are the most valuable features of GPUs?
  - High FLOP/s per Watt, high peak FLOP/s, high bandwidth memory, high FLOP/s per Dollar, latency tolerance, …?
- Could these features be integrated into, for example, CPUs?
  - Yes, mostly – see A64fx, Sapphire Rapids HBM, SiPearl Rhea, …
- What are the main drawbacks?
  - Hard to achieve the same peak FLOP/s per Watt without heterogeneity, however, performance on real codes is a different story…

Promising heterogeneous computing developments

- Heterogeneous-aware software stacks (ECP's etc)
- Cross-platform programming standards
  - ISO C++ / Fortran, OpenMP, SYCL, Kokkos, …
  - See talk by Jeff Hammond and Filippo Spiga on Friday 10th at 10am
- Rate of adoption of heterogeneous systems at the high end
  - All the top 10 systems are heterogeneous, except Fugaku and Sunway TaihuLight
  - 150 systems in the Nov 2021 Top500 now use accelerators (mostly NVIDIA GPUs) – that's 30% of the total list

The future of heterogeneous computing…

- Likely we'll rely on heterogeneous systems for quite some time
  - But modestly diverse, not extremely, diverse
- Closer integration between CPUs and GPUs (cache coherent, in-package etc)
- CPUs will keep integrating some of the best parts from GPUs
  - HBM, lots of cores with wide vectors, in-core accelerators (matrix etc)
- The programming situation is improving, but still has a long way to go
- The long tail of users and codes will not (and should not) go away



University of BRISTOL

https://www.theregister.com/Debates/2021/12/06/software_environment/

ExCALIBUR 10

Key takeaways

- **Increased heterogeneity** is the order of the day (esp. for Exascale)
  + Better peak FLOP/s per Watt
  ? Performance per Dollar on real codes
  - Harder to program, reduced developer productivity
  - Lack of vendor agreement on programming models a severe hindrance
- Some of the best features of heterogeneous systems are being integrated into homogeneous ones
- We have a very long tail of developers running thousands of different, complex, continually evolving codes, in dozens of different programming languages and parallelism models → most of this may never be ported to accelerators

For more information

Bristol HPC group: https://uob-hpc.github.io/

Email: S.McIntosh-Smith@bristol.ac.uk

Twitter: @simonmcs

ExCALIBUR: https://excalibur.ac.uk

Isambard: https://gw4-isambard.github.io/

Professor Mark Parsons (EPCC Director at The University of Edinburgh / EPSRC Director of Research Computing)

The UK Exascale Project

Abstract: The Exascale era is upon us. China is already operating two Exascale systems and the first Exascale system in the USA – Frontier – will go live soon. The UK Government has committed to have a UKRI Exascale supercomputer in operation by 2025. This talk will summarise the UK Exascale Project with insights into the hosting of such a system and what the system is likely to look like from a technical standpoint. The Exascale system will enable the UK's internationally respected computational science community to cement the UK as a global science and technology superpower.

Bio: Mark Parsons is Director of EPCC, the supercomputing centre at the University of Edinburgh. He has a Personal Chair in High Performance Computing at the University and has worked for EPCC since 1994 following his PhD in Particle Physics at CERN in Geneva. In addition to being in charge of the most of the UK's national supercomputers at EPCC's data centre outside Edinburgh, he also has responsibility for the Edinburgh International Data Facility, a key part of the Edinburgh & SE Scotland City Region Deal. One day per week he works for UK Research & Innovation as EPSRC's Director of Research Computing. He is best known internationally, particularly in the Europe, for his work with the industrial applications of HPC with a specific focus on SMEs. He also has had the honour in 2021 of being the Chair of the ACM Gordon Bell Prize committee.

UK Research and Innovation

The UK Exascale Supercomputer Project

Professor Mark Parsons
EPSRC Director of Research Computing

10th December 2021

UK Research and Innovation

THE UK EXASCALE SUPERCOMPUTER PROJECT

Professor Mark Parsons

EPCC Director
Dean of Research Computing

epcc

History of the project

- In 2017 establishment of EuroHPC was announced at 60th Anniversary of Treaty of Rome celebrations in Rome
- Towards end of 2018, UK declined to join EuroHPC and relinquished its “observer” status on EuroHPC Governing Board
- Exascale Project Working Group set up in late 2018 to develop Outline Business Case for Government
  - Draft OBC first completed in late 2019
  - In parallel Supercomputing Science Case completed and published
- **Since 2020 has moved into UKRI as a cross-Research Council development project within DRI Programme**

UK Research and Innovation

Exascale Requirements from Government

- System should support both **traditional Modelling & Simulation** and **Artificial Intelligence / Deep Learning** applications
  - Technology choices may be impacted by this
  - But future technologies blur the distinction
- System should support both **scientific user communities** and **industry users**
  - A greater focus is proposed with regard to industry use for research
  - Pay-per-use production access will be supported
  - Specific support for SMEs
- System should be **operational around time of EU systems - 2024**

The Exascale era – worldwide progress

Country or Region	Timescale	Detail
Japan 	2020	Fugaku : based on Fujitsu A64FX Arm processo
China 	2021	Two systems in operation - next generation Sunway and Tianhe 3 system. Third system delayed.
USA 	2021 2022	Frontier : based on AMD EPYC CPU + AMD GPU Aurora : Intel Sapphire Rapids CPU + Intel Ponte Vecchio GPU
Europe 	2021/2 2023/4	Pre-Exascale systems in Finland / Italy + possibly Spain Two Exascale systems in 2024

41 million cores!

UK Research and Innovation

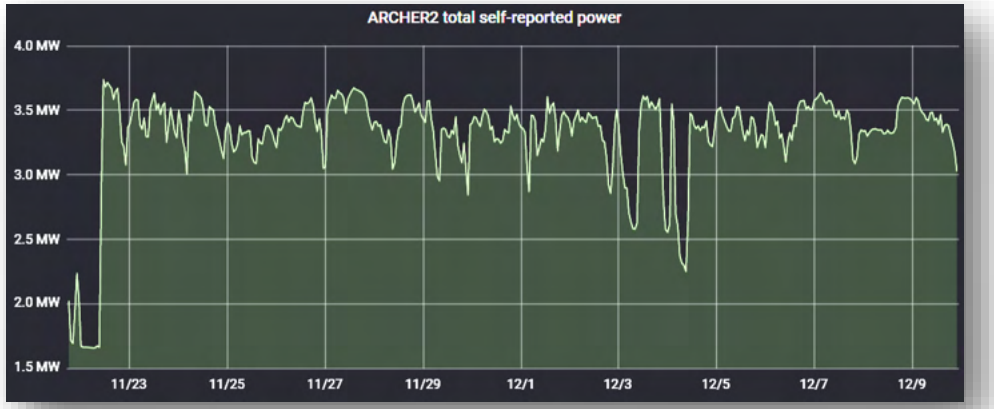
epcc

… Fugaku wears the crown

- Fugaku became the world's fastest supercomputer in June 2020 with a cores-only approach based on the Fujitsu A64FX Arm CPU
- Processor developed in long-term co-design (10 years) with Japanese computational science community led by Riken CCS
- 7,630,848 Arm CPU cores
- R_{peak} = 573.2 Petaflop/s
- R_{max} = 442.0 Petaflop/s
- Power = 29.9 MW
- Single precision > 1 Exaflop

… and ARCHER2 is finally here

- The 23 cabinet system finally opened for all users on 22nd November
- Very difficult 18 months
- Performance of the system is now good – we hope users agree
- Busy from Day 1 – and has remained busy



UK Research and Innovation

Exascale in the EU

- EuroHPC Joint Undertaking established to co-fund Pre-Exascale and Exascale systems with Member States
  - Long-term plan – including development of EU processor by SiPEARL
  - Funding of €7billion from 2021-2027
- Three sites chosen for pre-Exascale systems in 2019 – Finland (CSC), Italy (CINECA) and Spain (BSC)
- Two pre-Exascale systems procured for Finland and Italy
  - Spanish procurement is being re-run
- Exascale systems planned for 2024/25
  - Hosting locations likely to be Germany and France

UK Research and Innovation

Recent EuroHPC announcements

- Finland (CSC) is hosting Lumi
  - 375 Petaflops (HPL) / 550 Petaflops (Peak)
  - €145 million
  - Supplied by HPE
  - AMD EPYC CPUs + AMD GPUs

- Italy (CINECA) will host Leonardo
  - 249 Petaflops (HPL) / 324 Petaflops (Peak)
  - €120 million
  - Supplied by ATOS
  - Intel Icelake CPUs + NVIDIA A100 GPUs

First racks installed

Delayed

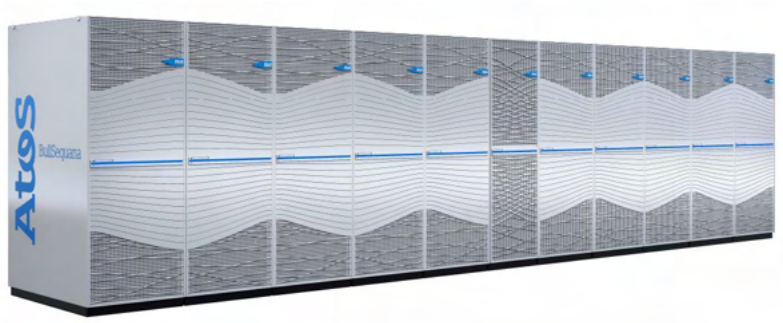
UK Research and Innovation

More detail on Lumi

- HPE Cray EX system
  - Same platform as ARCHER2
- GPU partition
  - 2,560 nodes – 1 AMD Trento CPU + 4x AMD MI250X GPUs
  - 10,240 GPUs and 16,384 cores
- CPU partition
  - 1,536 nodes – 2x AMD Trento CPUs
  - 196,608 cores
- 375 PFlops (HPL) / 550 PFlops (Peak)

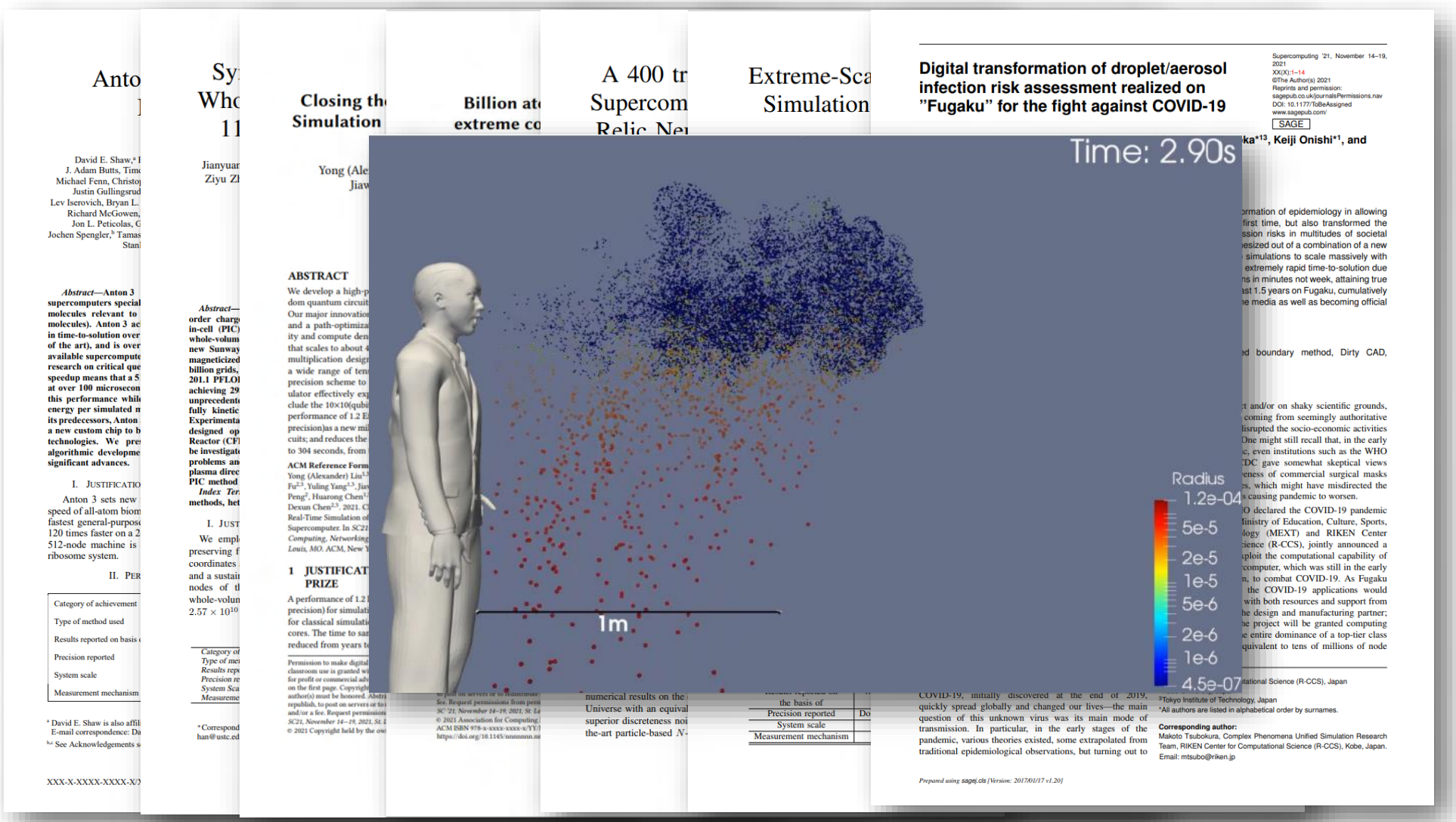
More detail on Leonardo

- ATOS BullSequana system
  - Two partitions – “Booster” and “Data Centric”
- GPU partition (Booster) – 3,456 nodes
  - 221,184 cores - Intel Icelake CPUs
  - 13,824 NVIDIA A100 GPUs
- CPU partition (Data Centric) – 1,536 nodes
  - 79,872 cores - Intel Sapphire Rapids CPUs
  - Local NVM (DCPMM?) for data analysis
- 249 PFlops (HPL) / 324 PFlops (Peak)



UK Research and Innovation

Scientific impact

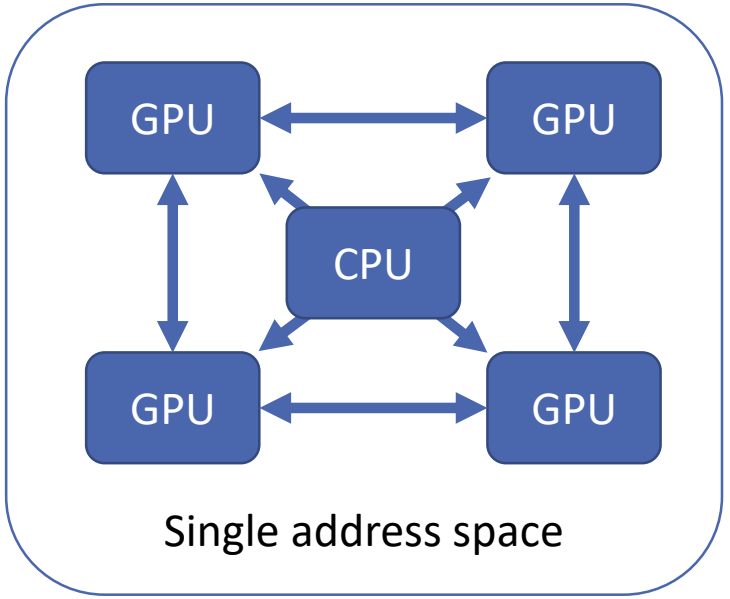


Provide the capability and scientists will use it

UK Research and Innovation

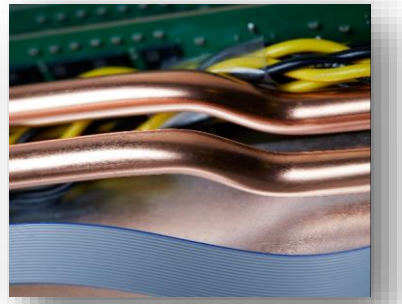
Technology –recent Exascale vendor briefings

- Memory is changing
  - Many Exascale blades include HBM
    - Some designs have no DRAM at all
  - But recently LPDDR5 is being mentioned more
- Four-way competition for CPUs and/or GPUs
  - Intel versus AMD versus Arm versus NVIDIA
- GPUs market is broadening
  - AMD is strongly competing with NVIDIA
- Cabinet energy densities are rocketing
  - Today's 80-100KW cabinets will be eclipsed by cabinets at 300KW+
- Multicore CPUs are also getting AI Deep Learning features



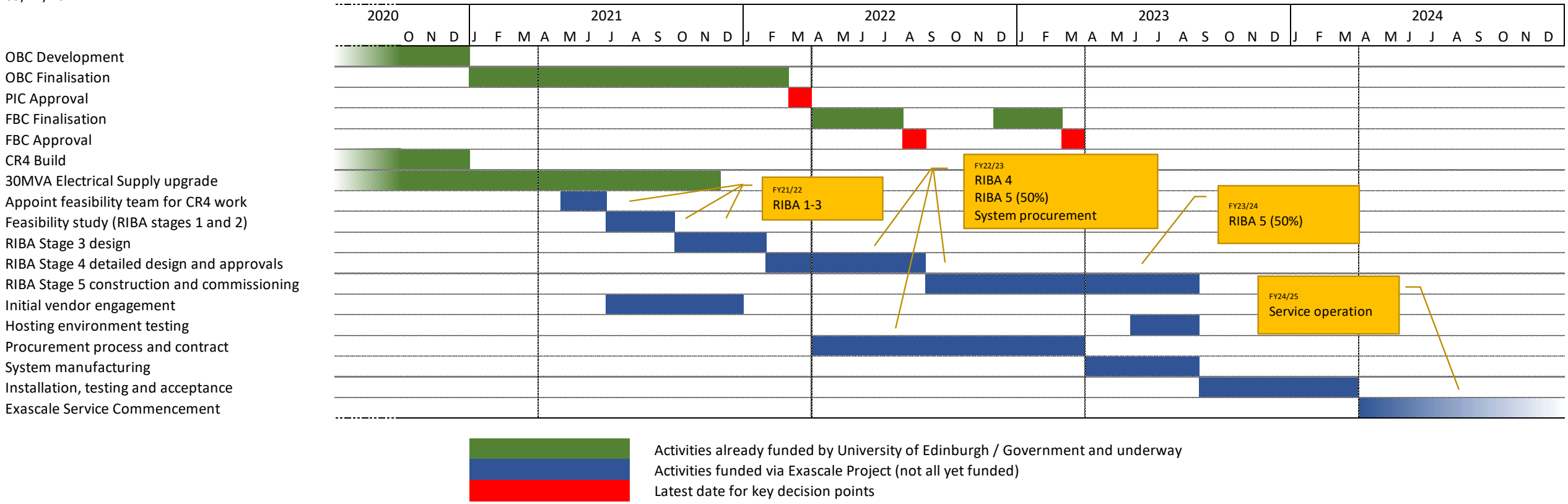
General design principles for UK Exascale Project

- 25MW system + 5MW support and cooling
- Single tightly coupled system
- Main compute power from GPU partition
  - Target 1 Exaflop/s R_{MAX}
- Remainder of space or power budget for CPU partition
  - Designed to provide attractive powerful resource for non-accelerated codes as they transition
- Large Software Programme envisaged
  - Multiple activities – Grand Challenge based to eCSE type activities
  - Lots of requirements gathering / consultation to do



Project timeline

UK Exascale Supercomputer Timeline
09/12/2021



Activities already funded by University of Edinburgh / Government and underway
Activities funded via Exascale Project (not all yet funded)
Latest date for key decision points

Entirely dependent on funding and UKRI prioritisation

UK Research and Innovation

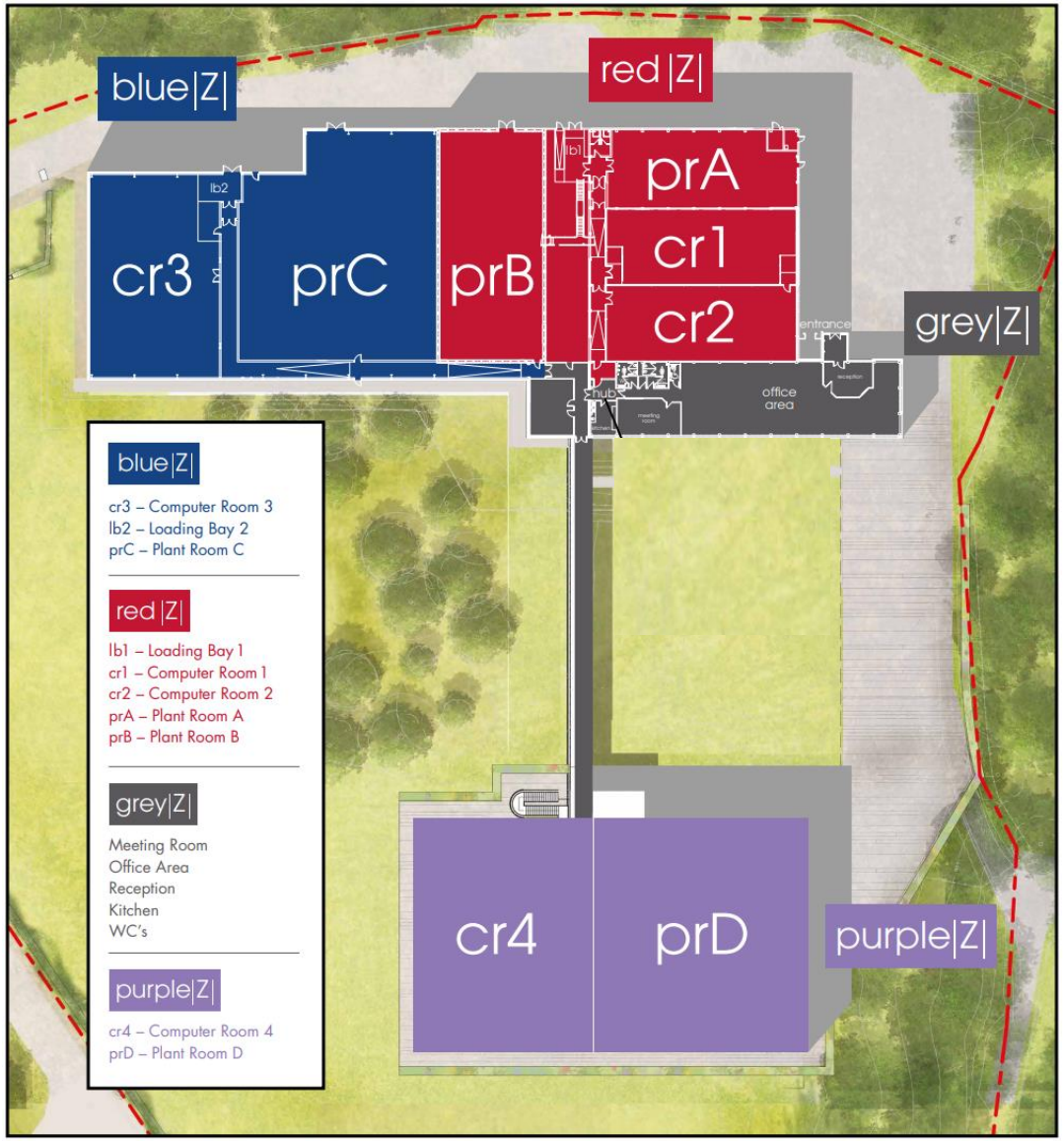
System Hosting and Outline Designs

Professor Mark Parsons

EPCC Director
Dean of Research Computing

epcc

EPCC's Advanced Computing Facility Data Centre



- Plant Room A and Computer Rooms 1&2 date back to 1970s
- Plant Room B added for HECToR
- Computer Room 3 and Plant Room C added for ARCHER – 4MW capability
- Computer Room 4 and Plant Room D added in 2020 – current configuration 6MW

epcc

Computer Room 4

£20m – CR 4 + PR D
£8.6m – 30MVA additional power
Space for 270 standard racks

Opened Dec 2020

epcc

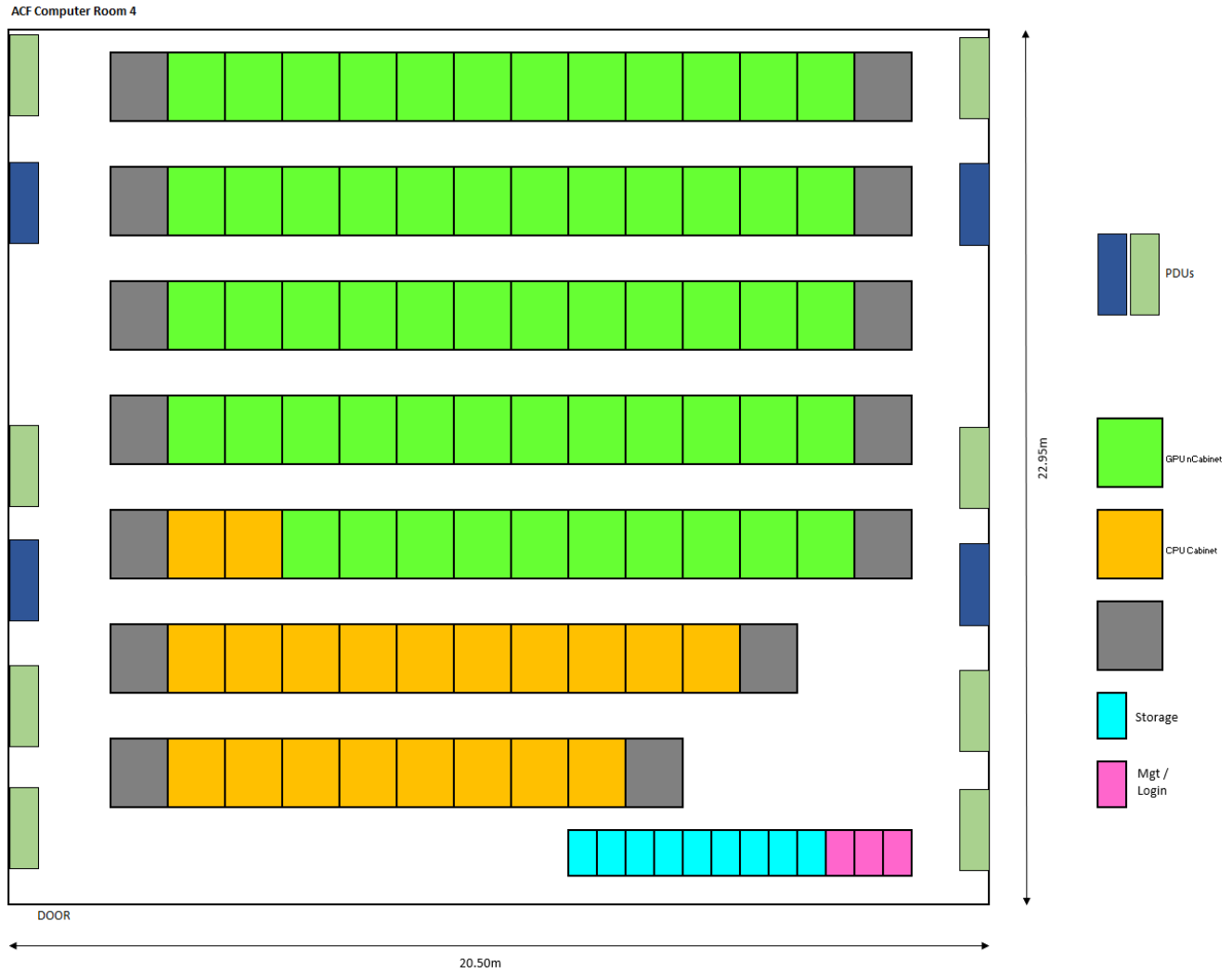
Preparing the ACF for the Exascale

- EPCC's … to … by 2024
- Very con… es…
- **New £20… 4…**
- **New £9r… ri… be live** by Easte…
- CR4 is a… m… ical and electrica… E…
- Process… y… y of Edinburg… dy… FI
- **Preparations must be complete by Q2 2023**

epcc

Example from RFI responses (obfuscated)

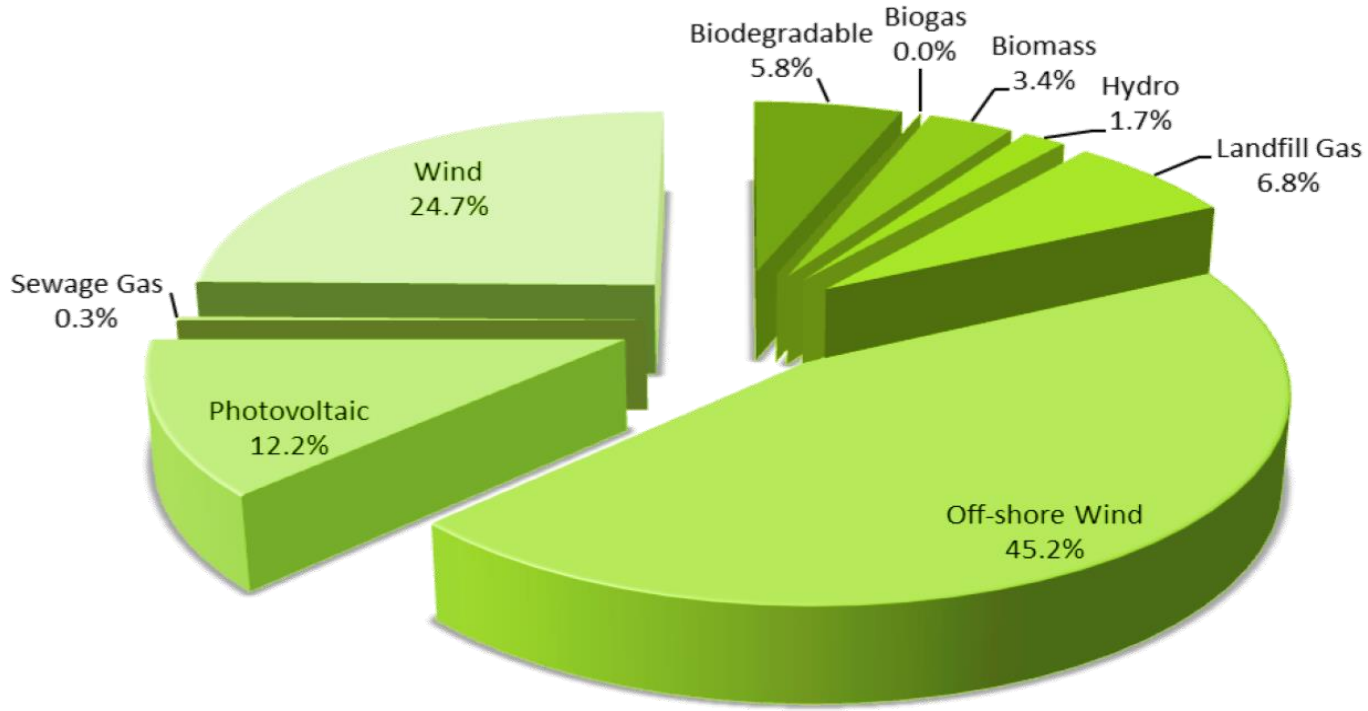
- Vendors asked to produce designs up to 25MVA
- Combination of
  - 1 ExaFlop HPL R_{max}
  - Cores-only partition
- Dual approach provides route from cores-only world to accelerated world
- GPU Partition
  - **24,000 GPUs**
  - **380,000 cores**
  - 60 racks to reach 1 ExaFlop HPL – 19MVA power
- CPU Partition
  - **1,000,000 cores**
  - 20 racks – 6MVA power (limit reached)
- Plus
  - 100PB storage system
  - Login and service nodes



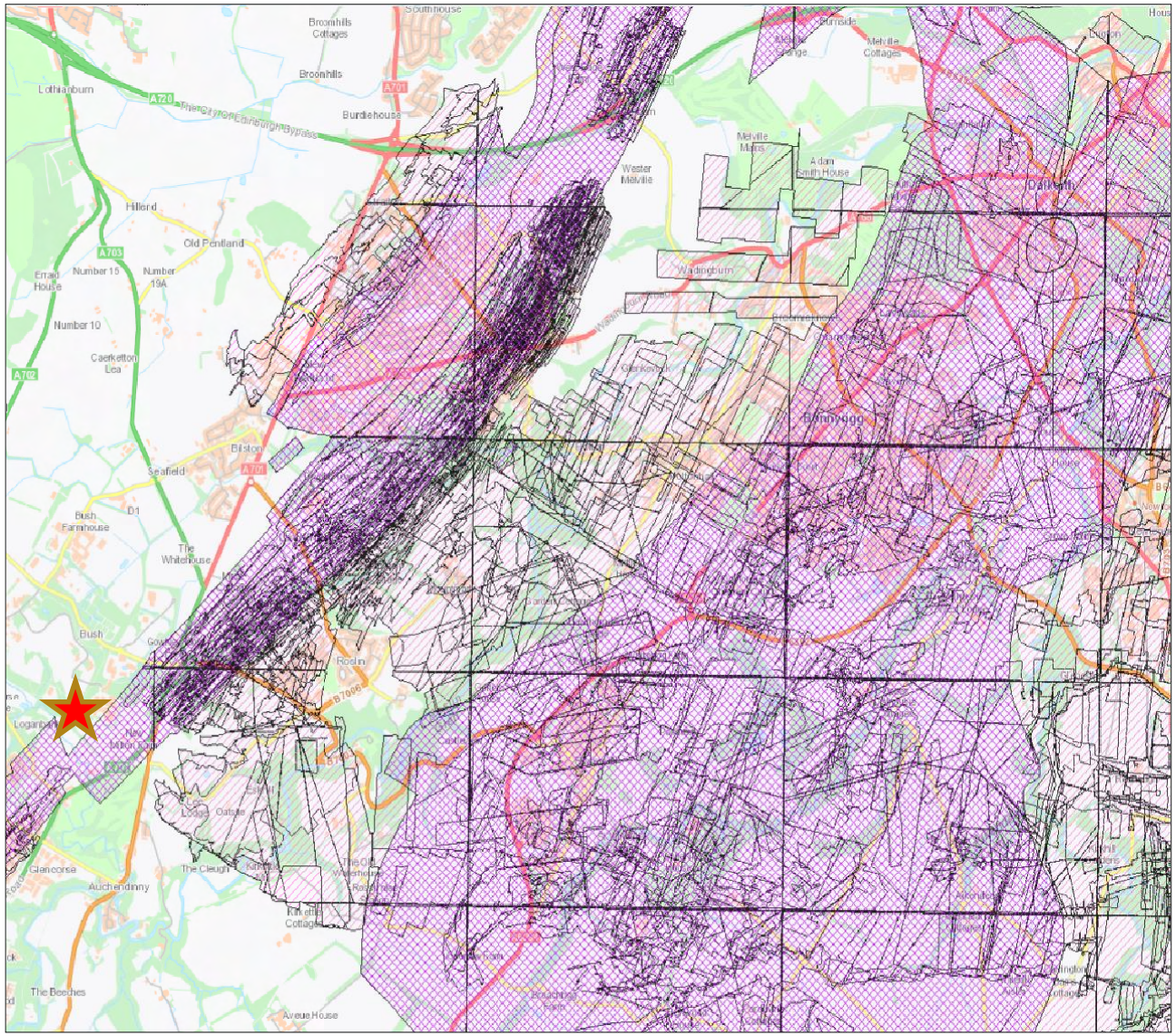
ACF CR4

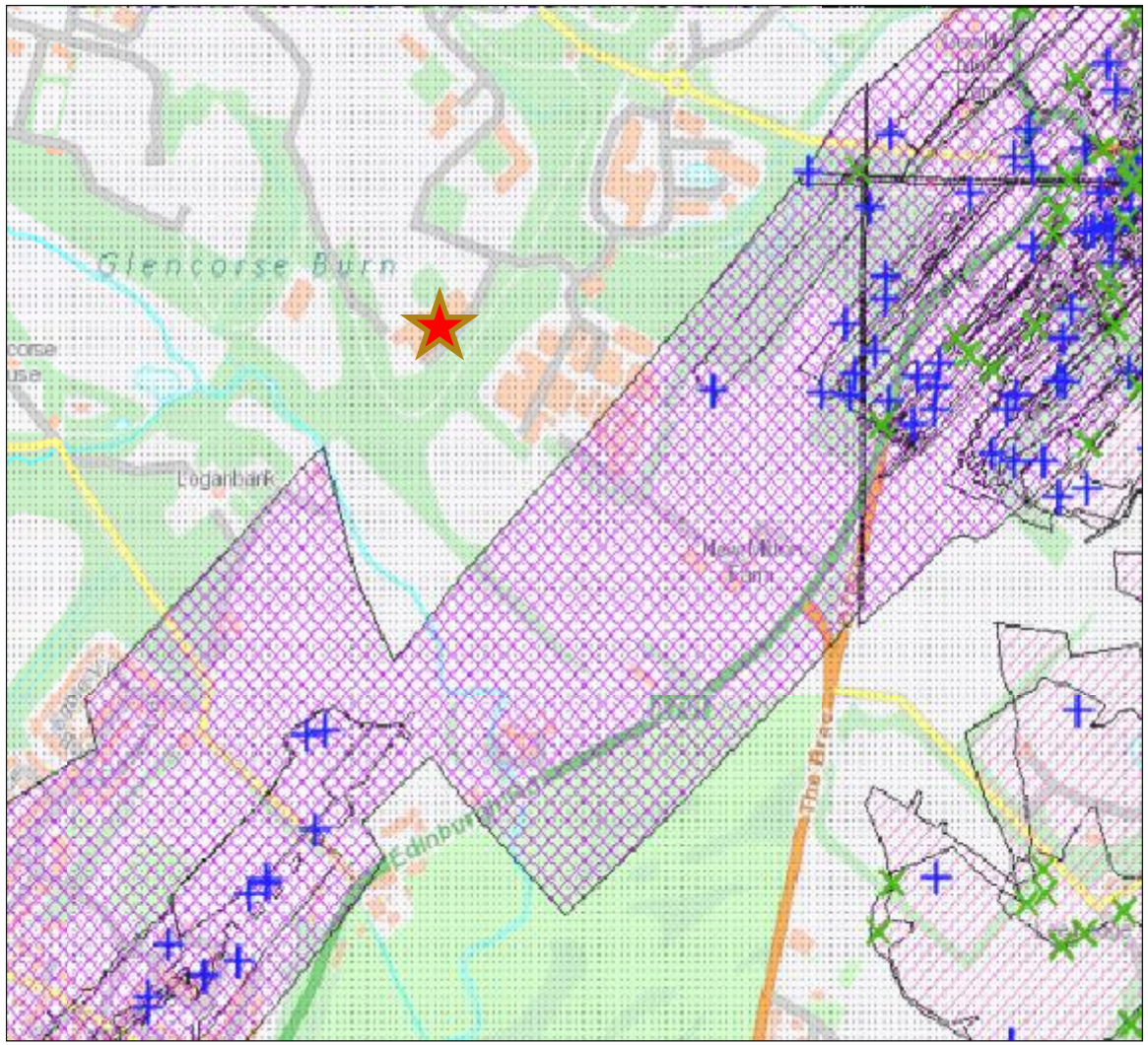
Aim for Net Zero - 100% Renewable Energy

- The University of Edinburgh is part of the Scottish Public Procurement contract for electricity
- We choose the 100% renewable energy option



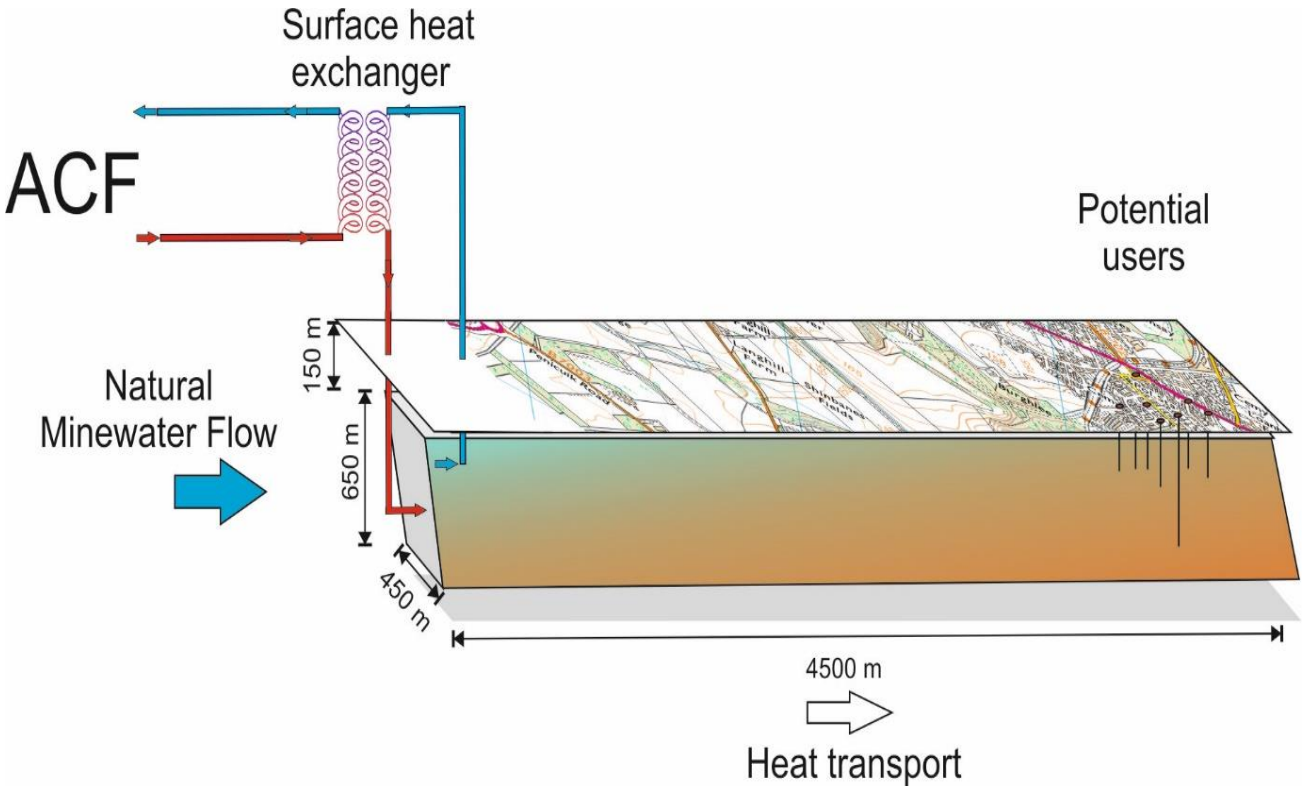
- The ACF consumed 24.46 GWhrs in FY2018/19 …
- With ARCHER 2 this will rise to ~50 GWhrs per annum


Broomhills Cottages
Broomhills
Lothianburn
A720
The City of Edinburgh Bypass
A701
Burdiehouse
Melville Cottages
Melville Mains
Adam Smith House
Wester Melville
Dalkeith
Hillend
Old Pentland
A703
Number 15
Number 19A
Number 10
Caerketton Lea
A702
Wadingburn
Bilston
A701
Seafield
B1
Bush Farmhouse
The Whitehouse
Bush
Roslin
Glencorse
Auchendinny
The Cleugh
The Beeches
Avenue House

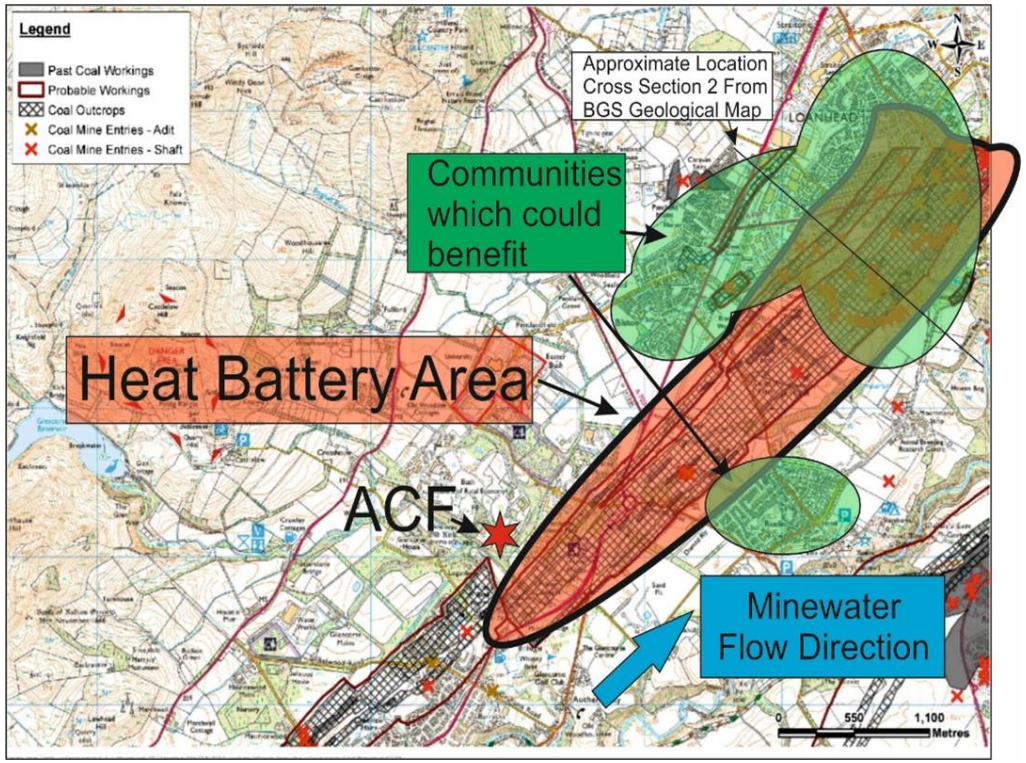

Glencorse Burn
Loganbank
Edinburgh

epcc

Aiming for better than Net Zero



- Detailed feasibility study now completed to use hot water to heat abandoned mine workings
- Will create geothermal heat battery for us by homes, public and commercial buildings
- Battery will extend into South Edinburgh



Bilston Glen Colliery, 670m, 15.0C, Minewater
Monktonhall, 866m, 25.5C, Rock
Lady Victoria, 768m, 18C, Minewater

epcc

Conclusion

- Delivering an Exascale capability will allow the UK's computational science community to compete with their international peers
- A true demonstration of the UK as a Science & Technology superpower
- But …
  - There is no guarantee funding will be made available
  - Timescales can easily slip
  - As many current projects are showing, these very large systems are not easy to procure, install or operate
- … however, if we don't try we'll never succeed!

Jeff Hammond and Filippo Spiga (NVIDIA)

Shifting through the Gears of GPU Programming: Understanding Performance and Portability Trade-offs

Abstract: This talk will show implementations of standard linear algebra algorithms in a range of programming models, including standard language parallelism, directives/pragmas, and CUDA, and how the performance and productivity varies across these. Unlike my GTC talk, this one will show Python results, in addition to Fortran, C and C++.

Bio: Jeff Hammond is a Principal Programming Model Architect at NVIDIA, working on open standards, the ARM HPC software ecosystem, and scientific applications. He is based in Helsinki, Finland. Previously, Jeff worked at Intel and Argonne on a wide range of HPC hardware and software projects. He received his PhD in Chemistry from the University of Chicago for work on NWChem.

Filippo Spiga works as EMEA HPC Developer Relations manager and Arm HPC ecosystem Alliance manager at NVIDIA. In these roles, he works closely with computational scientists from several science domains to understand their needs and help them prepare their software to run efficiently on current and future Arm-based GPU-accelerated. Prior NVIDIA, Filippo worked at Arm Research, University of Cambridge, ICHEC, CINECA and IBM Research. He has been developing and contributing in various HPC codes (mainly Physics, Chemistry and Engineering) for more than a decade. He is based in Cambridge (UK).

SHIFTING THROUGH THE GEARS OF GPU PROGRAMMING: UNDERSTANDING PERFORMANCE AND PORTABILITY TRADE-OFFS

JEFF HAMMOND, PRINCIPAL ENGINEER

ABOUT ME

I joined NVIDIA in 2021 as part of the NVHPC software team, with a special focus on ARM CPUs in HPC and Fortran.

I've been around HPC for a while, having worked at Argonne and Intel on a range of projects related to MPI, scientific computing, linear algebra, and supercomputing architecture.

My background is computational chemistry and I've been working on HPC applications for more than 15 years.

I am based in Helsinki, Finland.

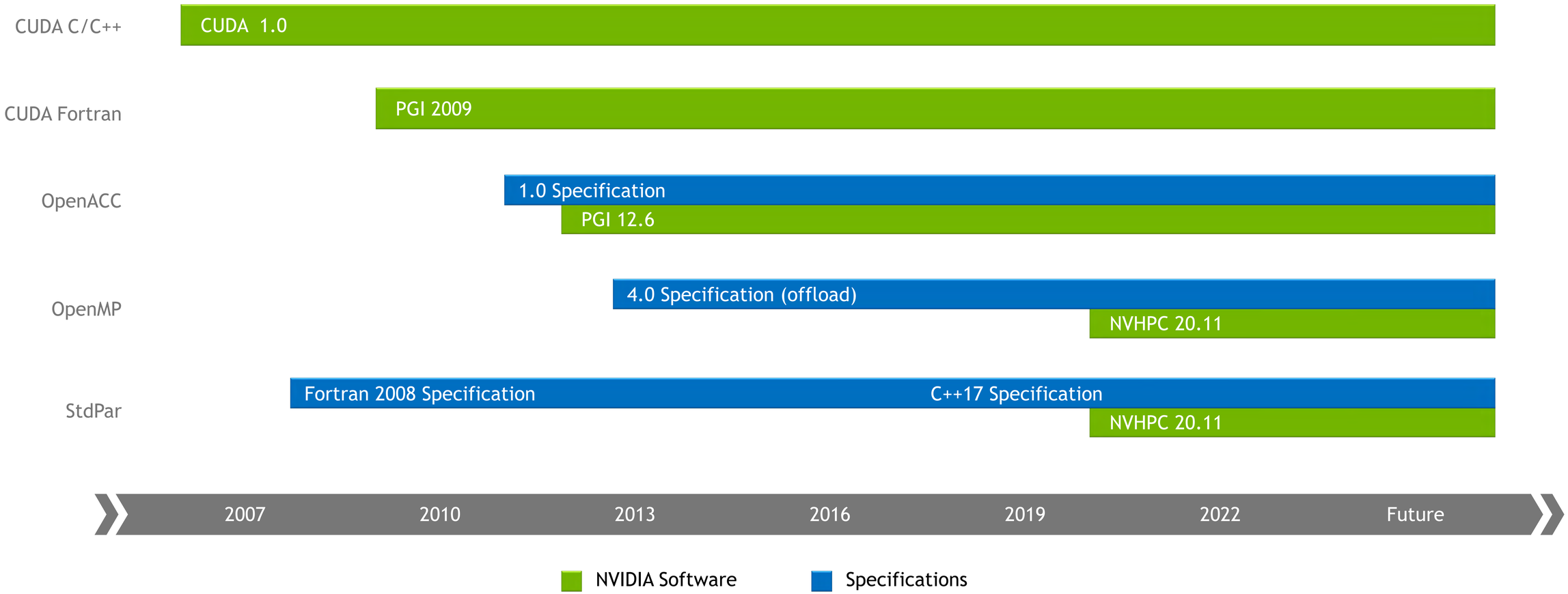
https://github.com/jeffhammond

SHIFTING THROUGH THE GEARS OF GPU PROGRAMMING: UNDERSTANDING PERFORMANCE AND PORTABILITY TRADE-OFFS

- A brief history of GPU programming models
- Overview of NVIDIA compiler and language support
- What is the GPU gearbox?
- Description of experiments with vector and matrix computations
- Performance results and analysis

GPU PROGRAMMING MODELS

A brief history



NVIDIA Compiler and Language Support

Accelerated Standard Languages

```
std::transform(par, x, x+n, y, y,
    [=](float x, float y){ return y + a*x;
}
);
```

```
do concurrent (i = 1:n)
   y(i) = y(i) + a*x(i)
enddo
```

```
import legate.numpy as np
…
def saxpy(a, x, y):
    y[:] += a*x
```

Incremental Portable Optimization

```
#pragma acc data copy(x,y) {
...
#pragma acc parallel loop
for (i=0; i<n; i++) {
  y[i] += a * x[i];
}
...
}
```

```
#pragma omp target data map(x,y) {
...
#pragma omp target teams loop
for (i=0; i<n; i++) {
  y[i] += a * x[i];
}
...
}
```

Platform Specialization

```
__global__
void saxpy(int n, float a,
           float *x, float *y) {
  int i = blockIdx.x*blockDim.x +
          threadIdx.x;
  if (i < n) y[i] += a*x[i];
}

int main(void) {
  ...
  cudaMemcpy(d_x, x, ...);
  cudaMemcpy(d_y, y, ...);

  saxpy<<<(N+255)/256,256>>>(...);

  cudaMemcpy(y, d_y, ...);
```

Core	Math	Communication	Data Analytics	AI	Quantum

Acceleration Libraries

https://developer.nvidia.com/nvidia-hpc-sdk-downloads

NVIDIA

https://commons.wikimedia.org/wiki/File:Ferrari_F355_Spider_-_Flickr_-_The_Car_Spy_(16).jpg

WHAT IS THE GPU GEARBOX?

The GPU gearbox is a mental model for thinking about programming models, to deliver the best performance at different levels of developer effort and specialization.

Think about torque, not speed…

First Gear

ISO standard parallelism: Easiest to adopt. Maximum portability. Good performance in a subset of use cases.

Second Gear

Performance libraries: Peak performance for supported features, which include a wide range of common patterns in linear algebra, machine learning and data analysis.

Third Gear

Directives and Pragmas: Easy to adopt. Good portability. Great performance in many use cases.

Fourth Gear

CUDA languages: Exposes full hardware capability and enables maximum performance. Supported on all NVIDIA GPUs.

https://commons.wikimedia.org/wiki/File:Ferrari_F355_Spider_-_Flickr_-_The_Car_Spy_(16).jpg

WHAT IS THE GPU GEARBOX?

The GPU gearbox is a mental model for thinking about programming models, to deliver the best performance at different levels of developer effort and specialization.

Think about torque, not speed…

First Gear

ISO standard parallelism: Easiest to adopt. Maximum portability. Good performance in a subset of use cases.

Second Gear

Performance libraries: Peak performance for supported features, which include a wide range of common patterns in linear algebra, machine learning and data analysis.

Third Gear

Directives and Pragmas: Easy to adopt. Good portability. Great performance in many use cases.

Fourth Gear

CUDA languages: Exposes full hardware capability and enables maximum performance. Supported on all NVIDIA GPUs.

https://commons.wikimedia.org/wiki/File:Ferrari_F355_Spider_-_Flickr_-_The_Car_Spy_(16).jpg

WHAT IS THE GPU GEARBOX?

The GPU gearbox is a mental model for thinking about programming models, to deliver the best performance at different levels of developer effort and specialization.

Think about torque, not speed…

First Gear

ISO standard parallelism: Easiest to adopt. Maximum portability. Good performance in a subset of use cases.

Second Gear

Performance libraries: Peak performance for supported features, which include a wide range of common patterns in linear algebra, machine learning and data analysis.

Third Gear

Directives and Pragmas: Easy to adopt. Good portability. Great performance in many use cases.

Fourth Gear

CUDA languages: Exposes full hardware capability and enables maximum performance. Supported on all NVIDIA GPUs.

https://commons.wikimedia.org/wiki/File:Ferrari_F355_Spider_-_Flickr_-_The_Car_Spy_(16).jpg

WHAT IS THE GPU GEARBOX?

The GPU gearbox is a mental model for thinking about programming models, to deliver the best performance at different levels of developer effort and specialization.

Think about torque, not speed…

First Gear

ISO standard parallelism: Easiest to adopt. Maximum portability. Good performance in a subset of use cases.

Second Gear

Performance libraries: Peak performance for supported features, which include a wide range of common patterns in linear algebra, machine learning and data analysis.

Third Gear

Directives and Pragmas: Easy to adopt. Good portability. Great performance in many use cases.

Fourth Gear

CUDA languages: Exposes full hardware capability and enables maximum performance. Supported on all NVIDIA GPUs.

https://commons.wikimedia.org/wiki/File:Ferrari_F355_Spider_-_Flickr_-_The_Car_Spy_(16).jpg

WHAT IS THE GPU GEARBOX?

The GPU gearbox is a mental model for thinking about programming models, to deliver the best performance at different levels of developer effort and specialization.

Think about torque, not speed…

First Gear

ISO standard parallelism: Easiest to adopt. Maximum portability. Good performance in a subset of use cases.

Second Gear

Performance libraries: Peak performance for supported features, which include a wide range of common patterns in linear algebra, machine learning and data analysis.

Third Gear

Directives and Pragmas: Easy to adopt. Good portability. Great performance in many use cases.

Fourth Gear

CUDA languages: Exposes full hardware capability and enables maximum performance. Supported on all NVIDIA GPUs.

SHIFTING THROUGH THE GEARS

Experiments with linear algebra primitives

VECTOR ADDITION
Easy

MATRIX TRANSPOSE
Medium

MATRIX MULTIPLICATION
Hard

$$\forall i: \; Z_i = a \times X_i + Y_i$$

$$\forall i,j: \; B_{i,j} = B_{i,j} + A_{j,i}$$

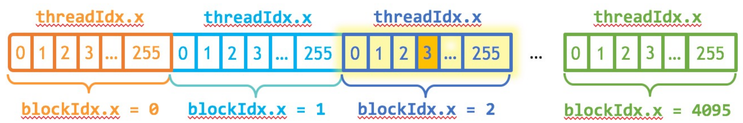
$$\forall i,j: \; C_{i,j} = \sum_k A_{i,k} \times B_{k,j}$$

SHIFTING THROUGH THE GEARS

Experiments with linear algebra primitives

VECTOR ADDITION

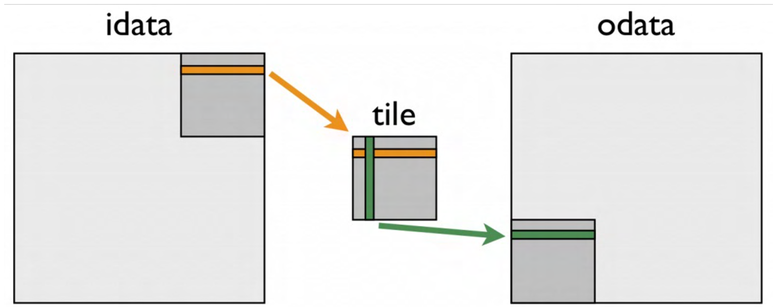
Memory bandwidth



https://developer.nvidia.com/blog/even-easier-introduction-cuda/

MATRIX TRANSPOSE

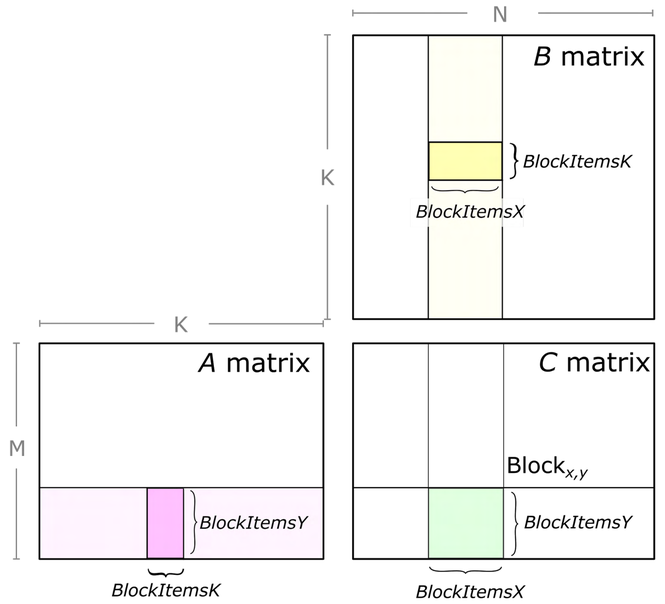
Memory bandwidth, shared memory, and coalescing



https://developer.nvidia.com/blog/efficient-matrix-transpose-cuda-cc/

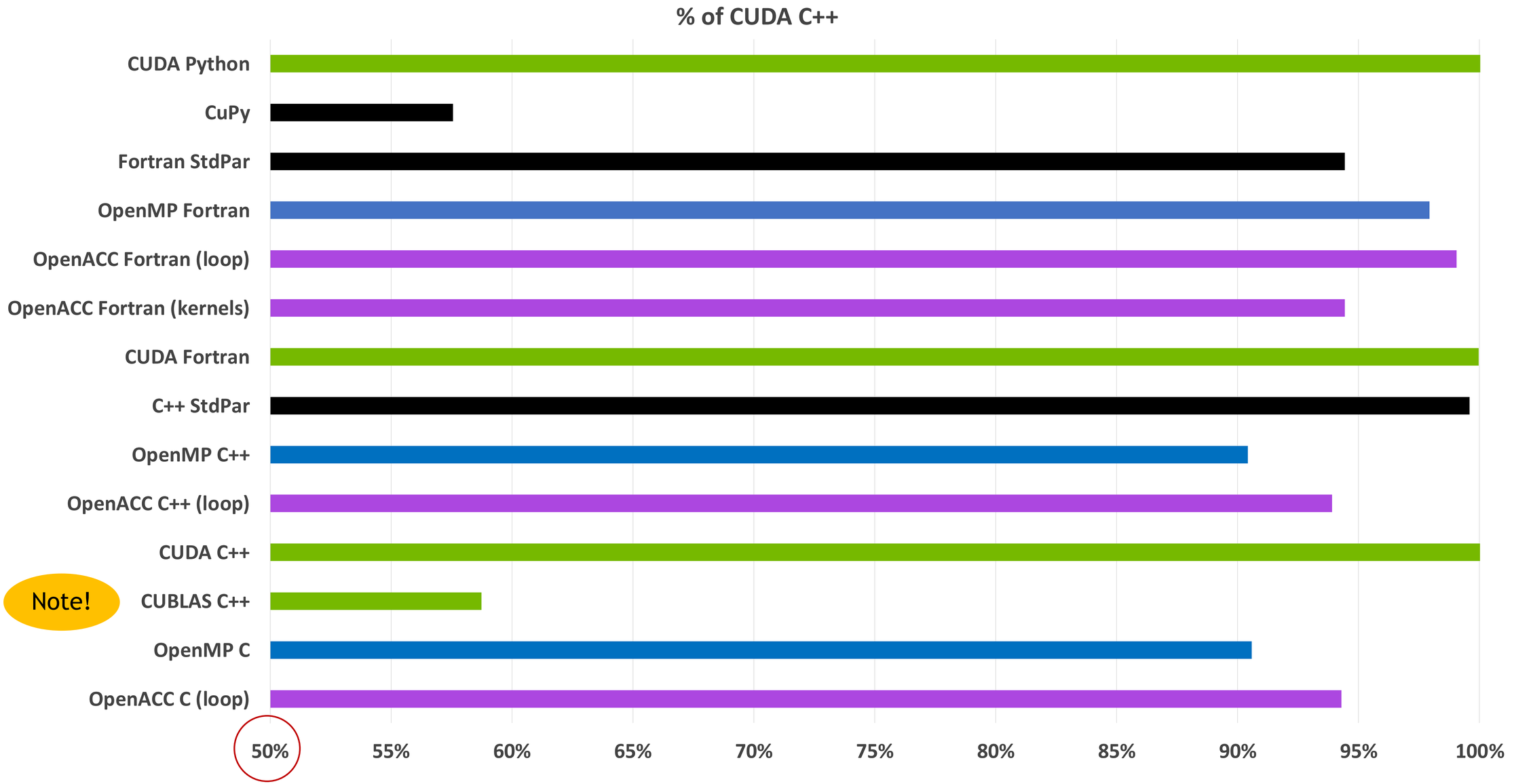
MATRIX MULTIPLICATION

Optimize everything…



https://developer.nvidia.com/blog/cutlass-linear-algebra-cuda/

Vector Addition: Z = a * X + Y



Vector Addition

Representative implementations

```
// CUDA C++

__global__
void saxpy(size_t n, T a, T * X, T * Y, T * Z)
{
    auto i = blockIdx.x * blockDim.x + threadIdx.x;
    if (i < n) {
        Z[i] = a * X[i] + Y[i];
    }
}

const int block_size = 256;
dim3 dimBlock(block_size, 1, 1);
dim3 dimGrid(length/block_size, 1, 1);

axpy<<<dimGrid, dimBlock>>>(length, a, X, Y, Z);
```

```
// C++17 standard parallelism

std::transform( std::execution::par_unseq,
                std::begin(X), std::end(X),
                std::begin(Y), std::begin(Z),
                [a](auto&& x, auto&& y) {
                    return a * x + y;
                }
);
```

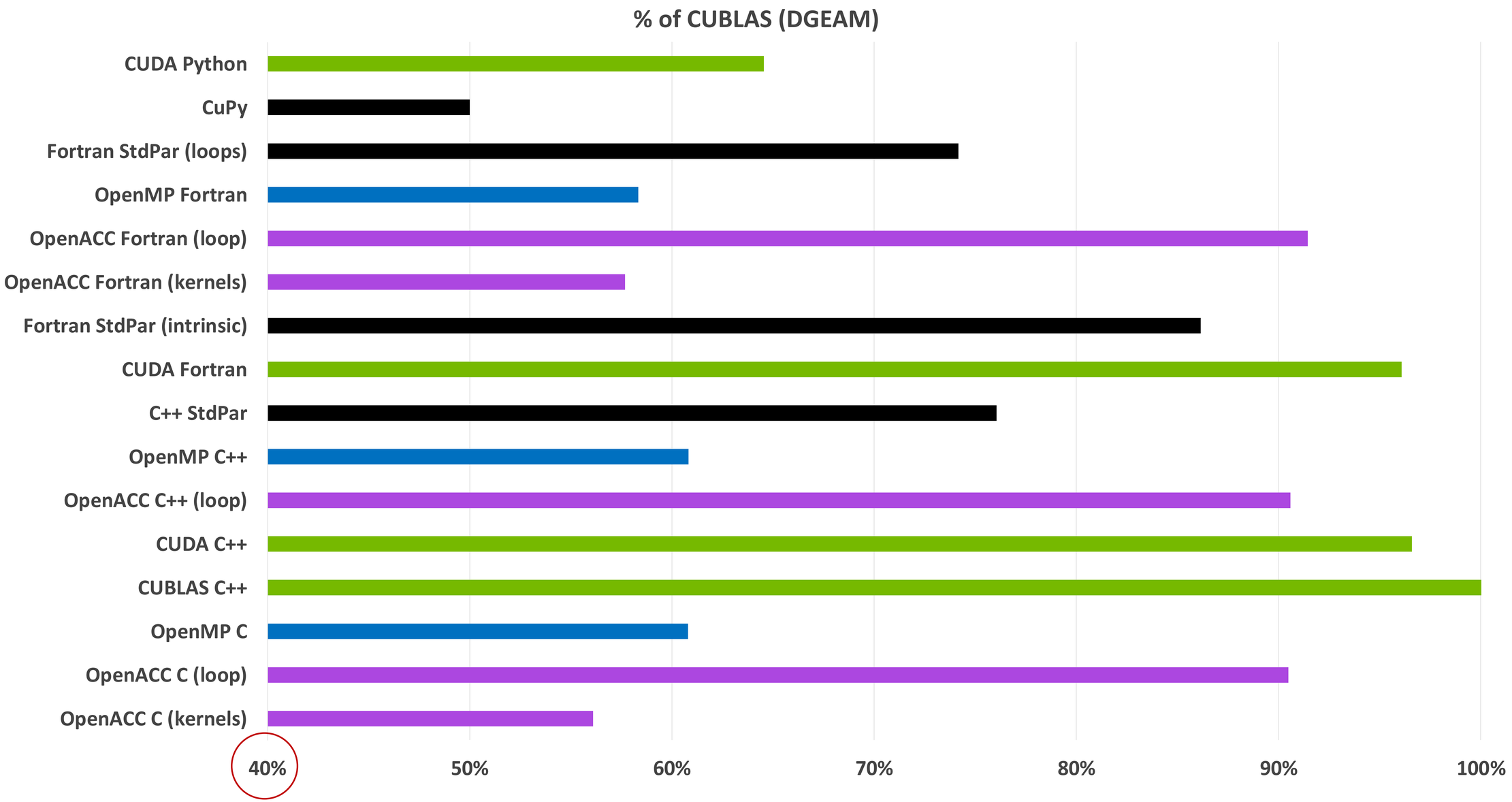
```
// OpenACC C++

#pragma acc parallel loop
for (size_t i=0; i<length; ++i) {
    Z[i] = a * X[i] + Y[i];
}
```

Vector Addition

Conclusions

- Most programming models achieve >90% of the performance of CUDA C++. In general, bandwidth-limited algorithms have a high degree of performance portability.
- For C++, first gear (ISO parallelism) delivers with maximum portability 99.6% of CUDA, which is close to 2 TB/s on A100 GPUs.
- Once you are able to take advantage of massive bandwidth on GPUs via StdPar, look for opportunities to increase compute intensity via loop/kernel fusion, higher-order methods, etc. in order to take advantage of higher gears.
- NVIDIA compilers support all of these models concurrently in an application, so you can start with first gear and shift up on a case-by-case basis.

Matrix Transpose: B = B + A^T



Matrix Transpose

Representative implementations

```
! CUDA Fortran

integer(kind=INT32), parameter :: tile_dim = 32
integer(kind=INT32), parameter :: block_rows = 8

attributes(global) subroutine transpose(N, A, B)
  implicit none
  integer(kind=INT32), intent(in), value :: N
  real(kind=REAL64), intent(inout) :: A(N,N)
  real(kind=REAL64), intent(inout) :: B(N,N)
  real(kind=REAL64), shared :: tile(33,32)
  integer :: x, y, j
  x = (blockIdx%x-1) * tile_dim + (threadIdx%x);
  y = (blockIdx%y-1) * tile_dim + (threadIdx%y);
  do j = 0,tile_dim-1,block_rows
    tile(threadIdx%x,threadIdx%y+j) = A(x,y+j);
  end do
  call syncThreads()
  x = (blockIdx%y-1) * tile_dim + (threadIdx%x);
  y = (blockIdx%x-1) * tile_dim + (threadIdx%y);
  do j = 0,tile_dim-1,block_rows
    B(x,y+j) = B(x,y+j) + tile(threadIdx%y+j,threadIdx%x)
  end do
end subroutine transpose
```

```
! Fortran standard parallelism

! Intrinsic version
B = B + transpose(A)

! Loop version
do concurrent (j=1:N, i=1:N)
  B(i,j) = B(i,j) + A(j,i)
enddo
```

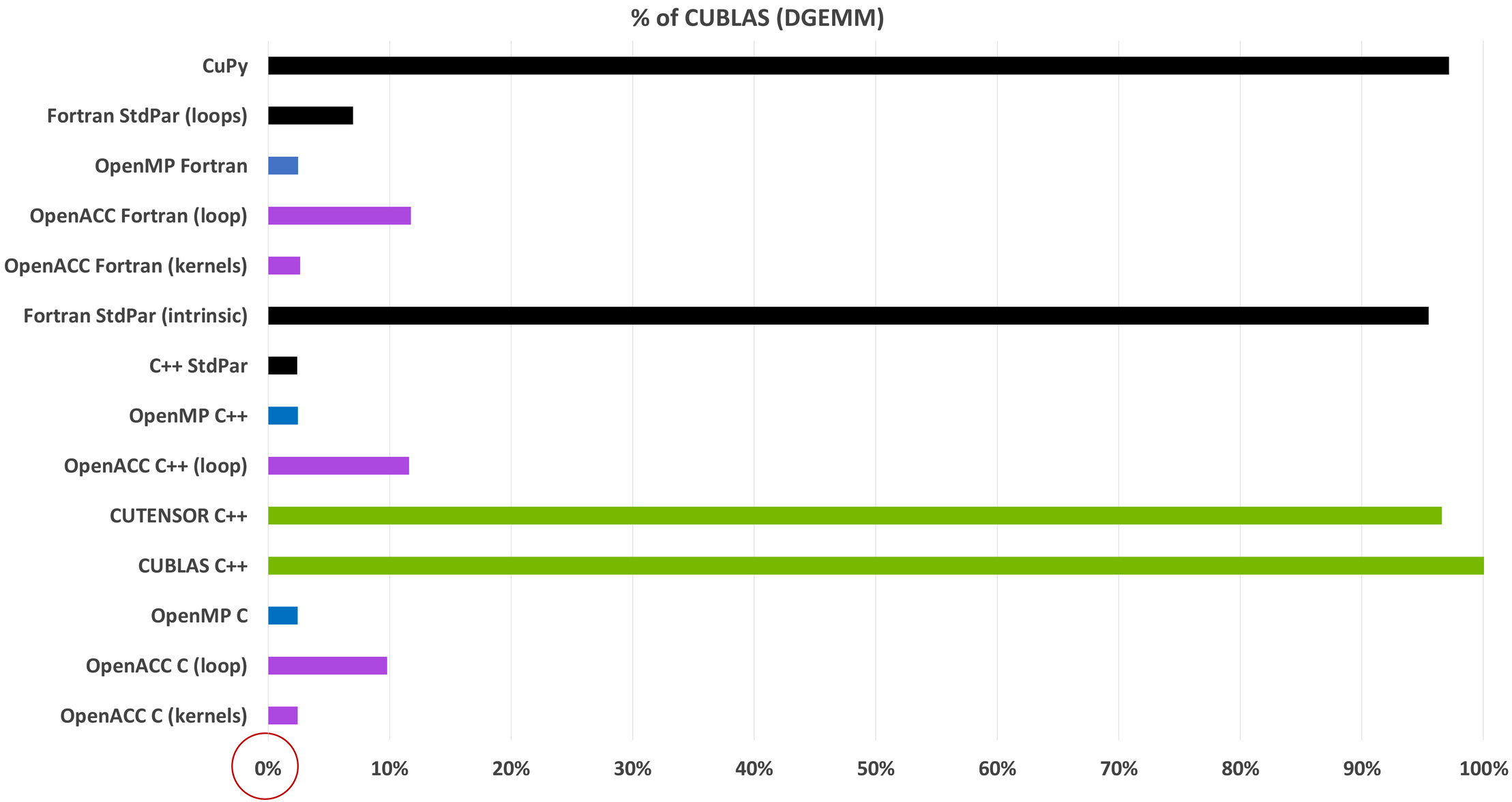
```
! OpenACC Fortran

!$acc parallel loop tile(32,32)
do j=1,N
  do i=1,N
    B(i,j) = B(i,j) + A(j,i)
  enddo
enddo
```

Matrix Transpose

Conclusions

- Calling CUBLAS is the best option if it supports your algorithm.
- CUDA C++ and Fortran provide the explicit control required to cause coalescing.
- OpenACC's tile construct allows it to achieve >90% of CUBLAS.
- Naïve implementations based on OpenMP or OpenACC kernels do not perform well.
- Fortran math intrinsics use NVIDIA performance libraries internally and perform well.
- C++17 patterns are not a good match for matrix transpose, but it still does a good job. Features in future versions of C++ improve performance.

Matrix Multiplication: C = C + A * B



Matrix Multiplication

Representative implementations

```
// CUBLAS C/C++

rb = cublasDgemm(handle,
                 CUBLAS_OP_N, CUBLAS_OP_N,
                 N, N, N,
                 &alpha, A, N
                 B, N
                 &one, C, N);
```

```
! Fortran standard parallelism

! Intrinsic version
C = C + matmul(A,B)
```

```
! OpenACC Fortran

!$acc parallel loop tile(32,32)
do j=1,order
  do i=1,order
    do p=1,order
      C(i,j) = C(i,j) + A(i,p) * B(p,j)
    enddo
  enddo
enddo
!$acc end parallel
```

```
! Loop version
do concurrent (j=1:order, i=1:order) local(T)
  T = C(i,j)
  do concurrent (p=1:order) ! Implicit reduction
    T = T + A(i,p) * B(p,j)
   enddo
   C(i,j) = T
enddo
```

Matrix Multiplication

Conclusions

- Calling performance libraries is the ONLY reasonable option for compute-bound matrix computations on any architecture.
- CUTENSOR achieves similar performance to CUBLAS and supports a wide range of operations found in quantum simulations and machine learning.
- Fortran math intrinsics use NVIDIA performance libraries internally and perform well.
- OpenACC's tile construct allows it to achieve ~10% of CUBLAS. Significant effort would be required to reach >50% with directives.
- If you can write CUDA that achieves peak performance in dense linear algebra operations, please visit https://www.nvidia.com/en-us/about-nvidia/careers/ ☺

Summary

The GPU gearbox is a mental model for deciding how to invest effort in your code.

- Standard parallelism in C++ and Fortran deliver a large fraction of peak when they are a good match for the algorithm. Unlike cars, top speed in first gear is feasible.
- Never do yourself what is available in performance libraries. NVIDIA performance libraries support a huge range of algorithms (e.g. 40-dimensional tensor contractions).
- OpenACC provide an easy way to achieve architecture-specific tuning using a portable model.

Footnote

"Look ma, no data copies!"

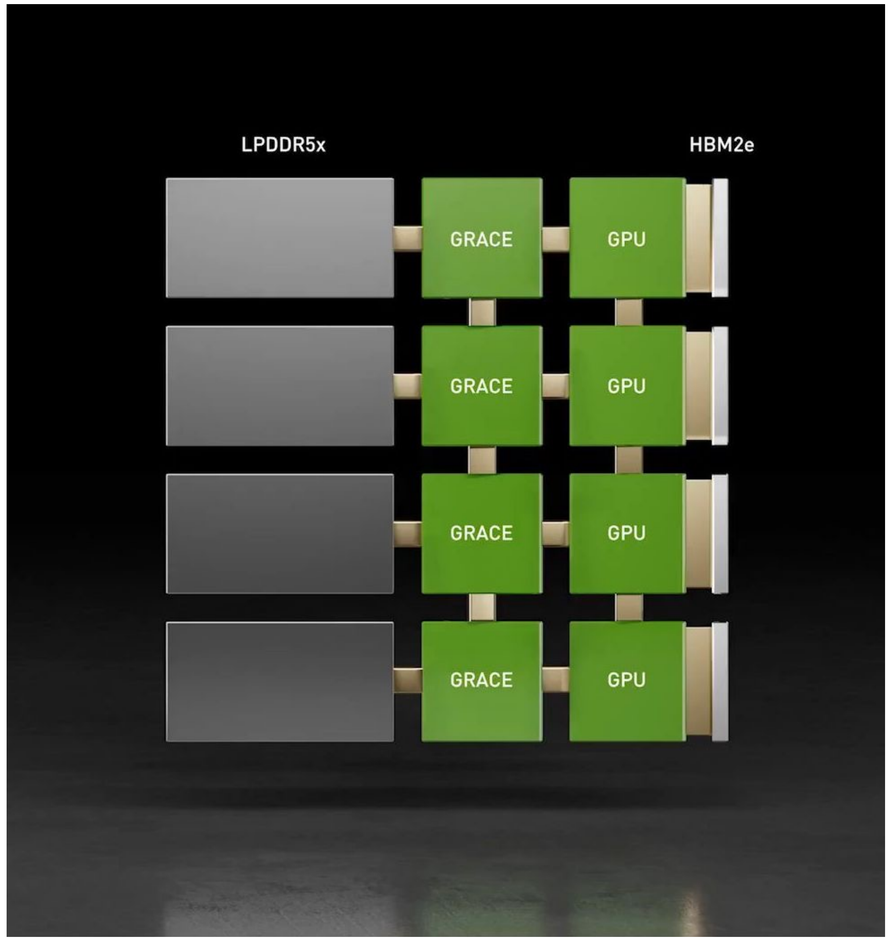
None of the codes used in this talk allocated device memory or performed explicit copies between host and device.

Every single example shown here used unified memory (UM), which shifts developer effort away from tedious memory management towards exploiting locality.

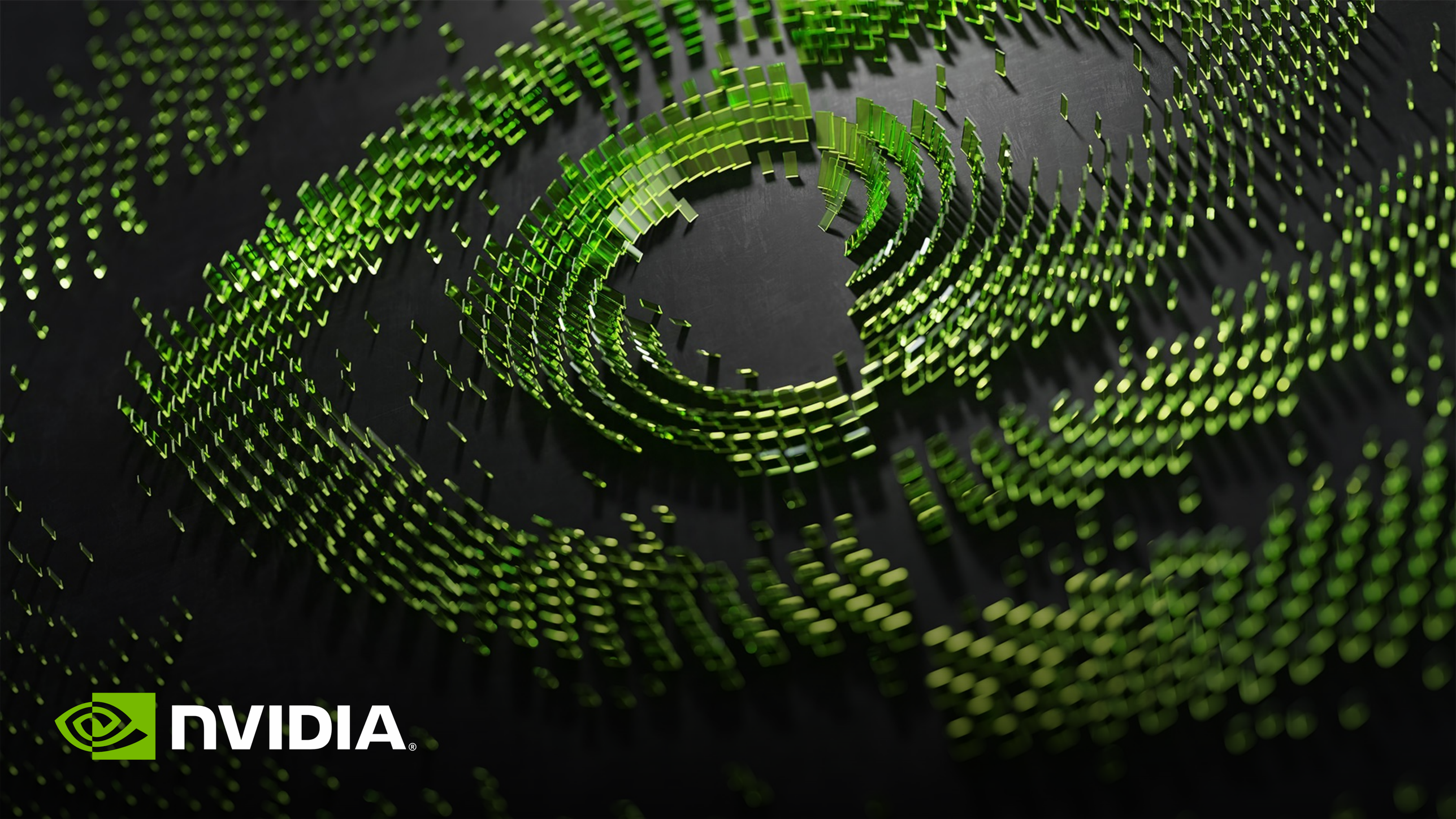
CUDA C++ code compiled with NVCC used **thrust::universal_vector**, which is backed by UM.

Everything was compiled with the NVHPC compilers using **-gpu=managed**, which enables UM in all standard allocations.

NVLink reduces the penalty – if any – of using UM everywhere.



https://www.nvidia.com/en-in/data-center/grace-cpu/


NVIDIA

CIUK 2021 Jacky Pallas Memorial Award Presentation

Dr Niall Jeffrey (Ecole Normale Supérieure & University College London)

Mapping dark matter with the Dark Energy Survey and AI

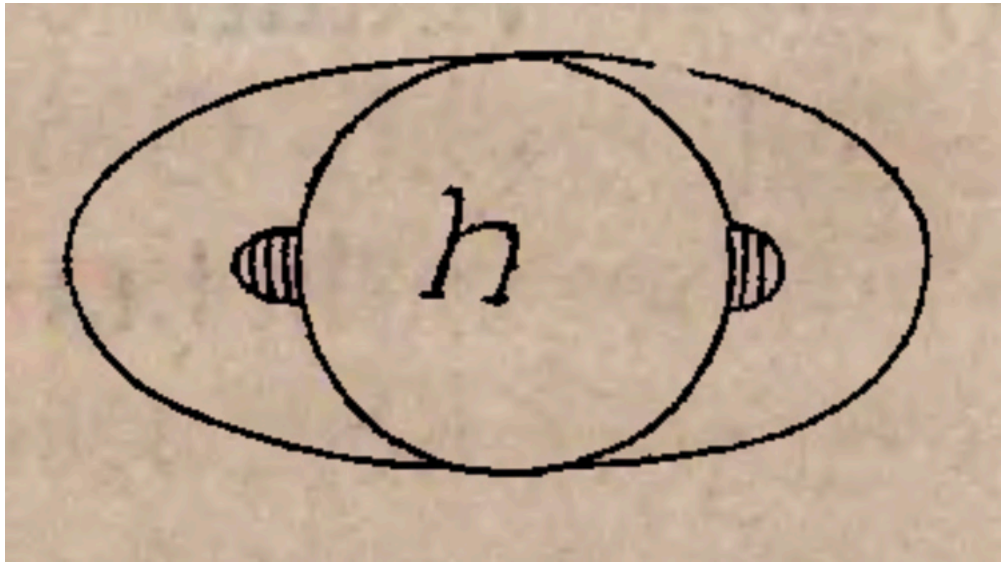
Abstract: In the Dark Energy Survey (DES), we have created the largest ever map of dark matter – invisible matter thought to account for 80% of the total matter of the Universe – using gravitational lensing of galaxies. I will share the exciting developments used for this cosmic cartography over a quarter of the Southern Hemisphere. Exploiting this map to understand the unknown physics of the Universe, in the DiRAC project “Likelihood-free inference with the Dark Energy Survey”, we combine GPU-accelerated cosmological simulations with novel artificial intelligence techniques. By using deep learning in a Bayesian framework, I will demonstrate how we can now quantify our belief in different cosmological models using information encoded in the new DES map.

Bio: Niall’s research interests combine cosmology, statistical methods and machine learning. After completing his MSci in Theoretical Physics at Imperial College London, he completed his PhD at University College London supervised by Prof. Ofer Lahav. It was during his PhD that he became involved in the Dark Energy Survey and the European Space Agency’s future Euclid mission.

The Jacky Pallas Memorial Presentation

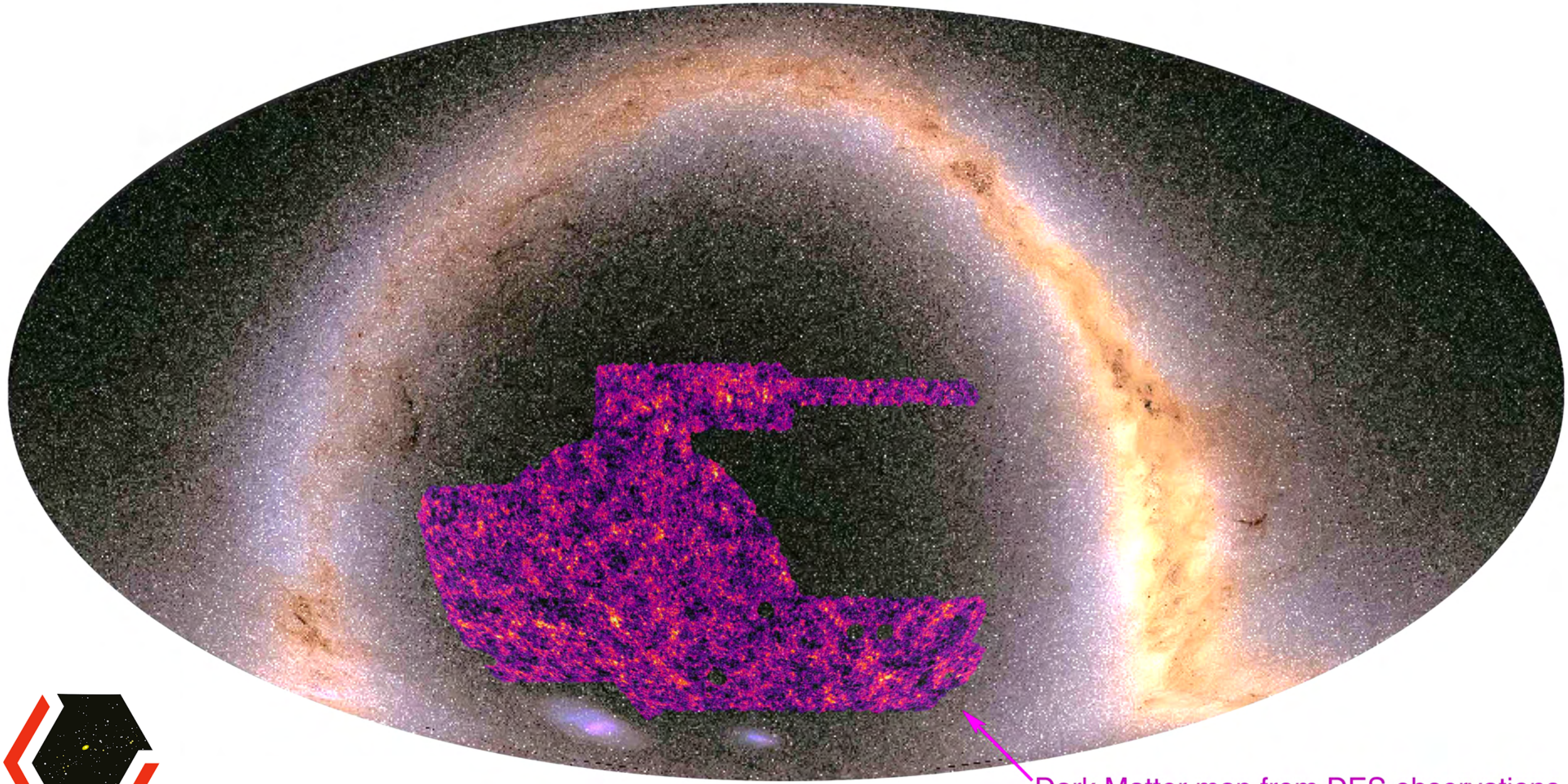
Mapping dark matter with the Dark Energy Survey and AI

Dr. Niall Jeffrey



THE DARK ENERGY SURVEY

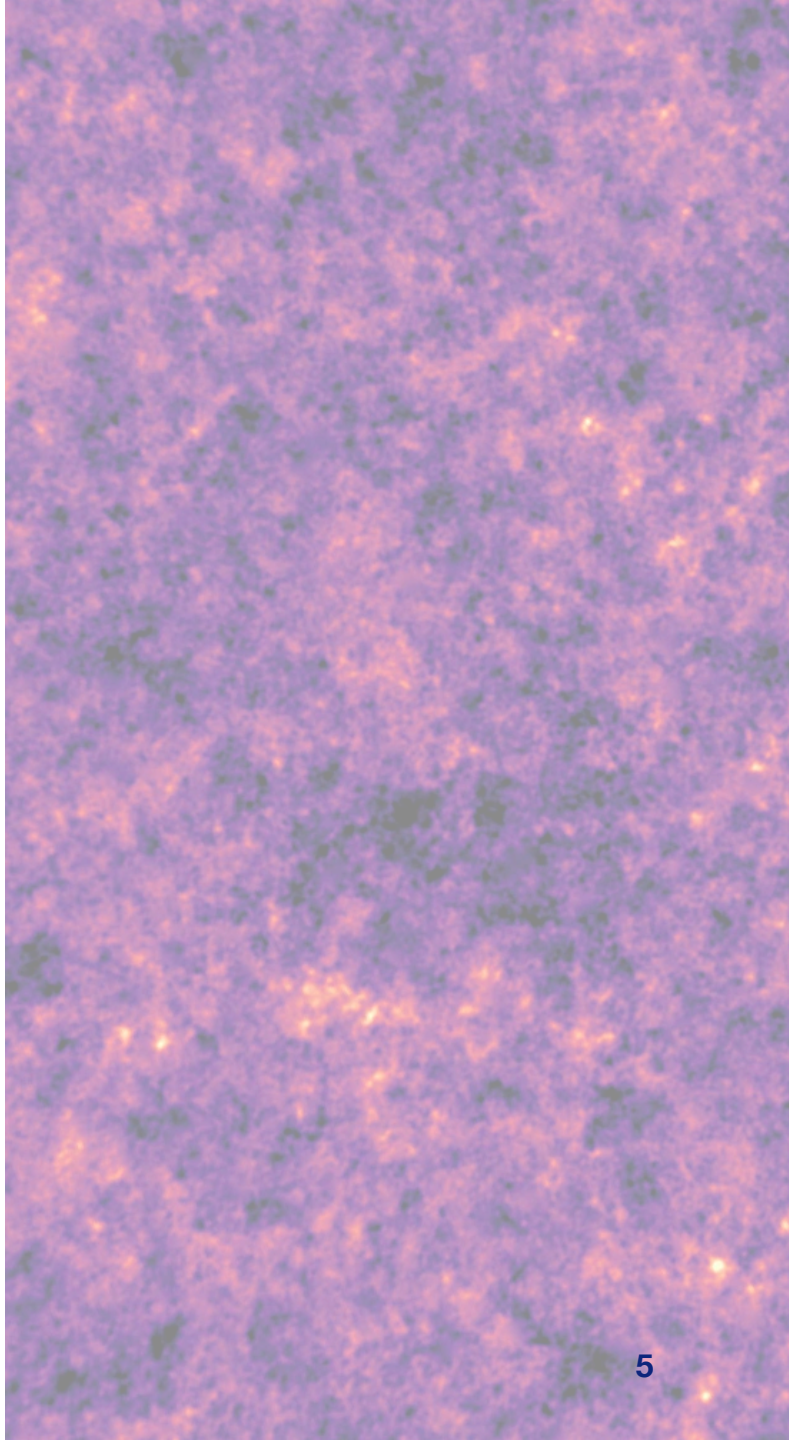

THE DARK
ENERGY SURVEY



THE DARK ENERGY SURVEY

Outline

1. Dark matter maps with DES
2. Understanding the Universe
3. Simulation-based inference
4. Future hopes with AI



THE DARK ENERGY SURVEY

01

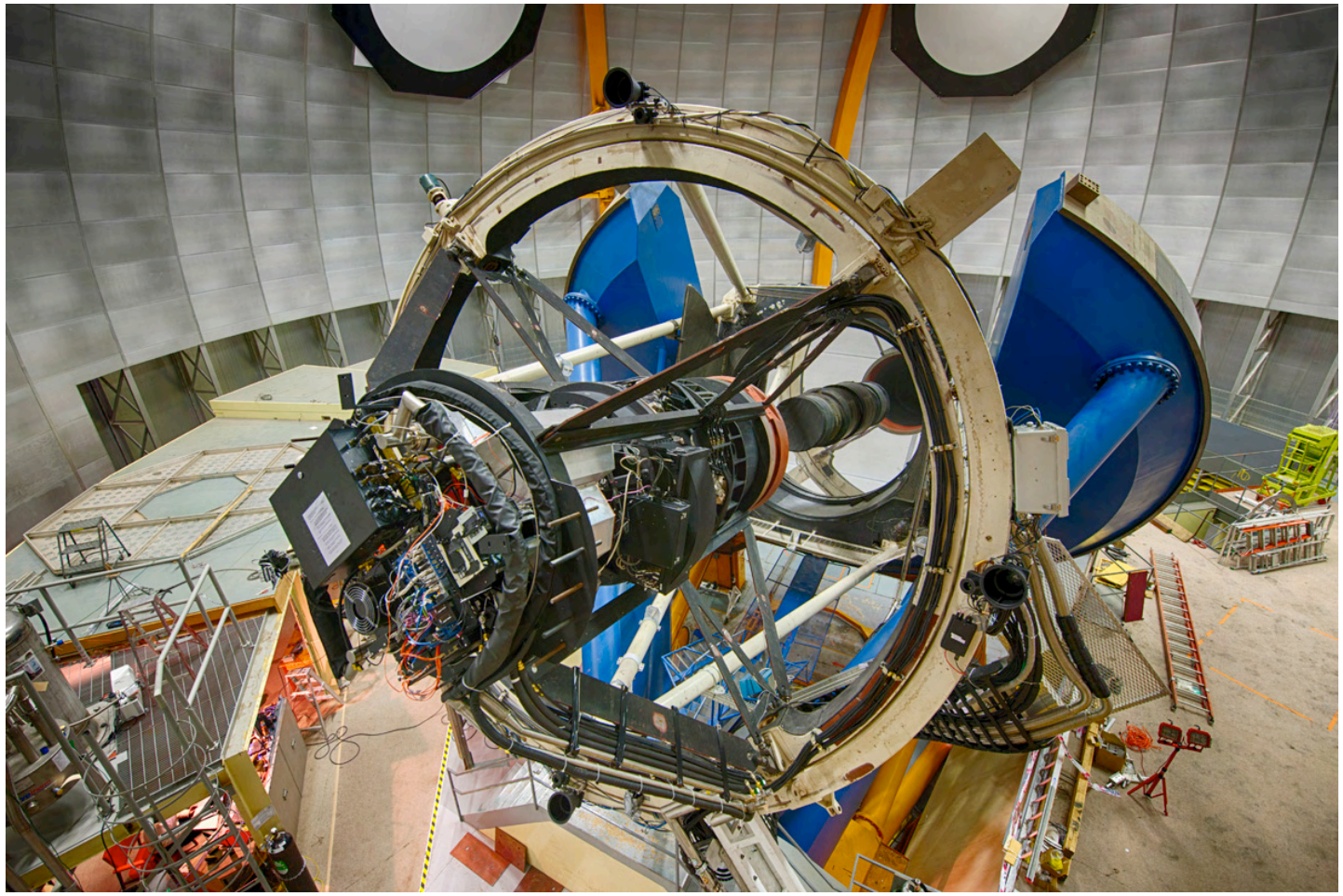
Dark matter maps with the Dark Energy Survey

Dark Energy Survey

Weak lensing data

THE DARK ENERGY SURVEY

Dark Energy Camera



THE DARK ENERGY SURVEY

First images (~2012)

I. 5-colour imaging

II. 6 years of observations

III. 100 million galaxies in Year 3 “gravitational lensing” data

THE DARK ENERGY SURVEY

How to think about “Gravitational Lensing”?

Dark Energy Survey

Weak lensing data

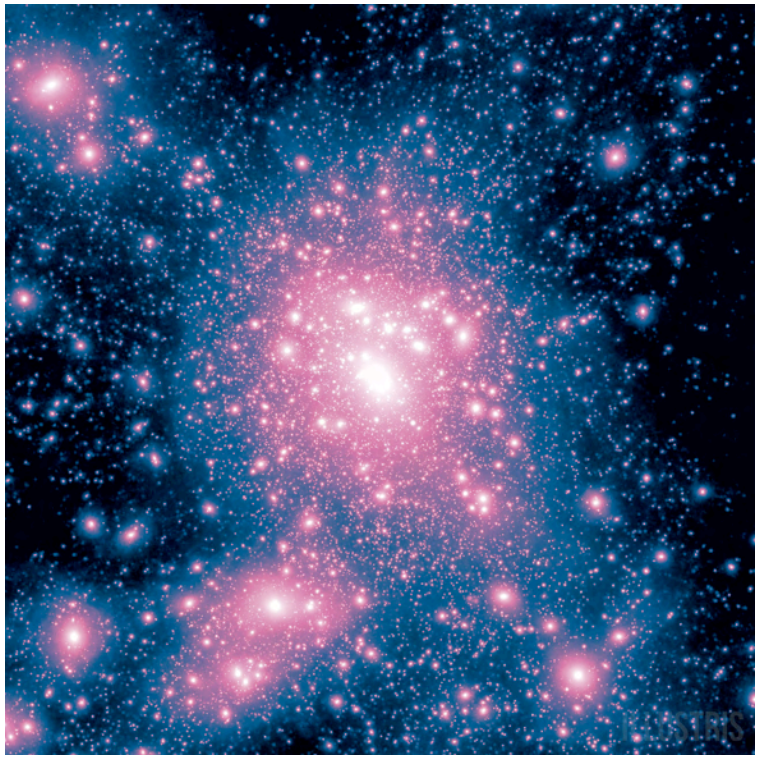
Dark Energy Survey

Weak lensing data

THE DARK ENERGY SURVEY

Dark Energy Survey

Weak lensing data



THE DARK ENERGY SURVEY

Bayesian Mass Mapping

$$p(\text{map}\,|\,\text{data}) \propto p(\text{data}\,|\,\text{map}) \times\ p(\text{map})$$

Bayesian Mass Mapping

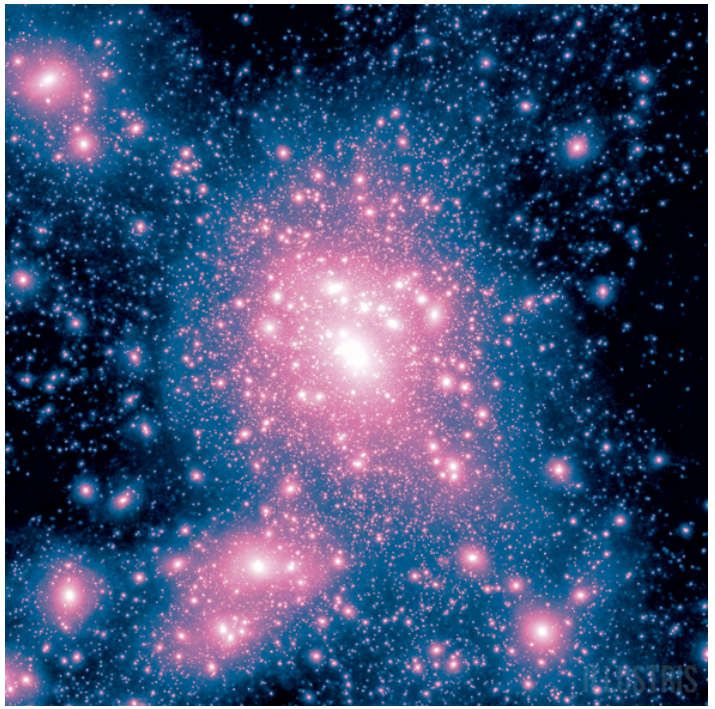
Prior choice?

$$p(\mathrm{map})$$

?

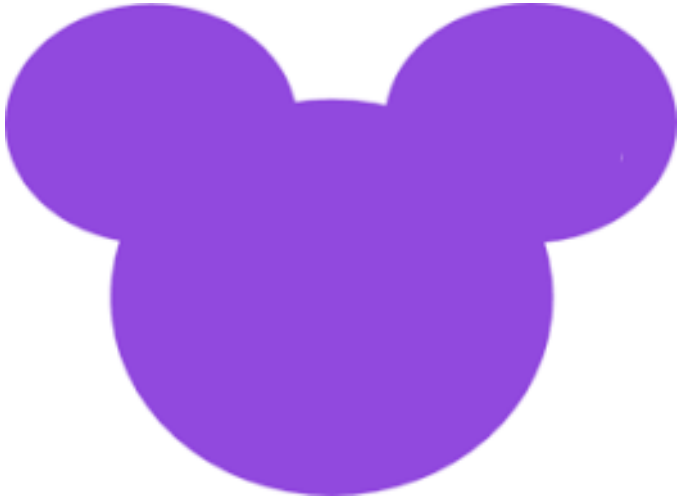
$$p(\text{map}) = \text{constant}$$

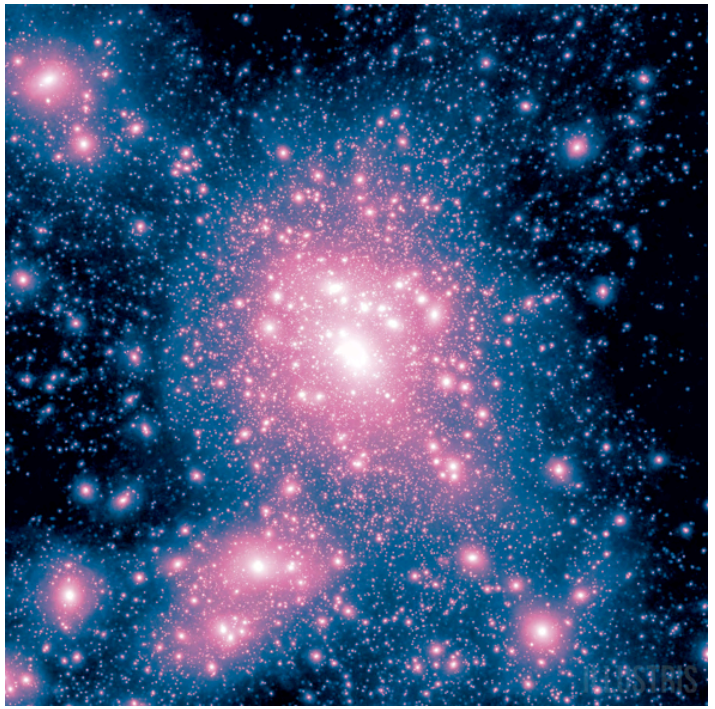
$$p(\text{map}) = \text{constant}$$



THE DARK ENERGY SURVEY

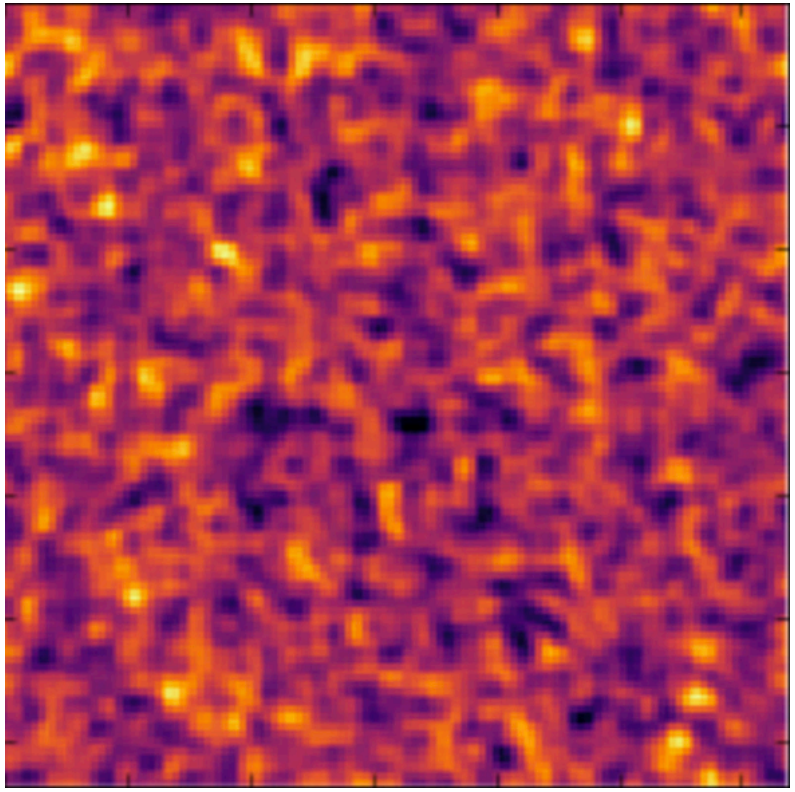
$$p(\text{map}) = \text{constant}$$



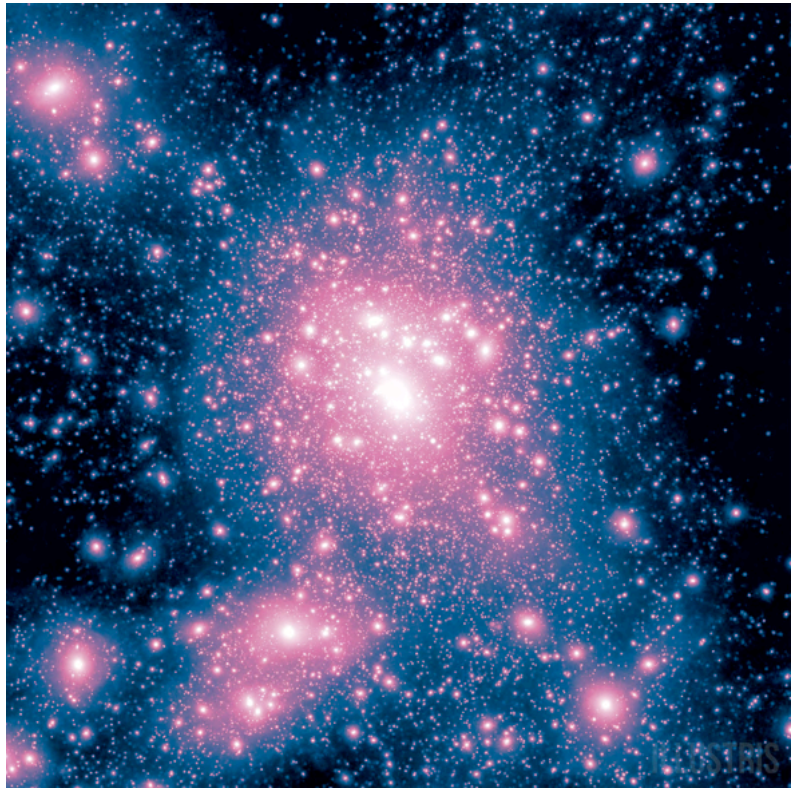


THE DARK ENERGY SURVEY

Gaussian & Sparsity priors



Wiener filter/posterior



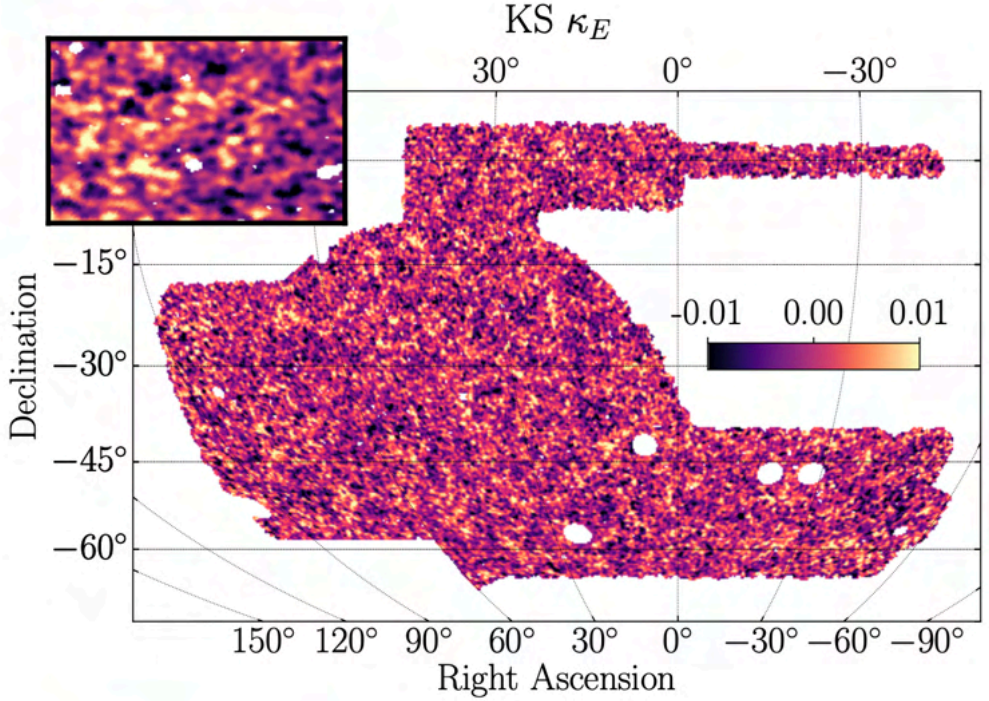
Dark matter halos

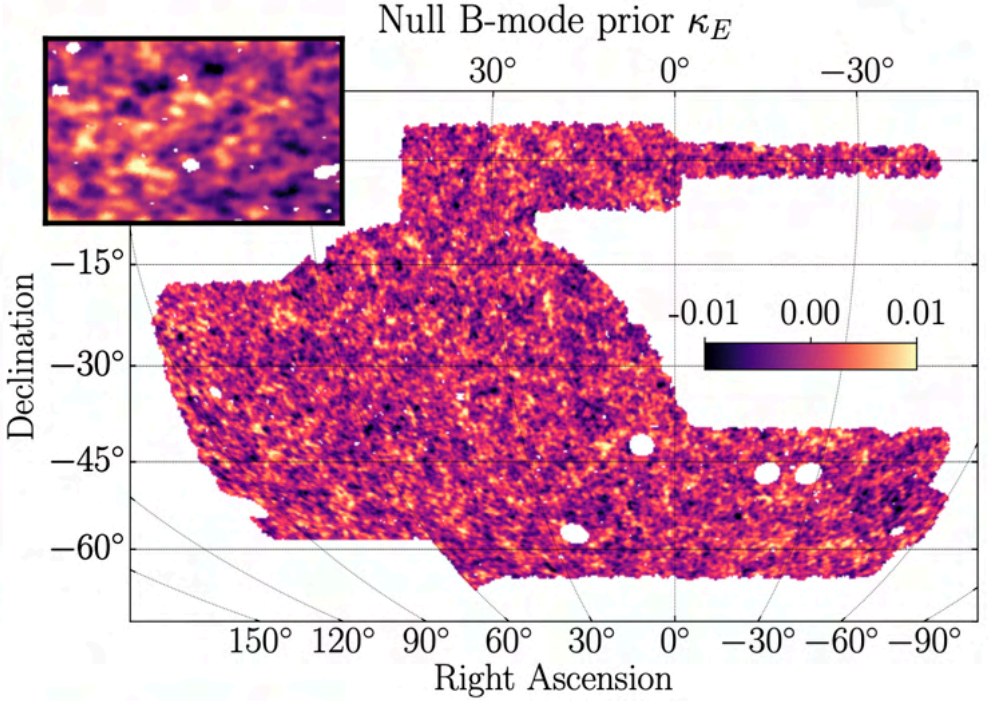
GLIMPSE

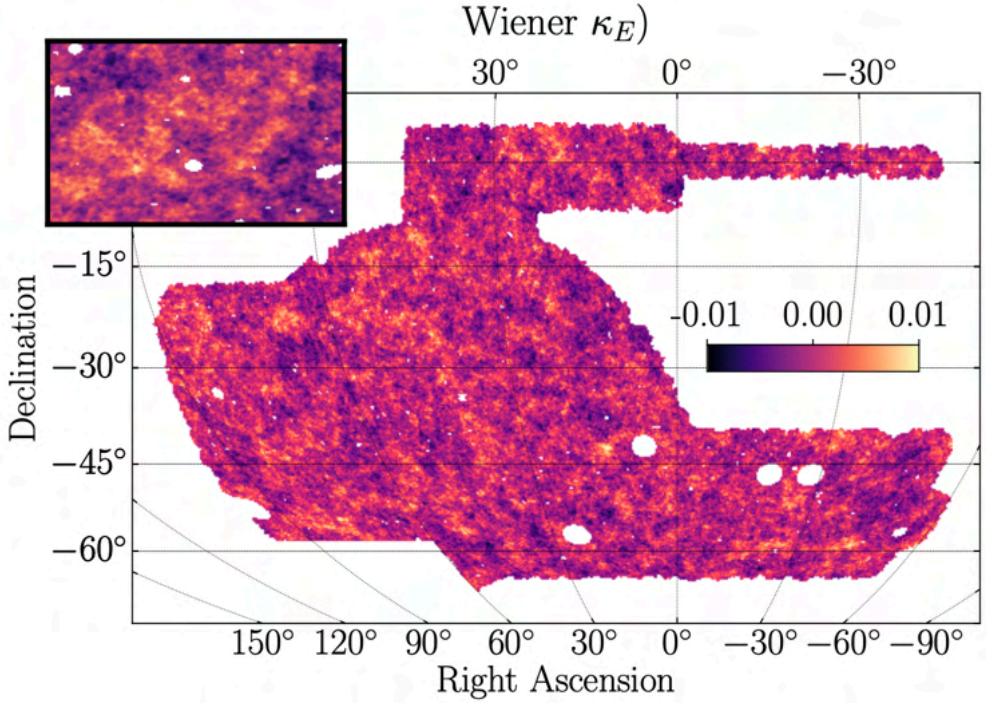
Results
(GPU-accelerated optimization)

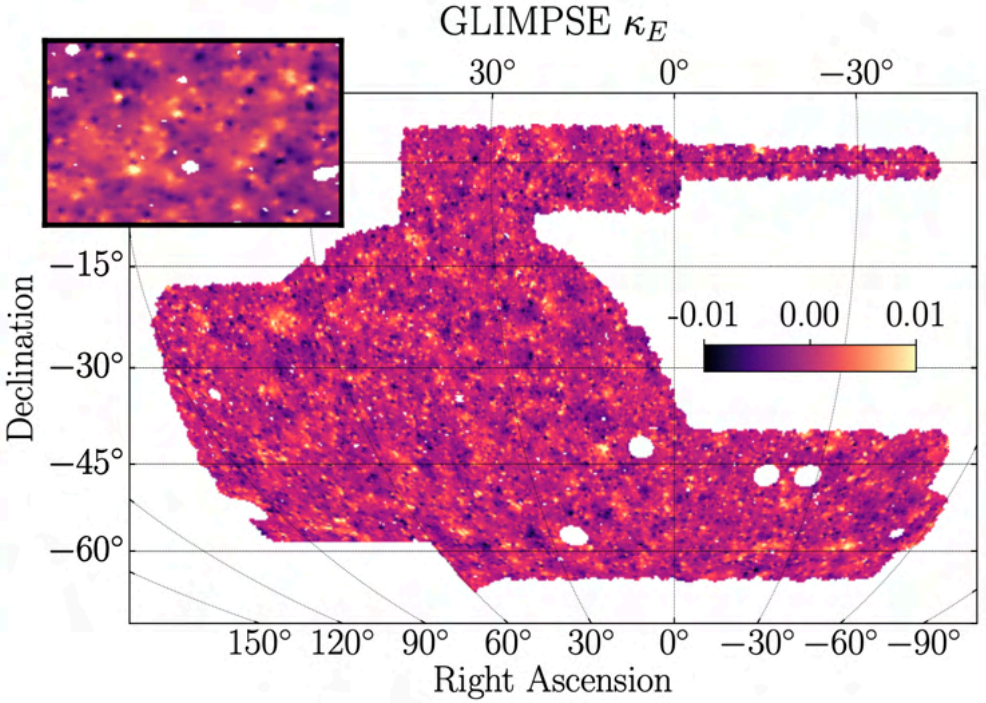
$$\log p(\text{data} \mid \text{map}) + \log p(\text{map})$$

Results









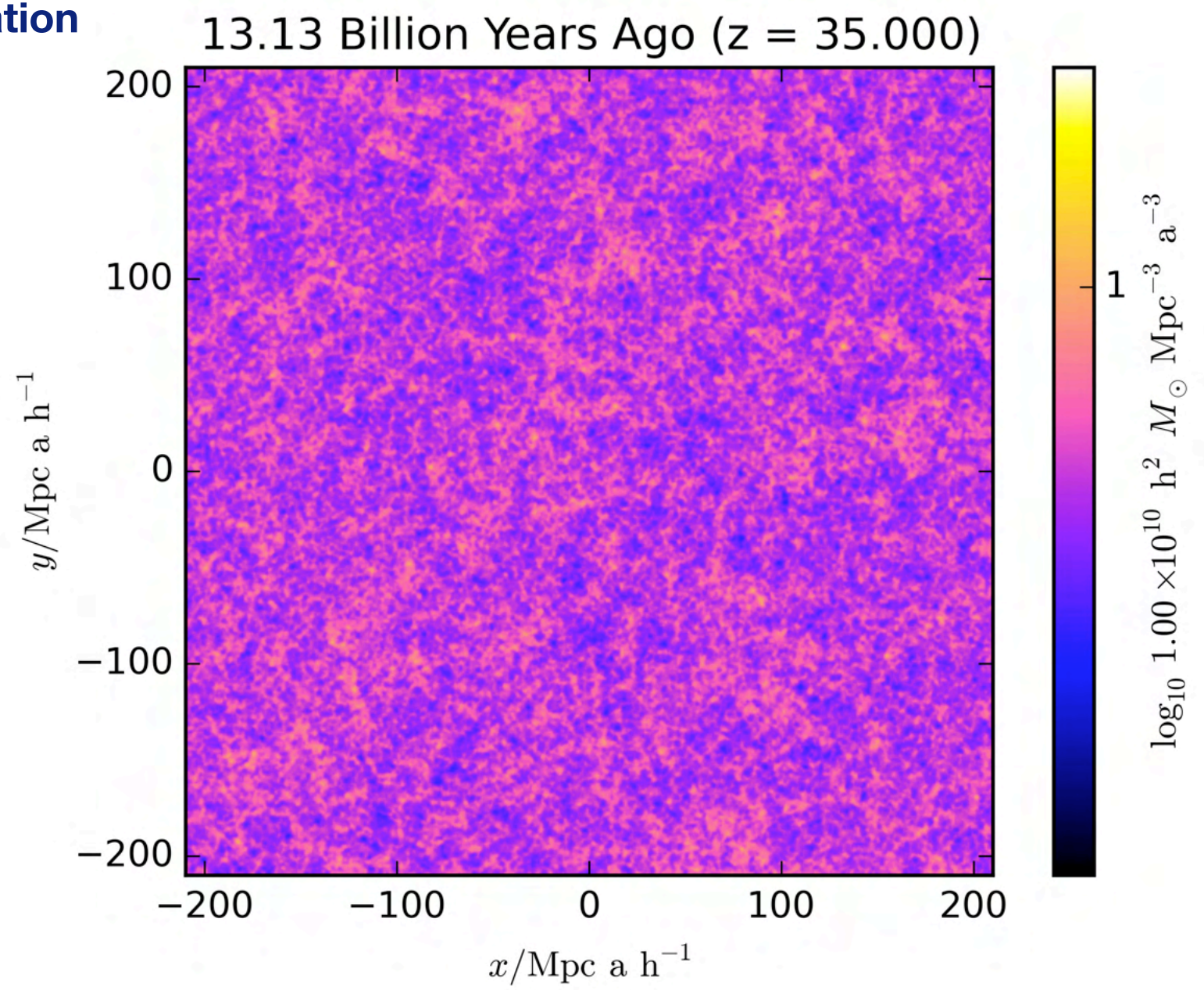
THE DARK ENERGY SURVEY

02

Understanding the Universe

The Cosmic Web

Cosmology simulation

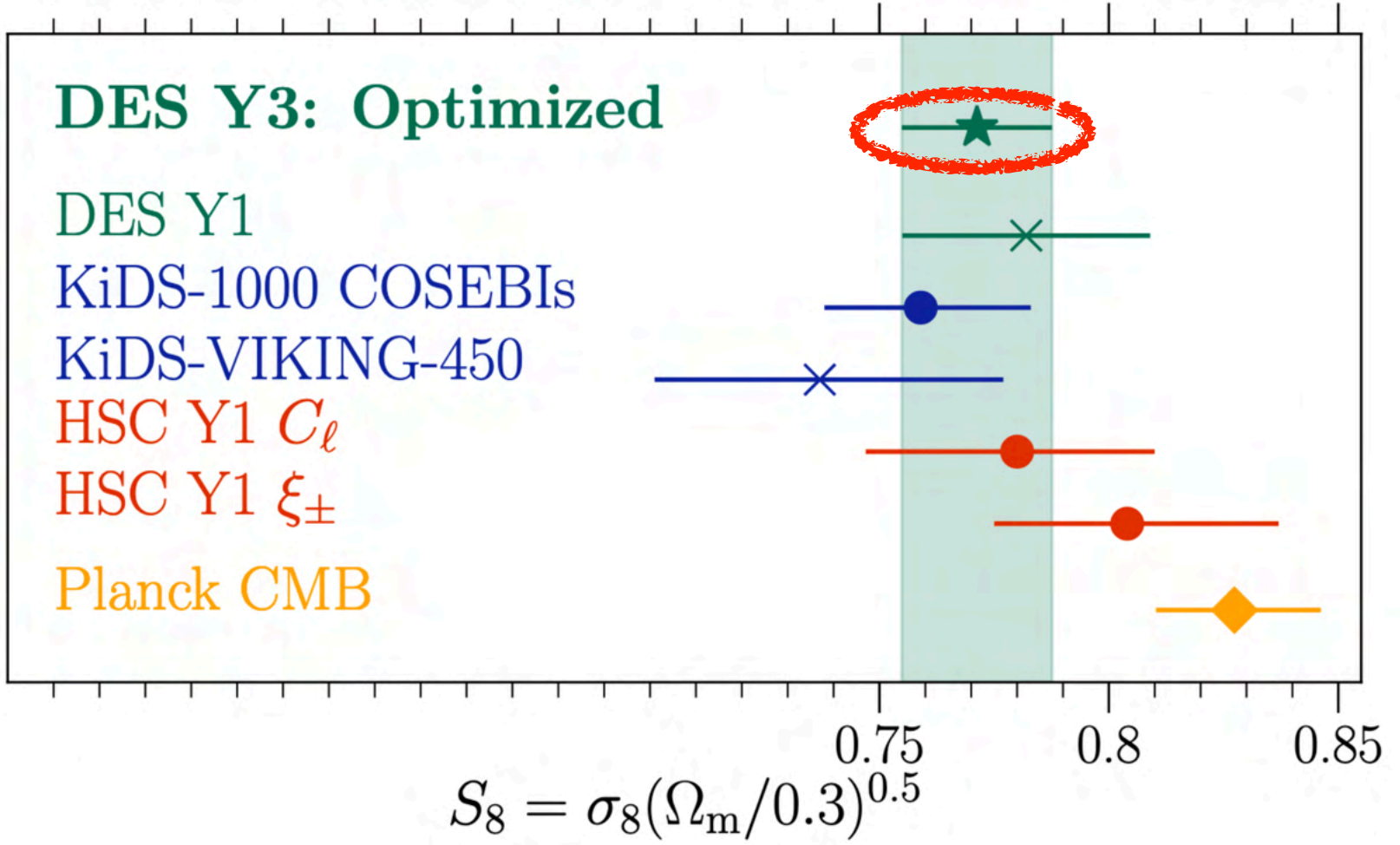


THE DARK ENERGY SURVEY

Dark Energy Survey

THE DARK ENERGY SURVEY

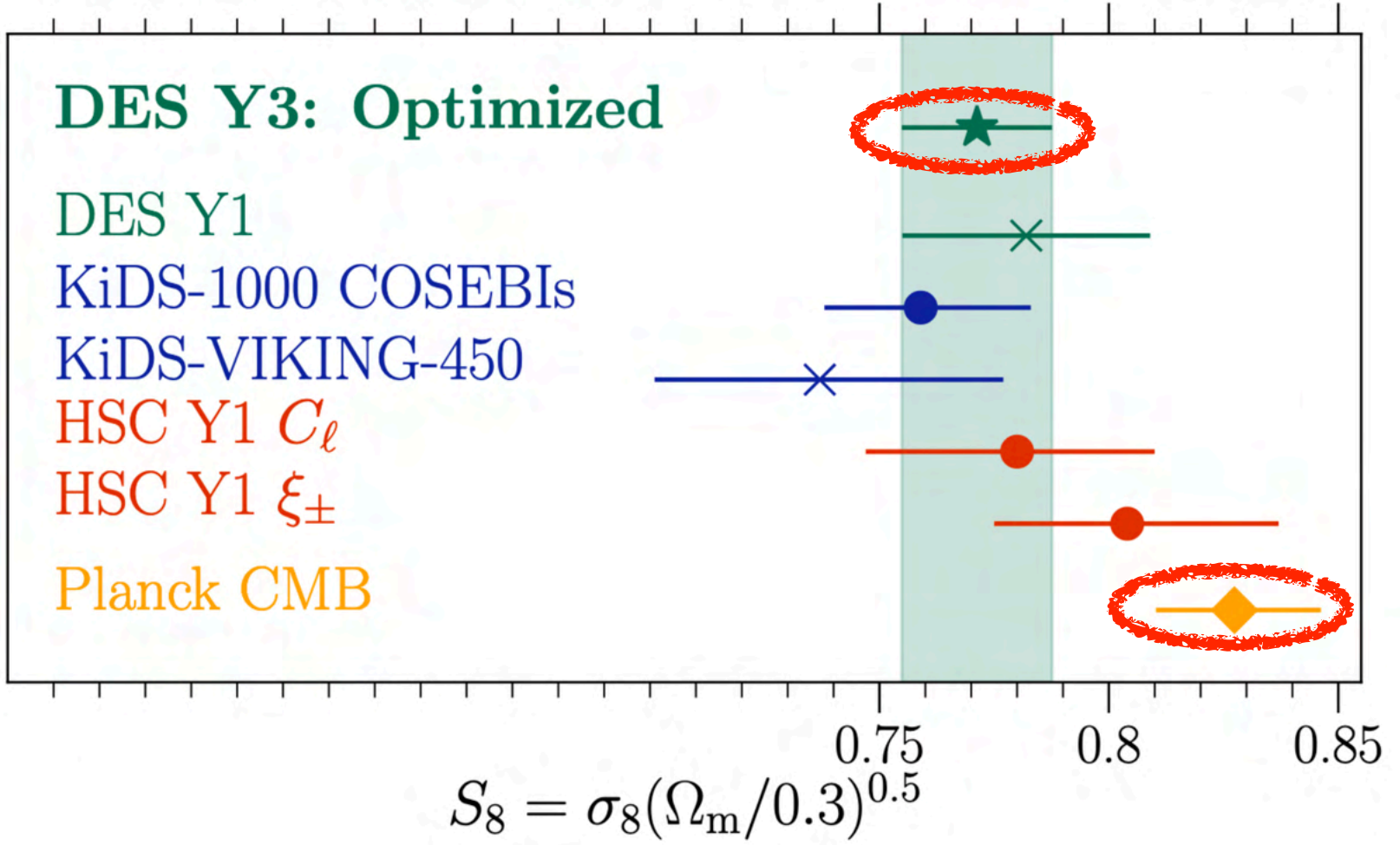
Lensing is low?



amount of "clumpiness"

THE DARK ENERGY SURVEY

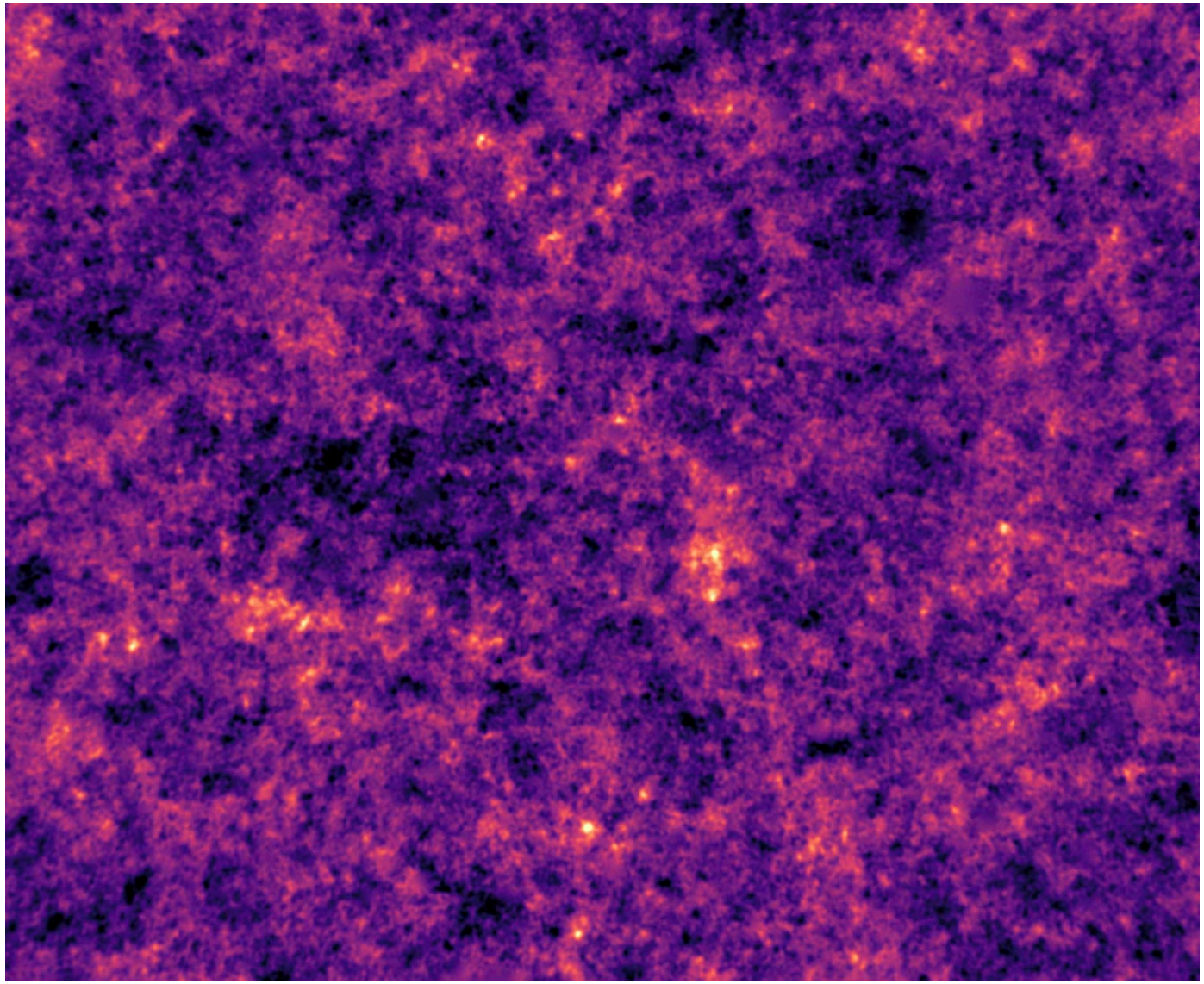
Lensing is low?



amount of "clumpiness"

THE DARK ENERGY SURVEY

Dark matter map:

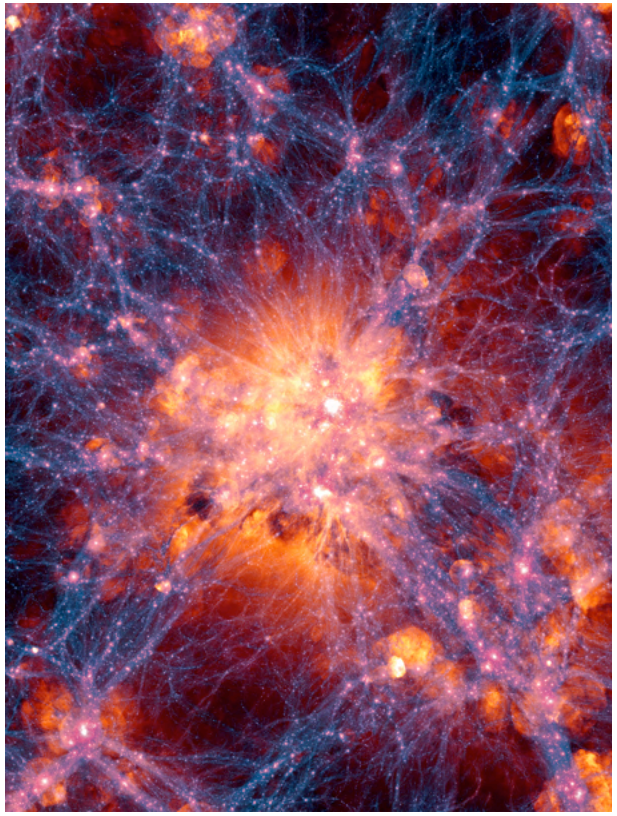


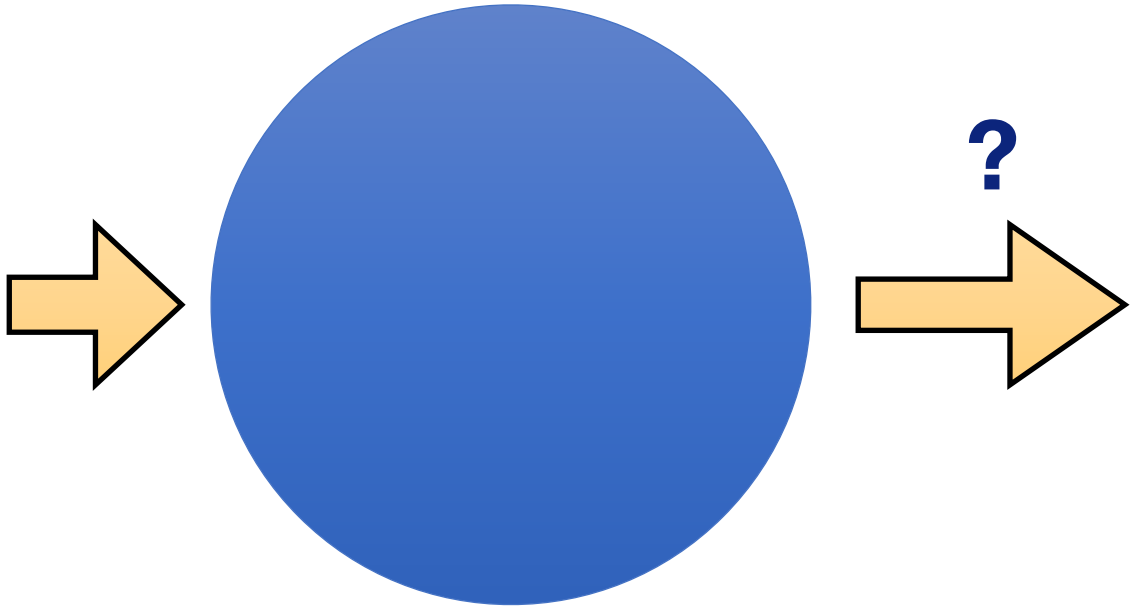
THE DARK ENERGY SURVEY

03 Simulation-based inference

Use Deep Learning to infer cosmology?

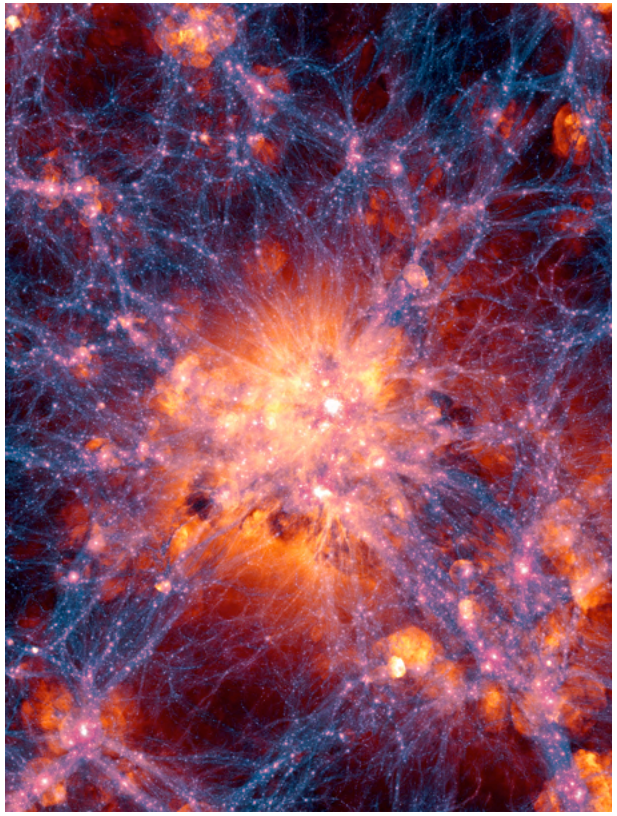
Use Deep Learning to infer cosmology?





THE DARK ENERGY SURVEY

Use Deep Learning to infer cosmology?



?

$\Omega_\Lambda = 0.68$

Dark Energy fraction

THE DARK ENERGY SURVEY

Uncertainty about what we have learned?

$$\Omega_\Lambda = 0.68$$

Dark Energy fraction

THE DARK ENERGY SURVEY

Parameter inference

1. Observed “data” d_o

2. Unknown parameters θ of a given model

Parameter inference

1. Observed “data” d_o

2. Unknown parameters θ of a given model

$$p(\theta \mid d_o) \propto p(d_o \mid \theta)\ p(\theta)$$

Likelihood-free inference

$$\{\mathbf{d}_i, \theta_i\}$$

I. Draw $\mathbf{d}_i$ from the distribution $p(\mathbf{d} \,|\, \theta_i)$ by running a simulation

Likelihood-free inference

$$\{\mathbf{d}_i, \theta_i\}$$

I. Draw $\mathbf{d}_i$ from the distribution $p(\mathbf{d} \mid \theta_i)$ by running a simulation

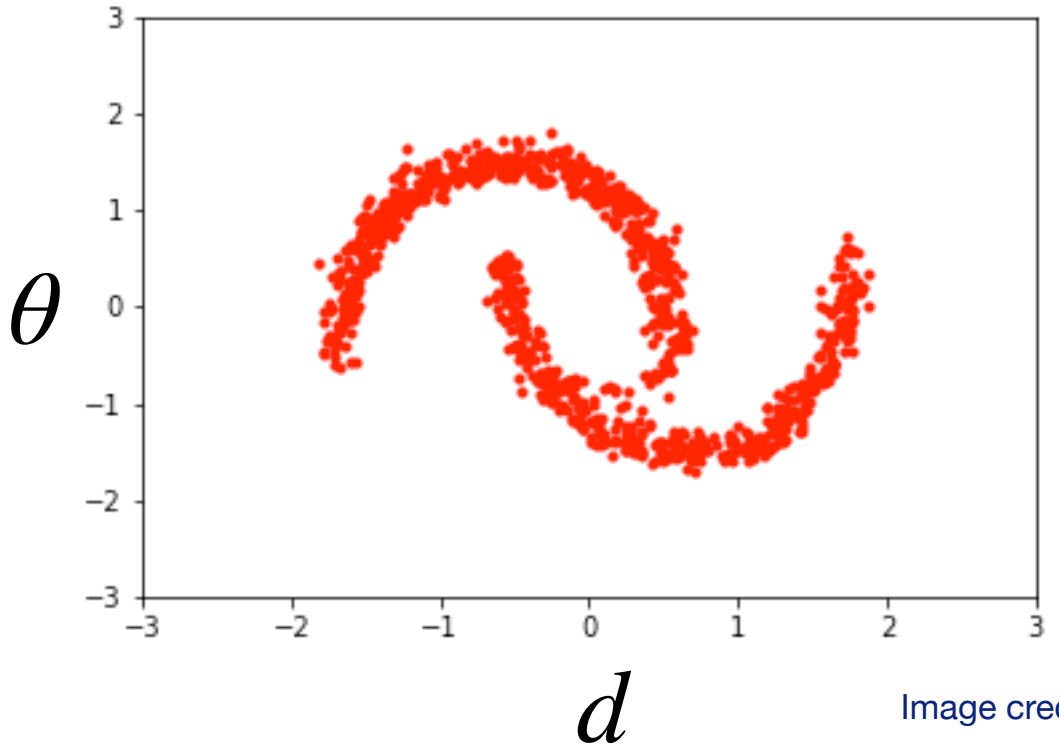
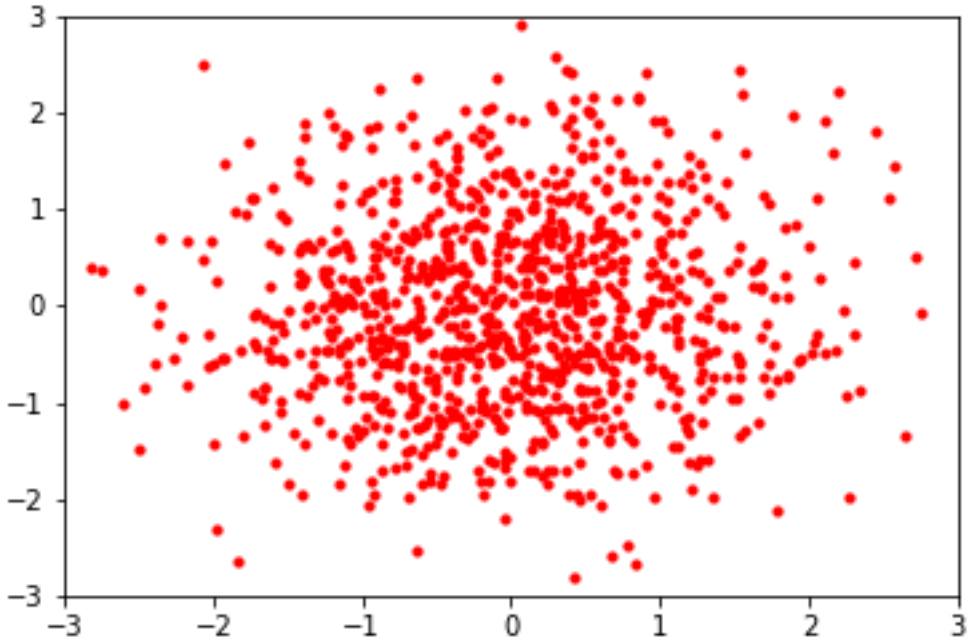


Image credit: Eric Jang

THE DARK ENERGY SURVEY

Neural density estimation

Normalizing flow from simple known distribution



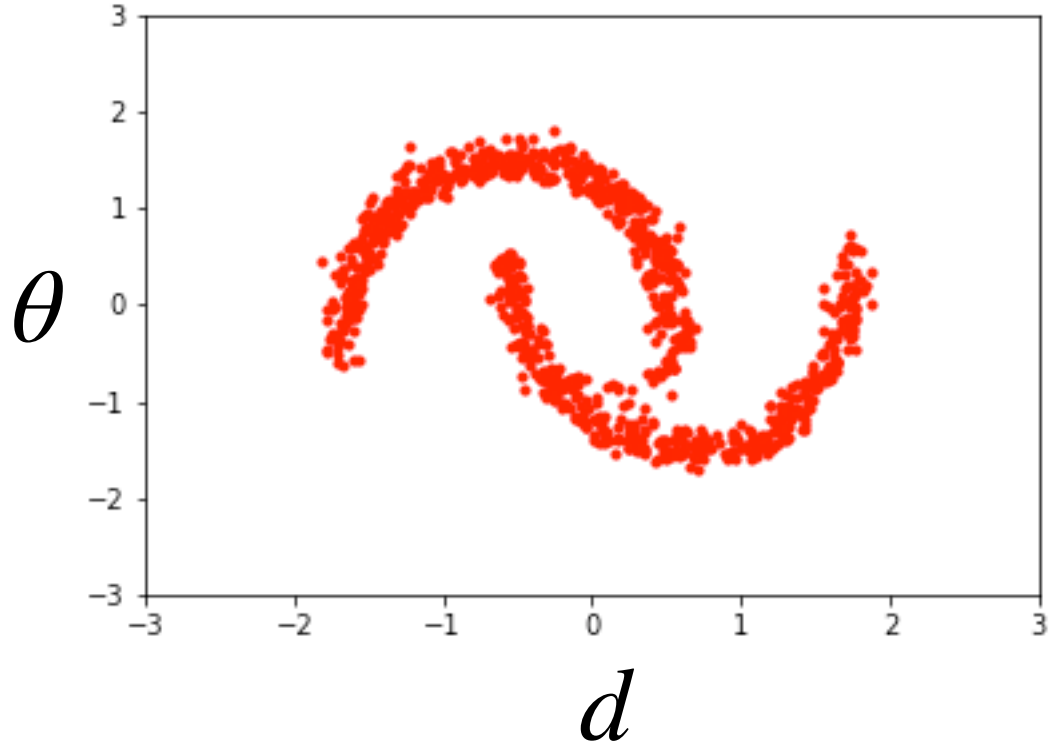
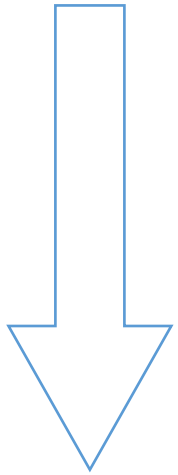


Image credit: Eric Jang

THE DARK ENERGY SURVEY

Likelihood-free inference with the Dark Energy Survey



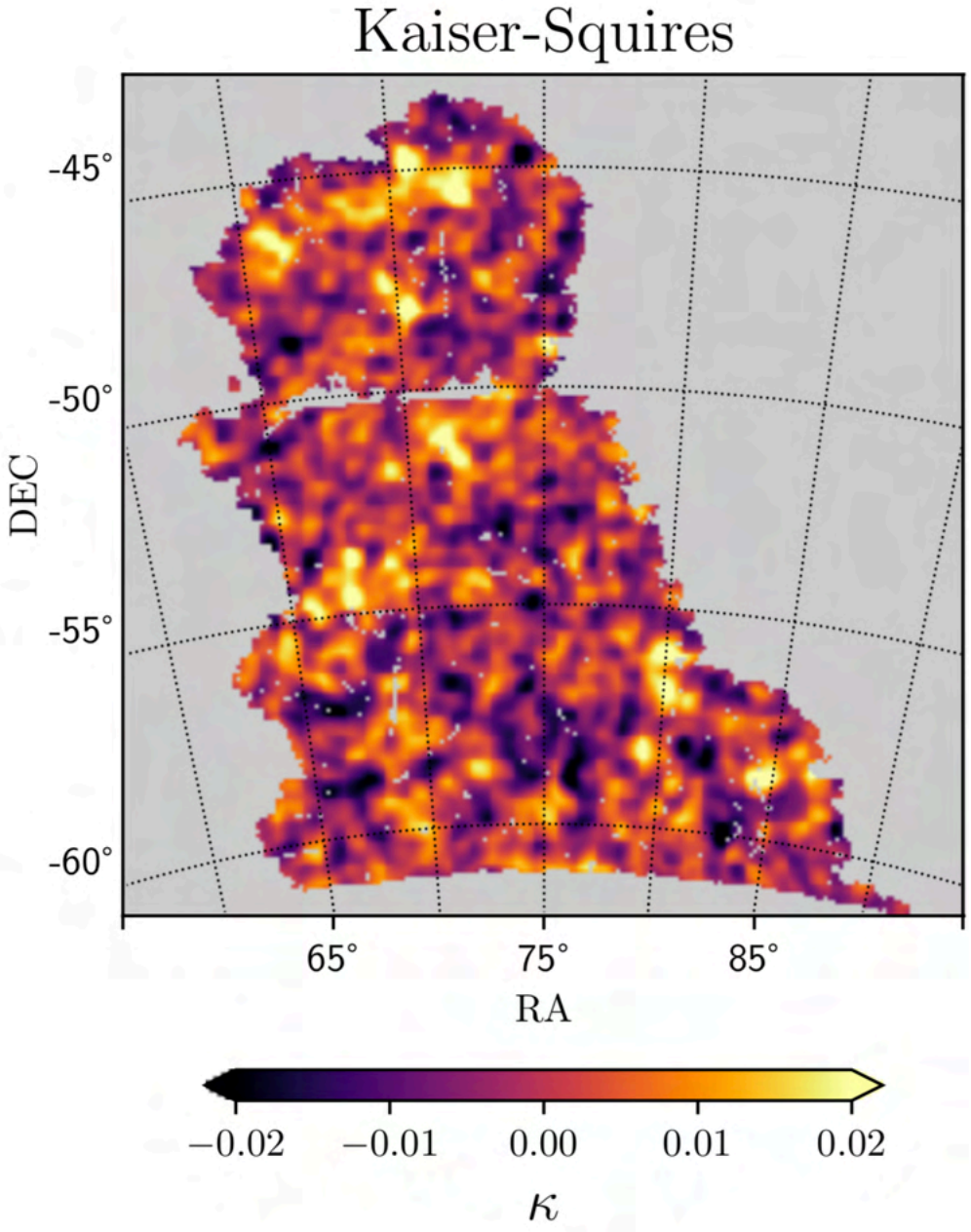
I. Simulations

II. Likelihood-free inference

III. Posterior probability for models given DES data

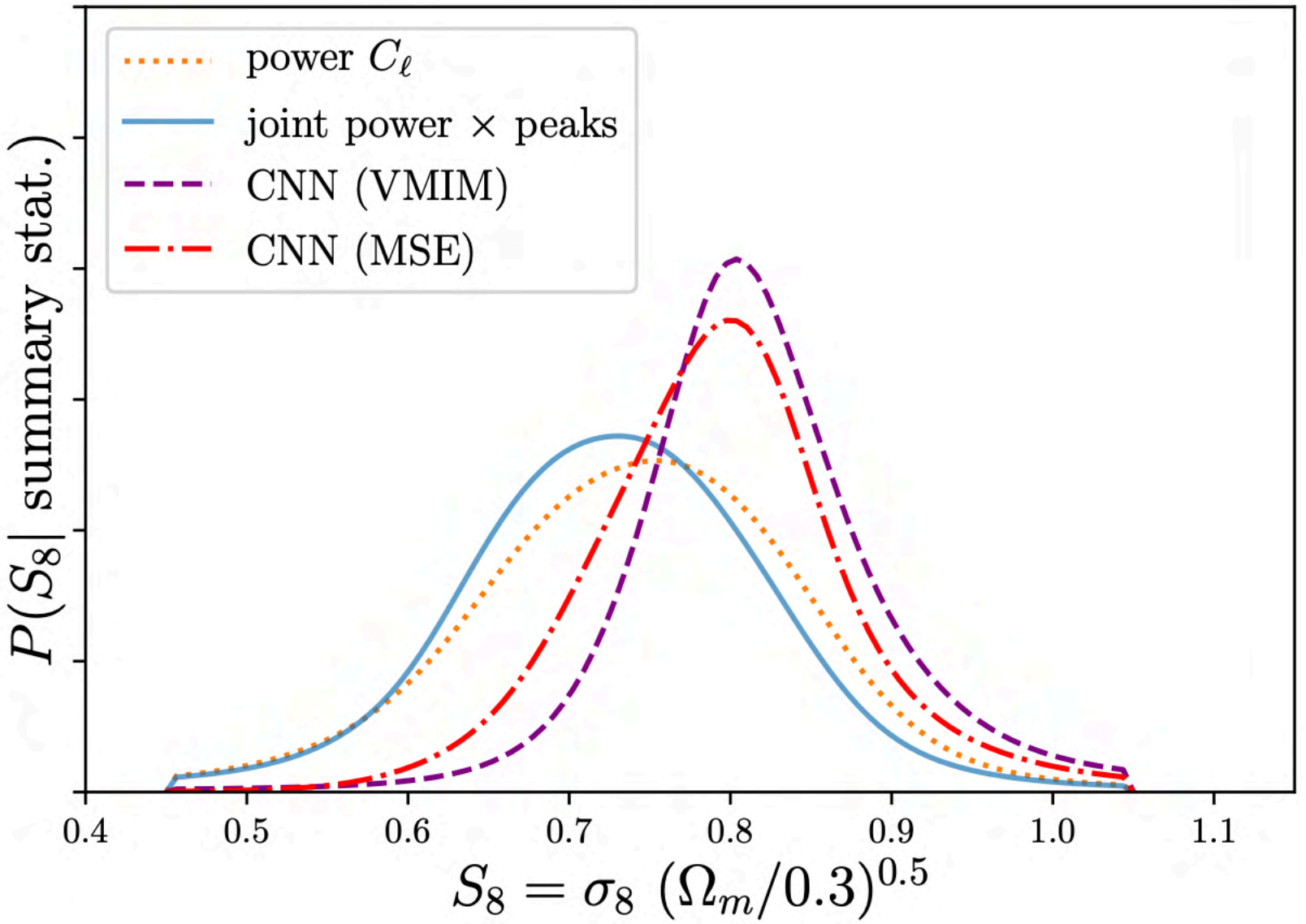
THE DARK ENERGY SURVEY

DES Science Verification data



LPENS
LABORATOIRE DE PHYSIQUE
DE L'ÉCOLE NORMALE SUPÉRIEURE

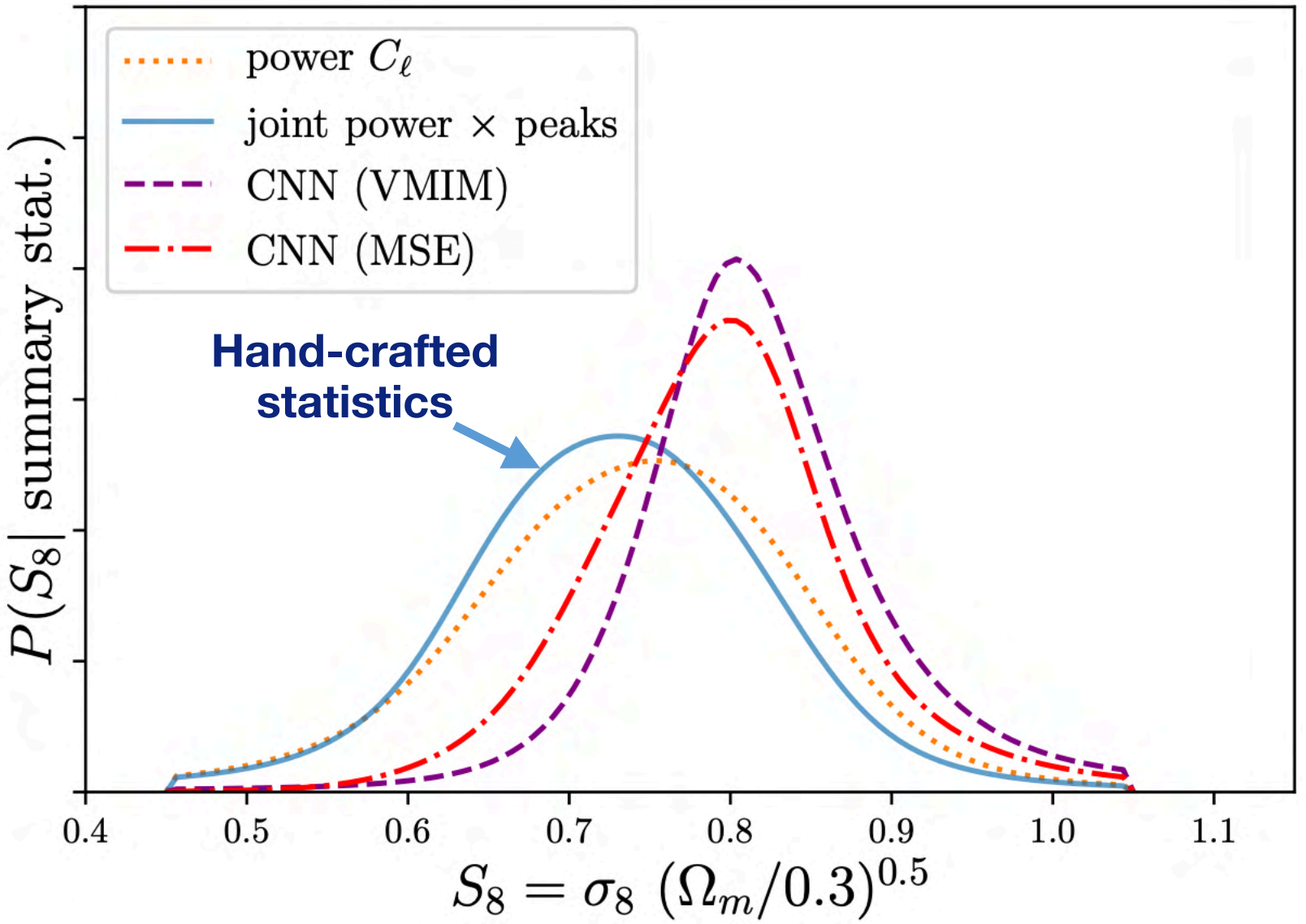
DES Science Verification results:



amount of "clumpiness"

LPENS
LABORATOIRE DE PHYSIQUE DE L'ÉCOLE NORMALE SUPÉRIEURE

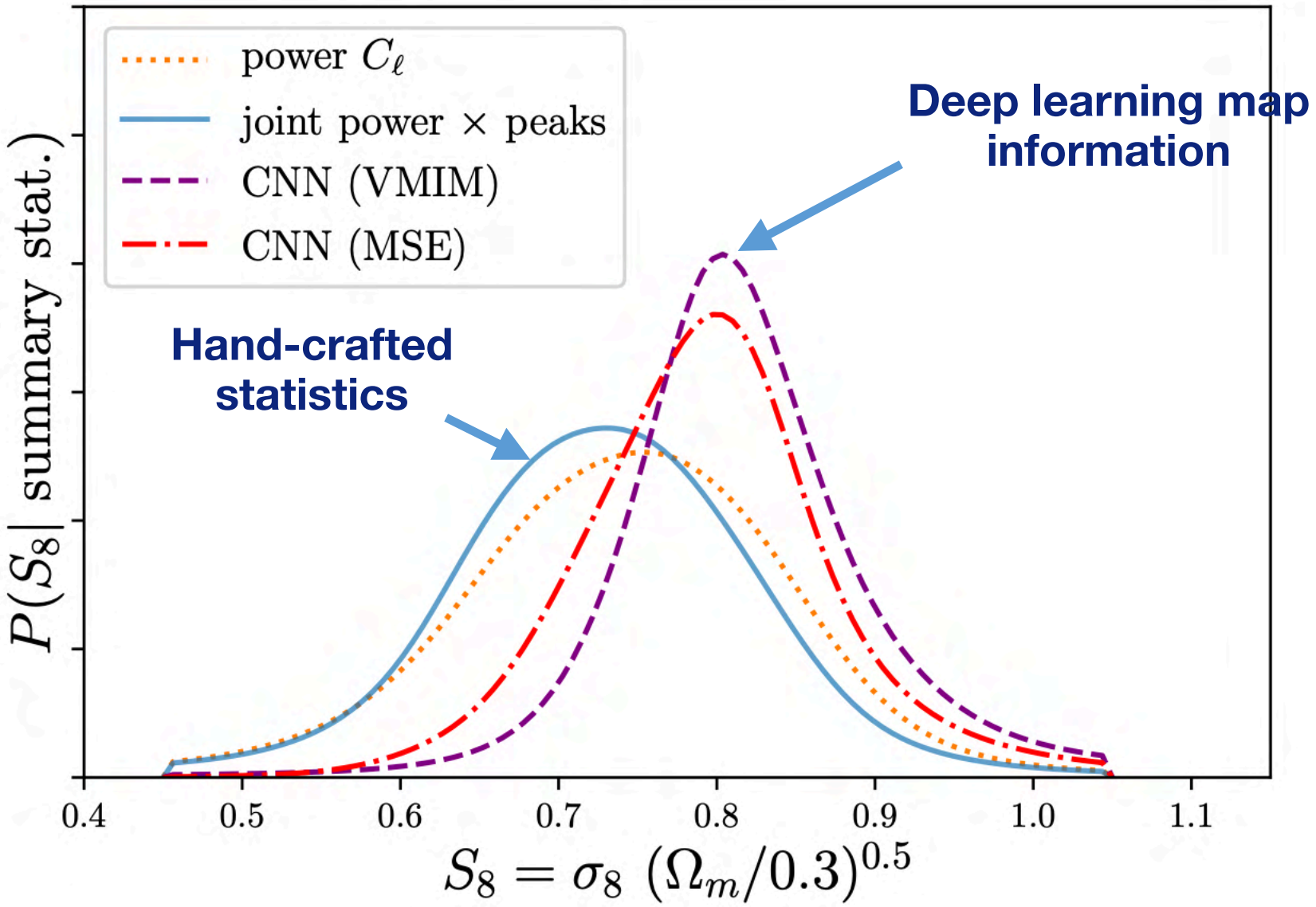
DES Science Verification results:



amount of "clumpiness"

LPENS
LABORATOIRE DE PHYSIQUE DE L'ÉCOLE NORMALE SUPÉRIEURE

DES Science Verification results:



amount of "clumpiness"

LPENS
LABORATOIRE DE PHYSIQUE DE L'ÉCOLE NORMALE SUPÉRIEURE

Likelihood-free inference with the Dark Energy Survey

I. Sample unknown cosmological parameters as input

II. Nbody dark matter only simulations: PKDGRAV3 code

III. Require ~800 full simulations

PKDGRAV code

I. Fast Multipole Method (FMM) Nbody highly optimised for GPU acceleration.

II. Fast Multipole Method Nbody highly optimised for GPU

THE DARK ENERGY SURVEY

DiRAC allocation

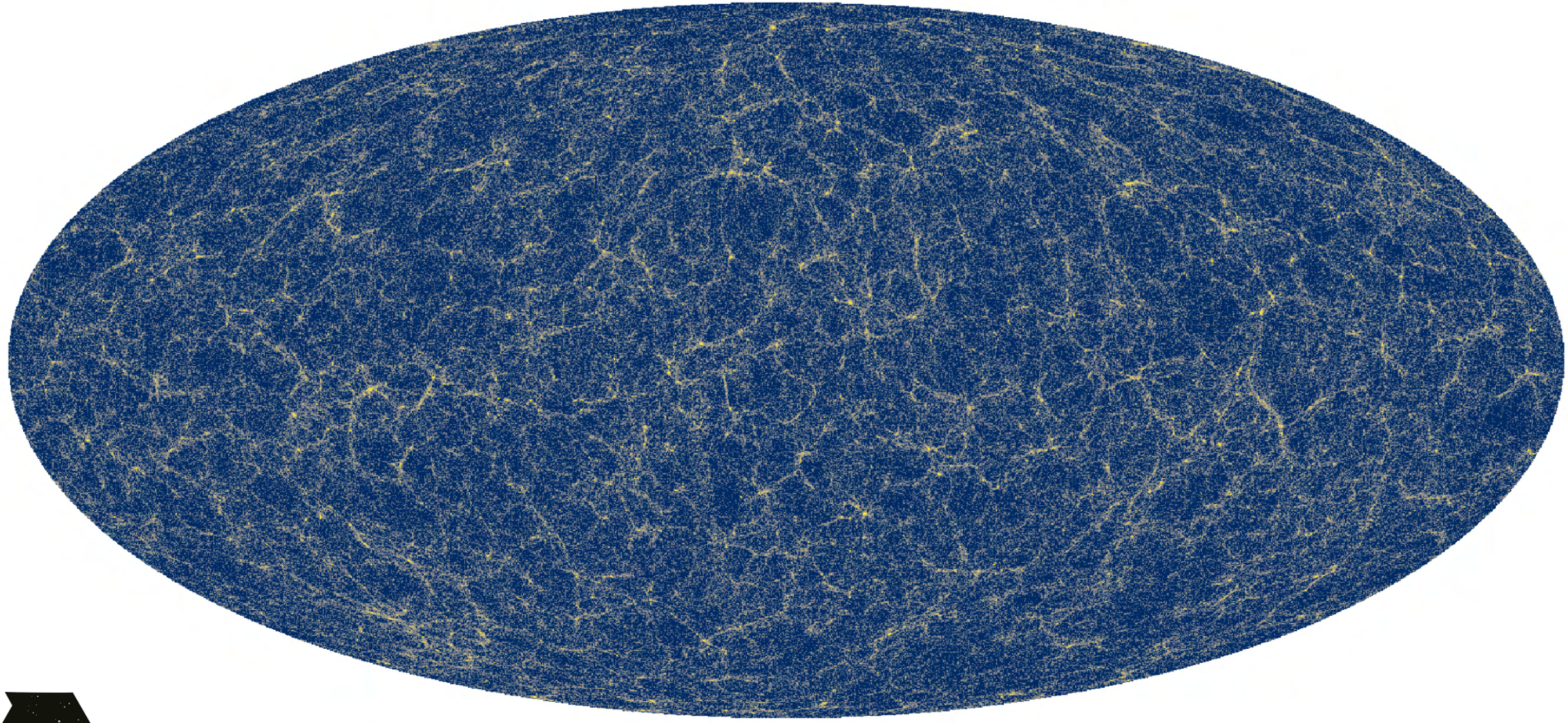
I. 47k GPUhrs initial allocation on Wilkes-2 (Cambridge) - NVIDIA A100

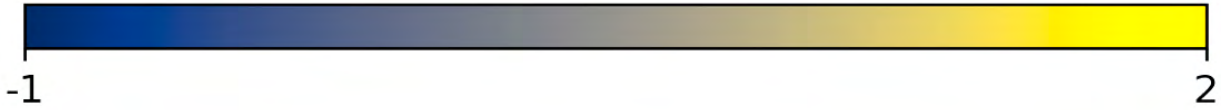
II. Additional uplift of 16k GPUhrs Extreme Scaling Service (Edinburgh)

III. Extremely parallel (one simulation per GPU) and IO efficient

THE DARK ENERGY SURVEY

PKDGRAV lightcone z=0.1: $\log_{10}\delta+1$



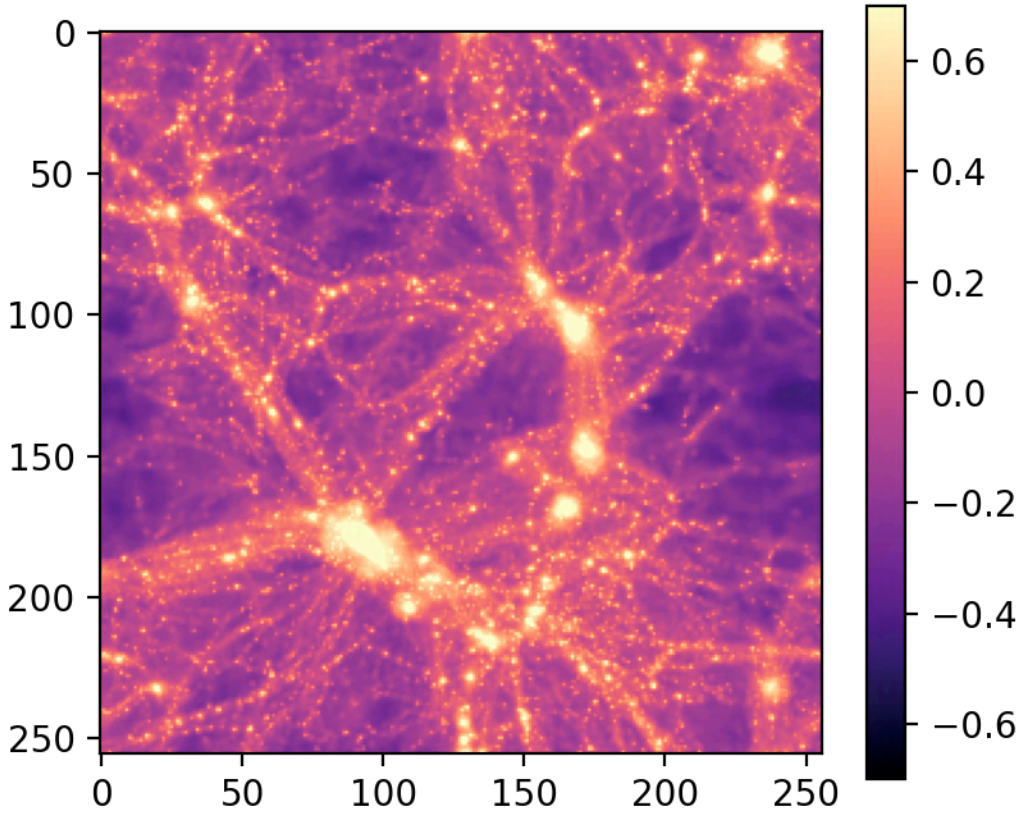


THE DARK ENERGY SURVEY

04 Future hopes with AI

Improving physics simulations with WPH synthesis

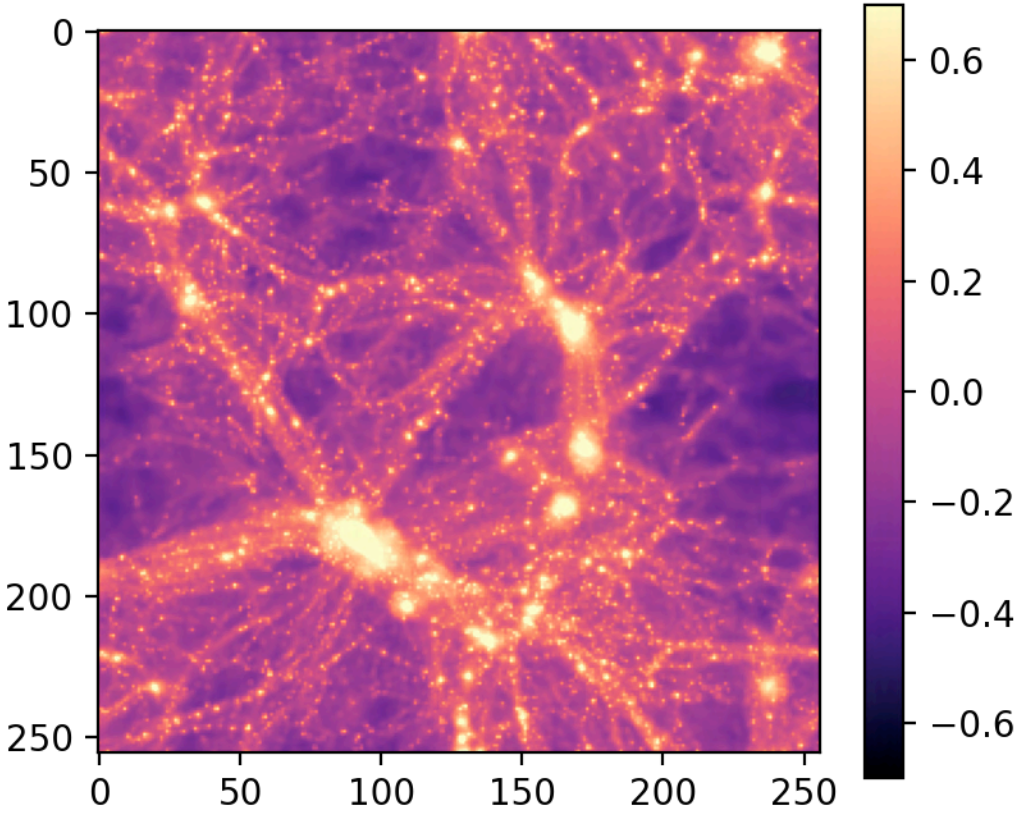
Dark matter



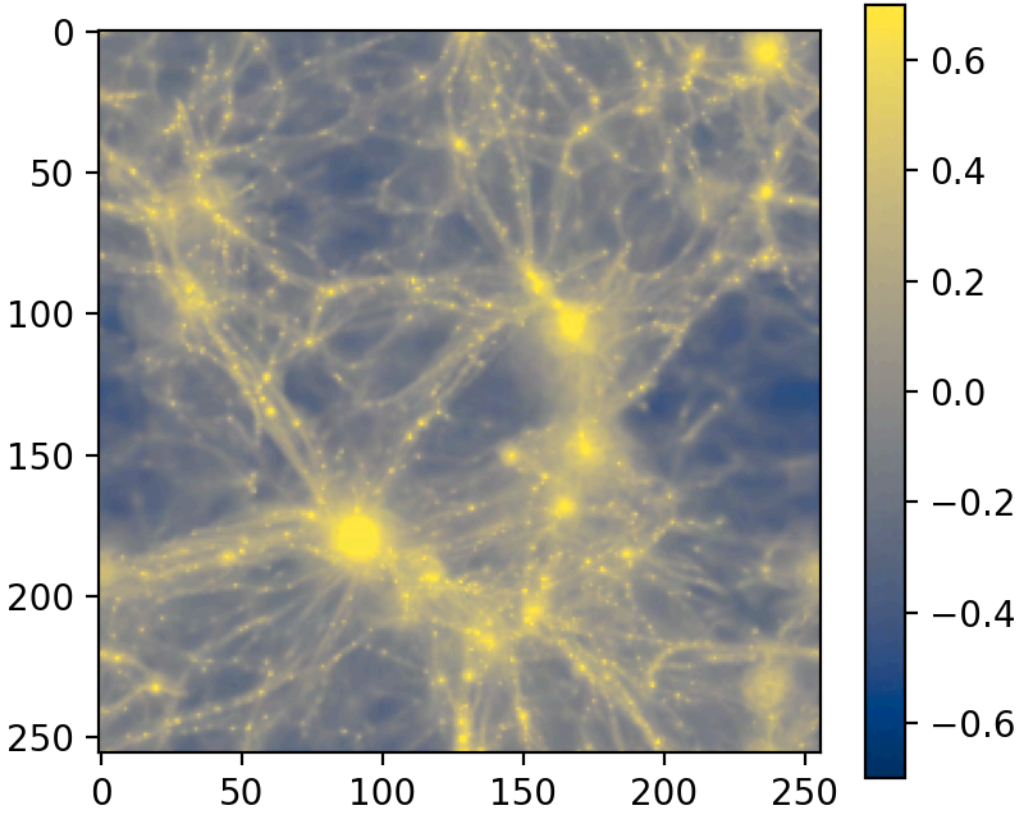
LPENS
LABORATOIRE DE PHYSIQUE
DE L'ÉCOLE NORMALE SUPÉRIEURE

Improving physics simulations with WPH synthesis

Dark matter

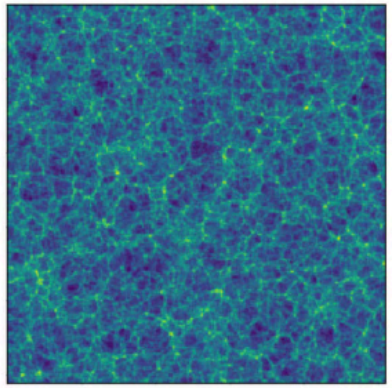


Gas



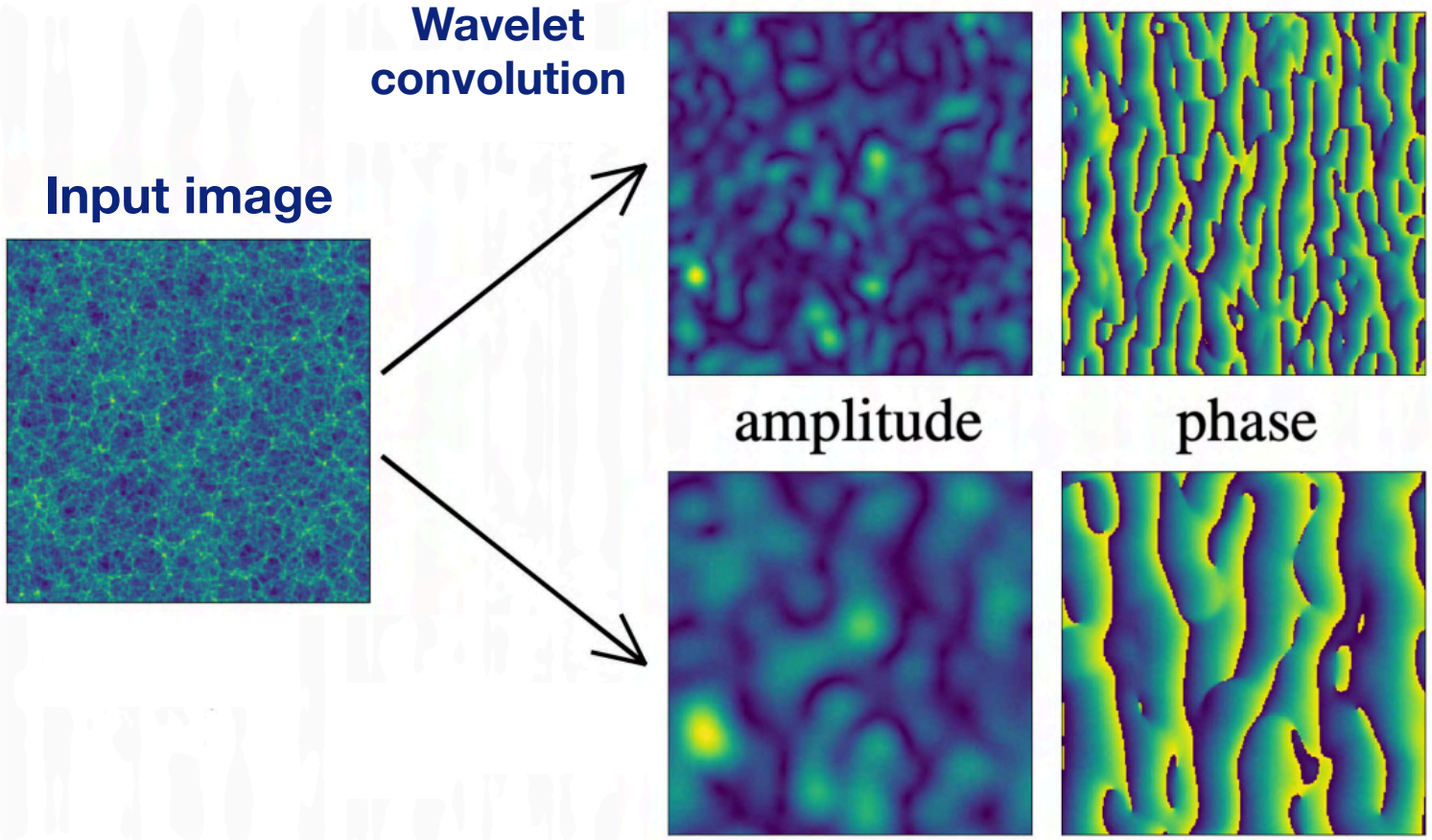
Wavelet Phase Harmonics
(Mallat, Zhang, Rochette 1810.12136)

Input image



Wavelet Phase Harmonics

(Mallat, Zhang, Rochette 1810.12136)

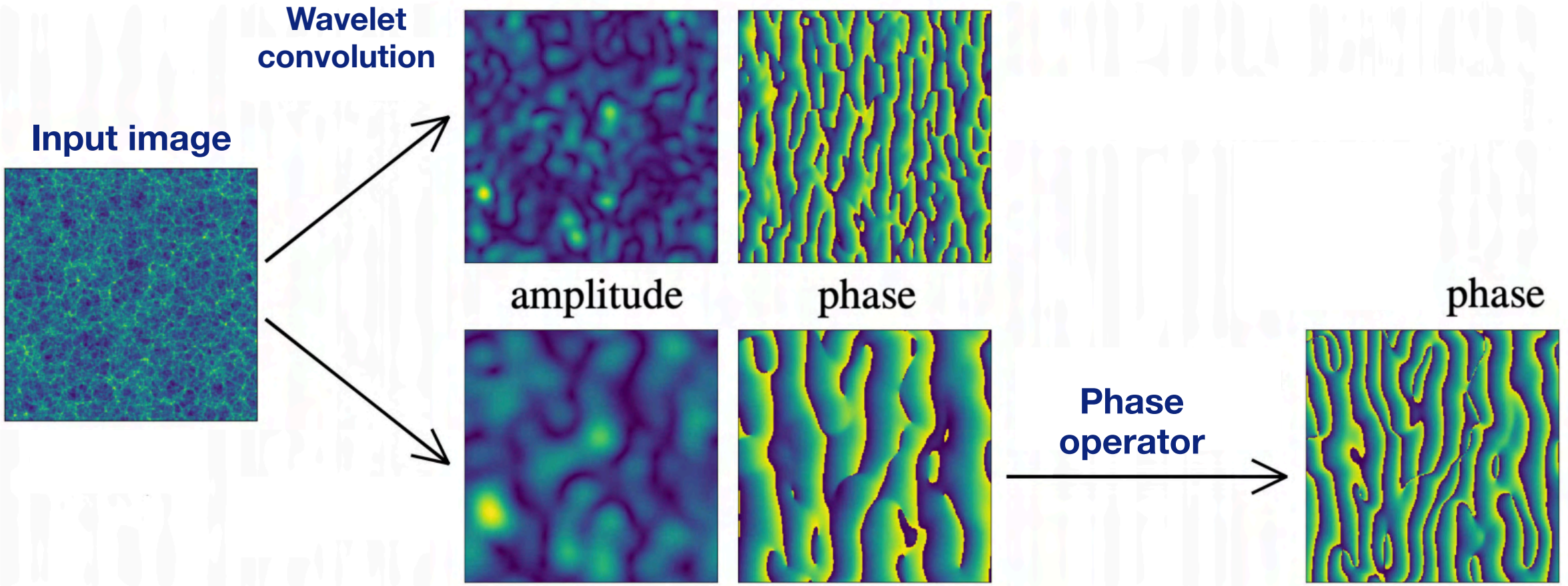


(adapted from Allys++ 2006.06298)

LPENS
LABORATOIRE DE PHYSIQUE DE L'ÉCOLE NORMALE SUPÉRIEURE

Wavelet Phase Harmonics

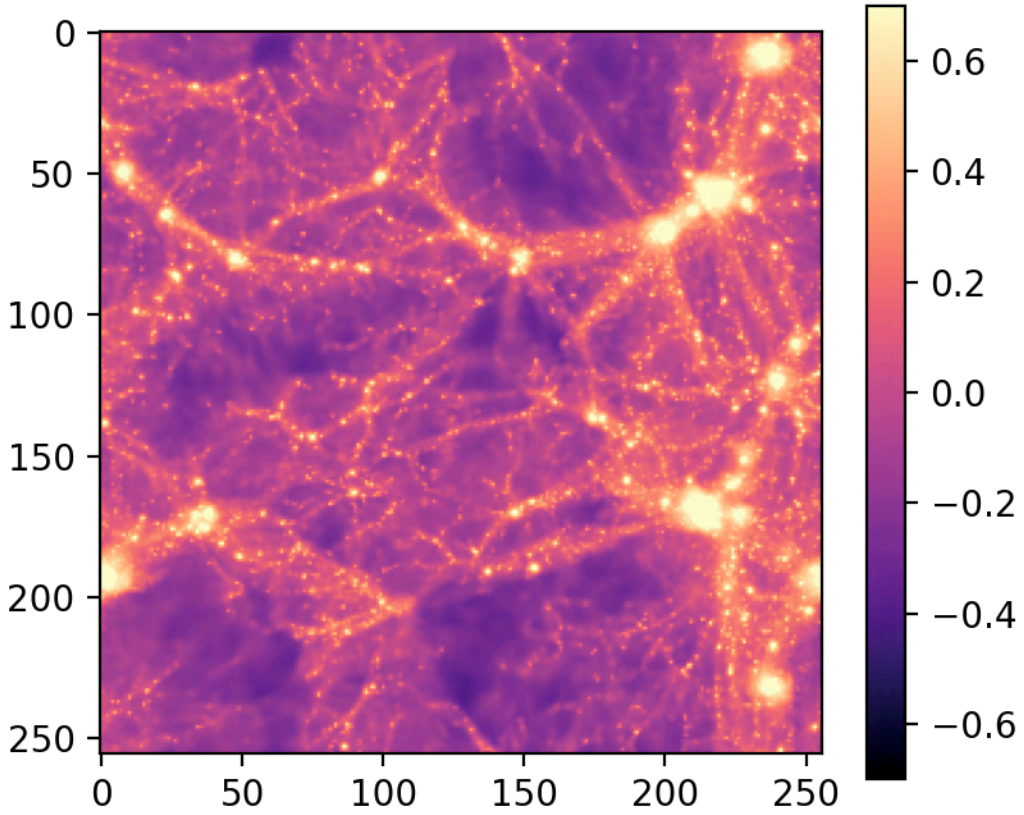
(Mallat, Zhang, Rochette 1810.12136)



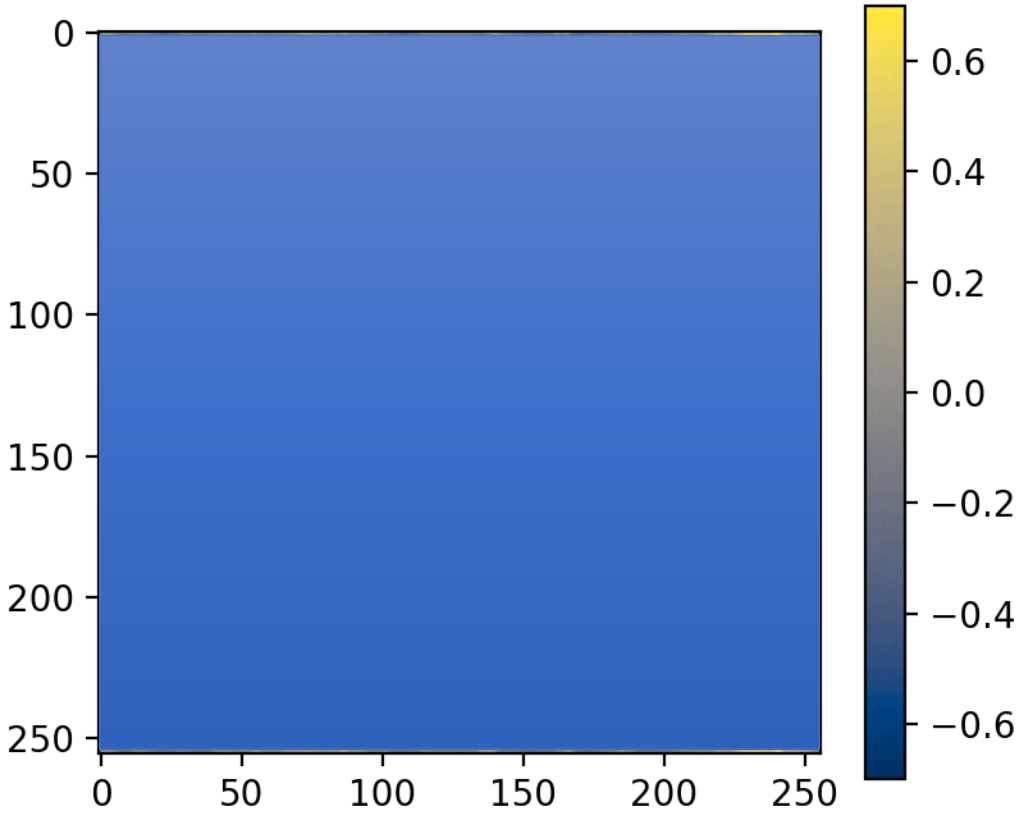
(adapted from Allys++ 2006.06298)

Conditional-WPH synthesis:

Dark matter

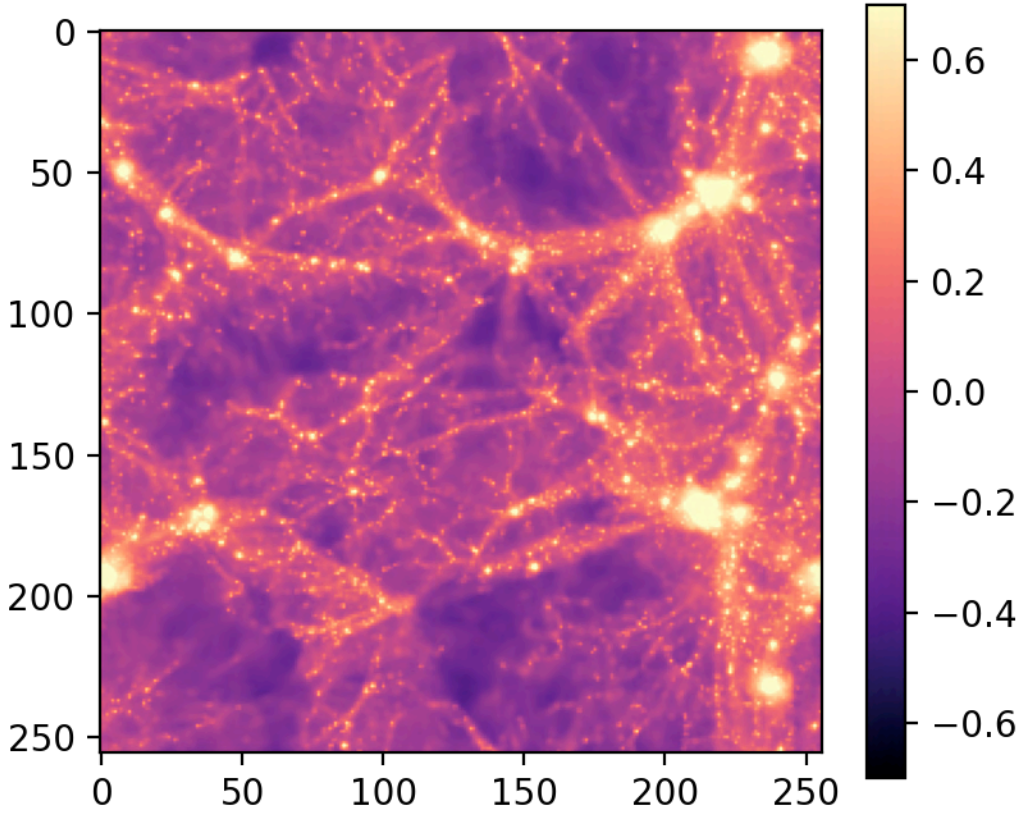


Gas

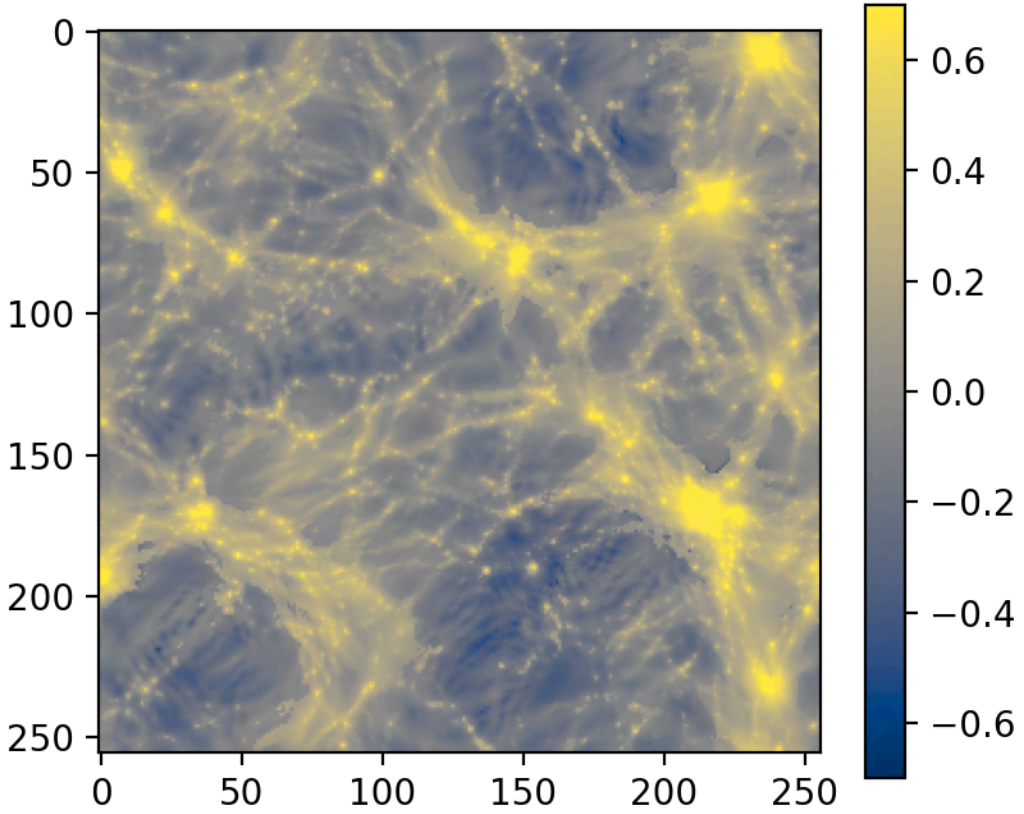


Conditional-WPH synthesis:

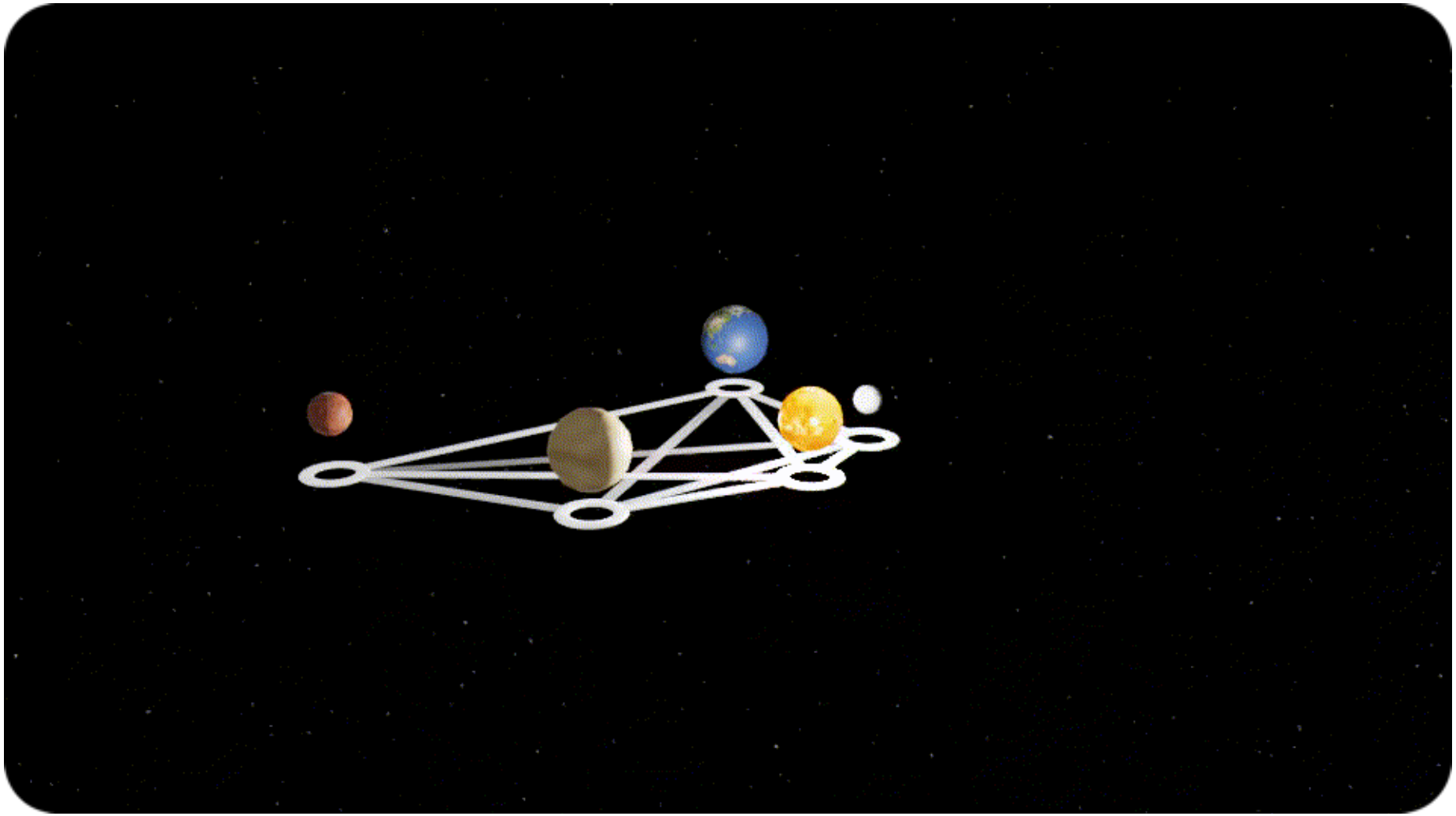
Dark matter



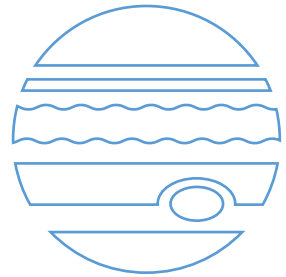
Gas

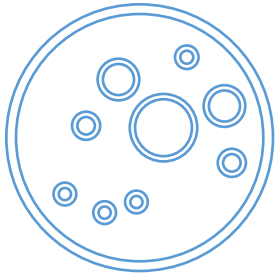


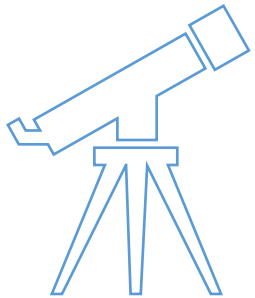
Learning the laws of the Solar System



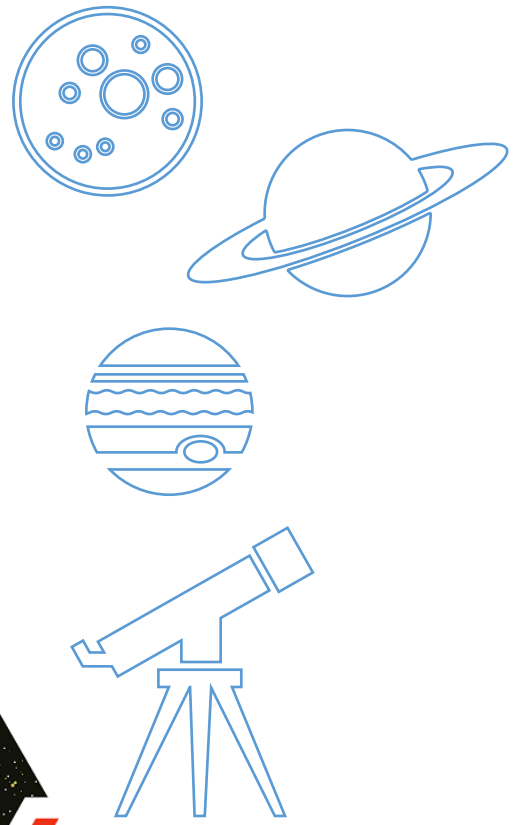
THE DARK ENERGY SURVEY





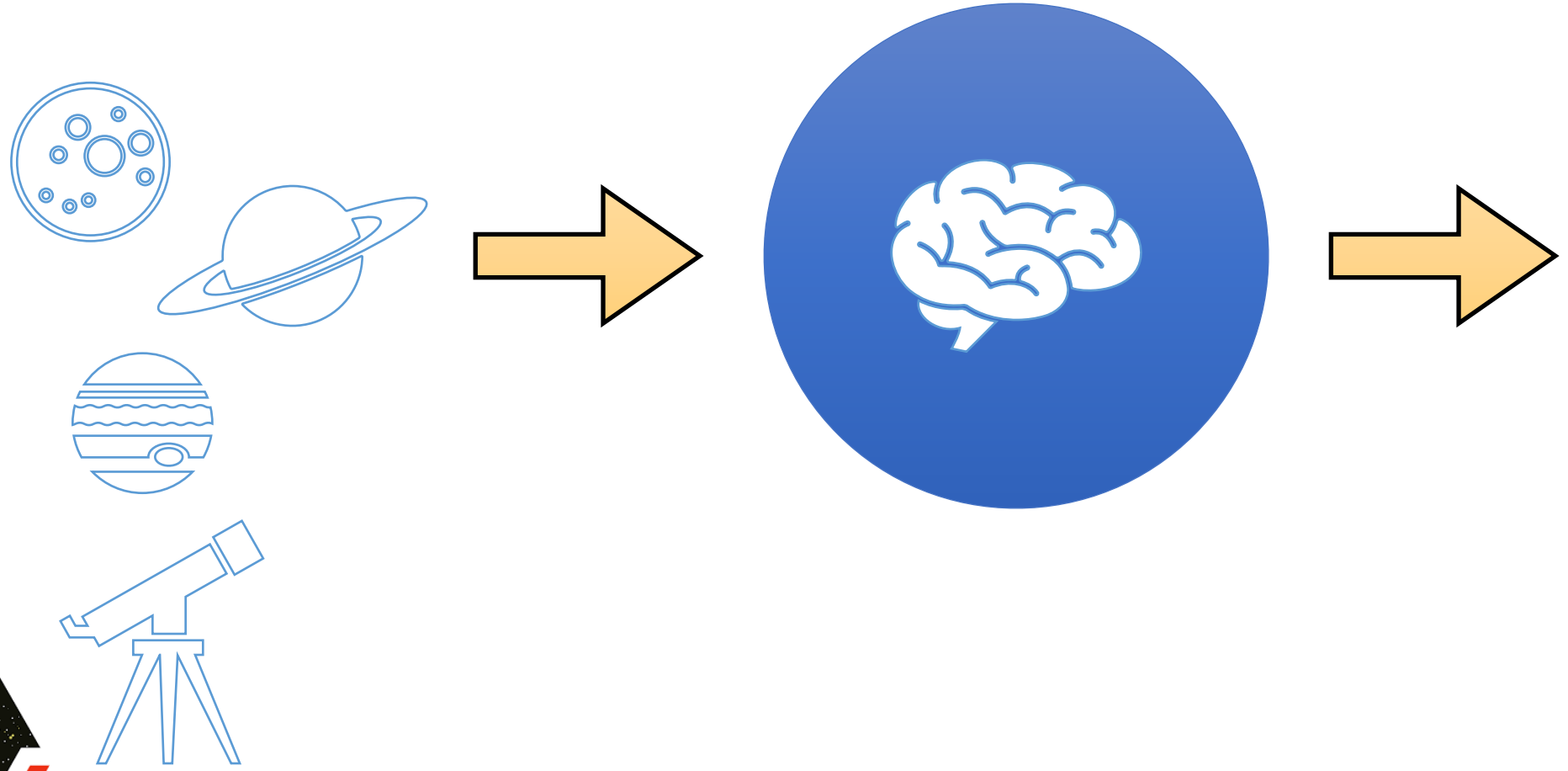


THE DARK ENERGY SURVEY



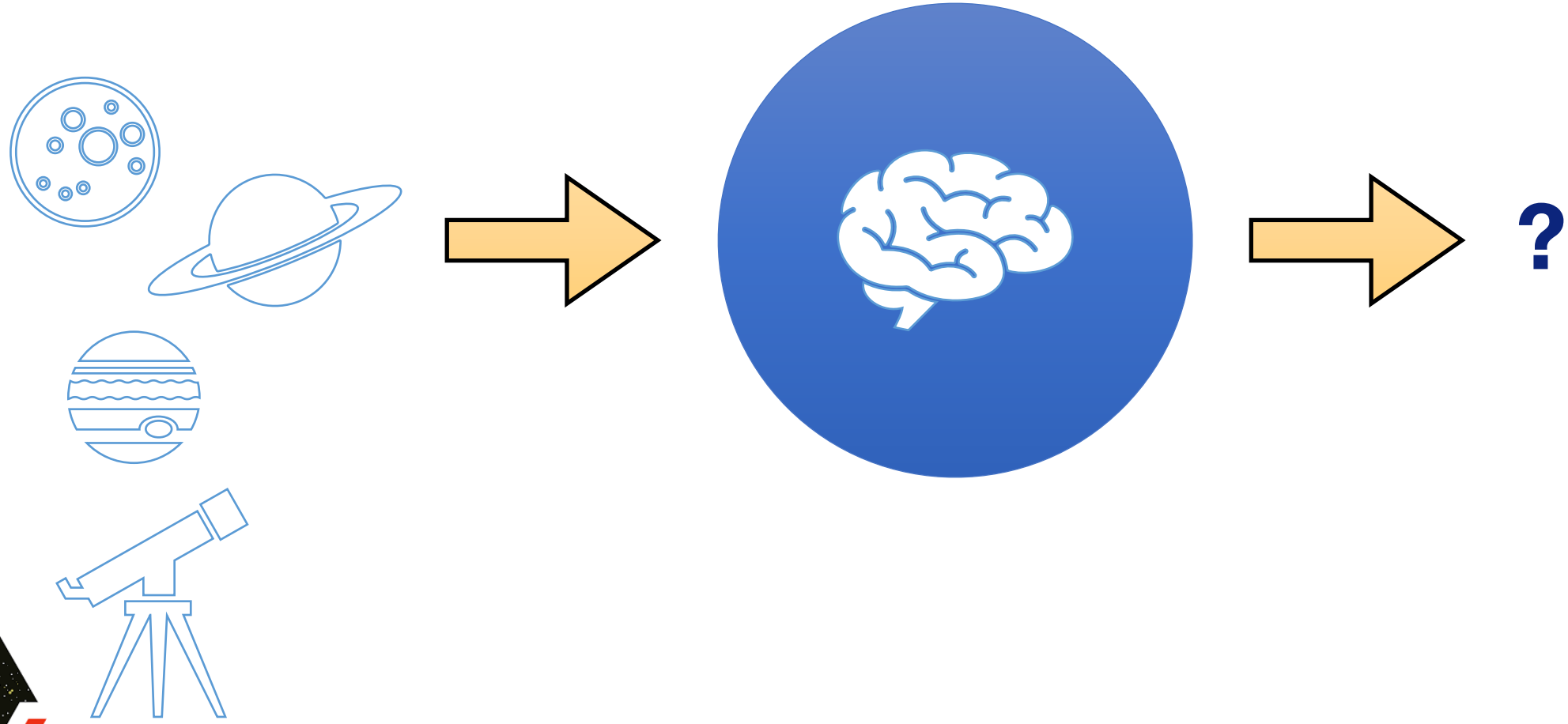
THE DARK ENERGY SURVEY

No simulations…
Learn physical laws



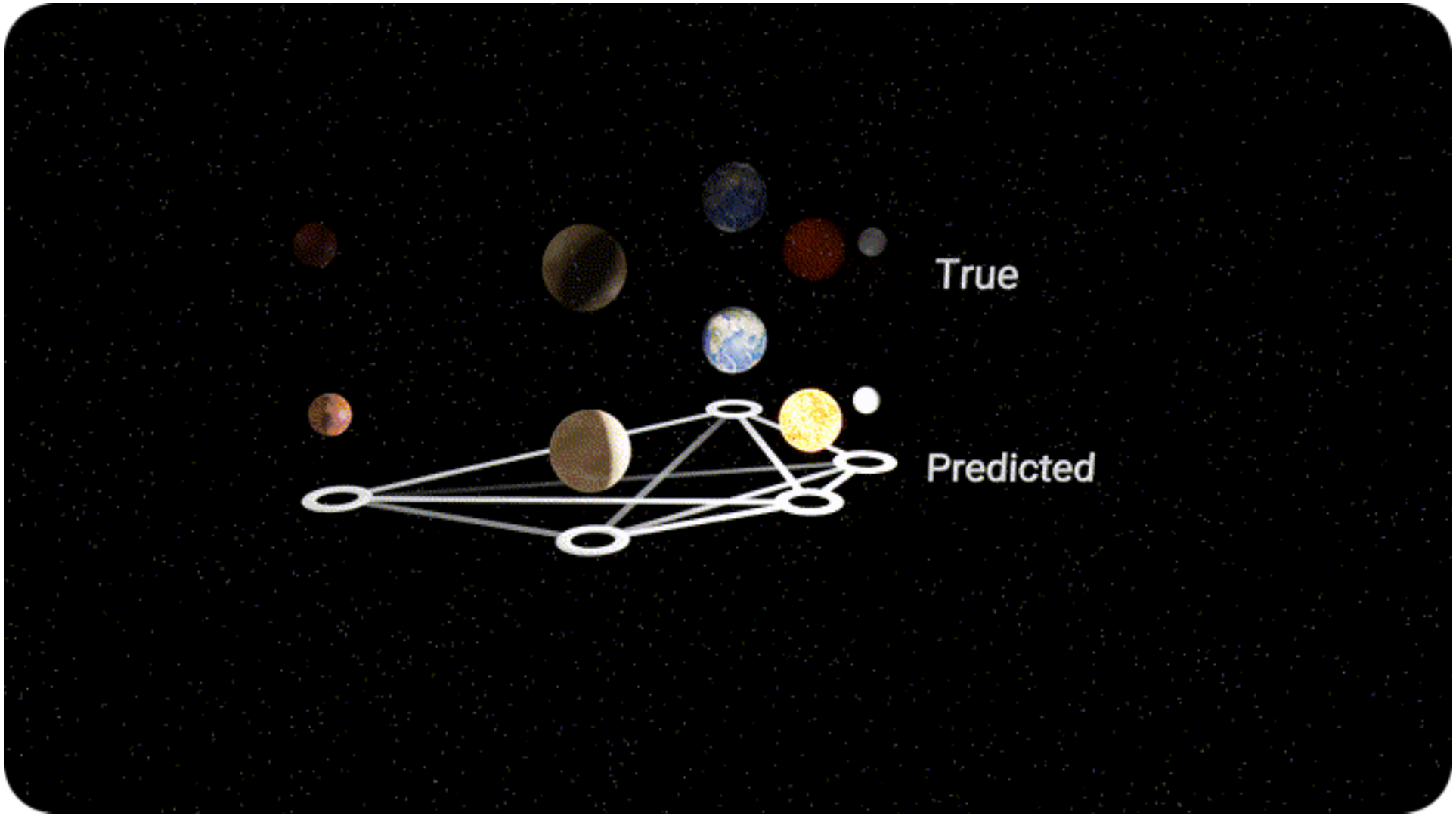
THE DARK ENERGY SURVEY

No simulations…
Learn physical laws

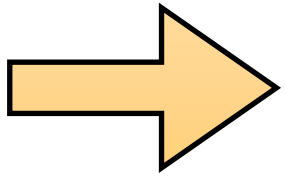


THE DARK ENERGY SURVEY

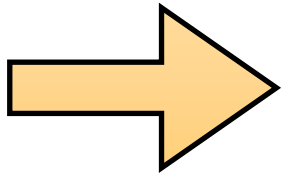
Learning the laws of the Solar System

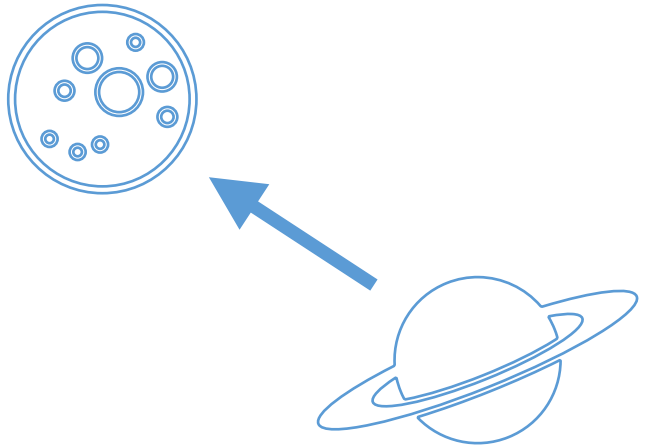


THE DARK ENERGY SURVEY



?

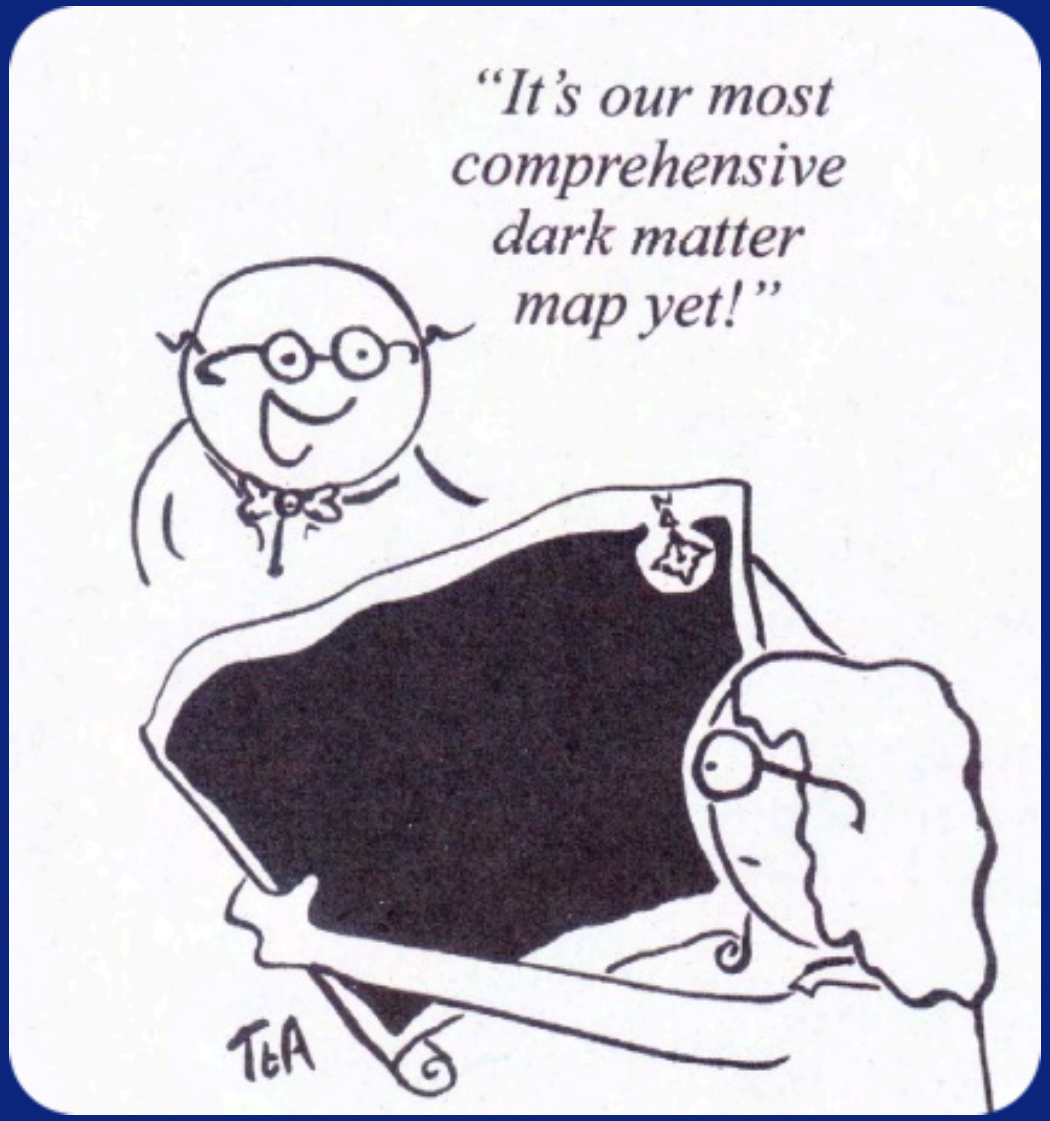




$$F = G\frac{Mm}{r^2}$$

THE DARK ENERGY SURVEY

Thank you !



The UK Skills Gap

A skills gap exists in the UK. Organisations are struggling to fill vacancies. Employees are moving out of research into industry. Can we attract new people? Can we ensure sustainable career development pathways? These questions - and many more - were discussed during a special session at CIUK 2021 chaired by Richard Gunn (UKRI) and including presentations from Christine Kitchen (Cardiff University), Michael Ball (BBSRC), Mark Wilkinson (DiRAC) and Andrew Medhurst (Inspire People). The session proved so popular with the conference attendees that the panel were unable to answer all of the questions during the allotted time and so they reconvened at a later date to discuss answers to the questions raised. Their response can be found here.

UK Research and Innovation

UKRI's Approach to supporting Digital Research Infrastructure (DRI) skills

Dr Michael Ball, Head of Research Infrastructure, BBSRC and UKRI Digital Research Infrastructure Committee

Digital Research Infrastructure

- Long term planning
- Driven by community requirements
- Environmental sustainability
- Partnership with government and industry
- Five cross-cutting themes

UK Research and Innovation

Turning data into knowledge

Catalysing breakthroughs and accelerating innovation and productivity

A foundation to enable UK researchers and innovators to harness the full power of modern digital platforms, tools, techniques and skills:

- A breadth and depth of capabilities and skills
- Seamless connection of communities to data, tools and techniques
- Accelerating productivity by enabling secure and easy access
- A step change in computational power
- Fostering collaboration across disciplines
- New capabilities and new communities of practice
- Environmentally sustainable

UKRI National Digital Research Infrastructure

Data services	Large-scale computing	Securing Trust in Data-intensive Research	Building Skills and Career Pathways	Foundational Tools and Techniques
■ Data infrastructure ■ High throughput computing ■ Sustainable platforms and technologies	■ Super computing ■ Cloud services	■ Trusted Research Environments	■ Skilled DRI professionals ■ Training ■ Career paths	■ Networks ■ AAAI ■ Best practice ■ Knowledge Transfer ■ Policy ■ Software

The inter-dependent themes of the digital ecosystem

Guiding principles

UKRI will:

- Be driven by the ambition of our diverse communities; we will provide a federated system of DRI that provides both a breadth and depth of capabilities and skills
- Build a DRI to catalyse breakthroughs and accelerate innovation
- Accelerate productivity by enabling consistent, appropriate, and secure access through common access procedures and authorisation technologies
- Understand the environmental impacts of our investments and make changes to improve our sustainability that contribute to the UK's commitment to Net Zero
- Work with partners to maximise the value from our investments
- Think and plan on a long-term horizon to evolve the UKRI's DRI in the context of other government investments and strategies in this space

Placing a premium on people

- People run our services
- People write our software
- People work with our university, institutes, cultural sector, and industrial base
- People help make sense of the data and create knowledge

UK Research and Innovation

Theme 4: Skills and Career Pathways

- To fully exploit the opportunities presented by DRI, the breadth and depth of skills support needs to be increased.
- This requires a new approach to engaging, developing and retaining expertise, as well as supporting rewarding, sustainable and flexible career pathways

‘DRI Professionals’

- An umbrella term that encompasses a wide range of roles and careers
- Although these roles are in demand across the whole of research and innovation, within some communities there are particularly urgent needs

- Some DRI professional roles:
  - Systems architects
  - System administrators
  - Programmers
  - Developers
  - Research Software Engineers (RSEs)
  - Information security professionals
  - Research operations engineers (ResOps)

…and many others!

UK Research and Innovation

Objective: Grow the numbers of DRI professionals

- **Training:** Ensure that professionals have access to, and funding for, training and professional development
- **Community Building:** Ensuring effective mechanisms for networking and community building for the benefit of all UK R&D.
- **Career paths and professional development:** Build recognised, attractive career paths, across all career stages.

UK Research and Innovation

Objective: Increase retention of DRI skills within the research and innovation ecosystem

- **Influence:** UKRI occupies a unique position in the landscape - work with Universities, Industry, Policy makers and the Learned Societies to ensure that appropriate training mechanisms, career pathways etc are created.
- **Recognition**: Ensuring these careers are recognised through the creation of local and national recognised hubs or centres of excellence.
- **Incentivise best practice:** Building on our Influence and recognition to reward best practice where it is happening to help further and broaden this support for DRI professionals, and ensure it becomes the norm and not the exception.

UK Research and Innovation

UKRI's Role

- As a funder……support training and recognise the contributions of DRI professionals
- As an employer…..support career pathways within our own centres/institutes/units
- As a partner…..develop and implement best practices
- As a convenor....support communities and help development of networks
- As a champion….continue to advocate for recognition of DRI professional roles, and for reward and recognition for research outputs such as data and software

UK Research and Innovation

A phased approach

- Phase 1 (2021/22) – **We have embarked on the first phase of developing a national DRI**, with £17 million invested in:
  - A portfolio of interventions to enhance our existing digital infrastructures
  - Initial investments in priority areas including Net Zero and Trusted Research Environments
  - Scoping activities to assess the communities' data and computing requirements in more detail
- Phase 2 and 3 (2022+) – Subject to funding, longer term allocations will allow us to expand capacity and capability appropriately and efficiently to meet community requirements, to **catalyse breakthroughs, and accelerate innovation and productivity**

UK Research and Innovation

UK Research and Innovation

Thank you

@UKRI_news

UK Research and Innovation

UK Research and Innovation

https://www.ukri.org/our-work/creating-world-class-research-and-innovation-infrastructure/digital-research-infrastructure/

Building the HPC community of tomorrow

CIUK 2021

Mark Wilkinson
DiRAC HPC Facility

DiRAC

Science & Technology Facilities Council

Why skills? What skills?

- The UKRI e-Infrastructure Roadmap highlights the need for increased numbers of skilled people to support the successful exploitation of HPC and AI in UK science and industry
- *Algorithms, Software & Skills* white paper (2019): “train a workforce in the skills needed to create digital products and services to raise the level of core programming and data handling skills across the research community”
  - Calls for new computational infrastructure must be accompanied by proportionate investment in skilled specialists.
- Critical shortages of RSEs, ResOps, DevOps, SysAdmins, Data Stewards
- Create the right incentives and environment to attract the best talent to the UK
  - Establish career pathways for skilled technical staff

Diverse science workflows require heterogenous architectures

DiRAC

Extreme Scaling
"Tursa"
(Edinburgh)

Atos

- Nvidia A100-based system
- large lattice-QCD simulations

Data Intensive
"DIaL" and "CSD3"
(Leicester & Cambridge)

Hewlett Packard Enterprise

DELL EMC

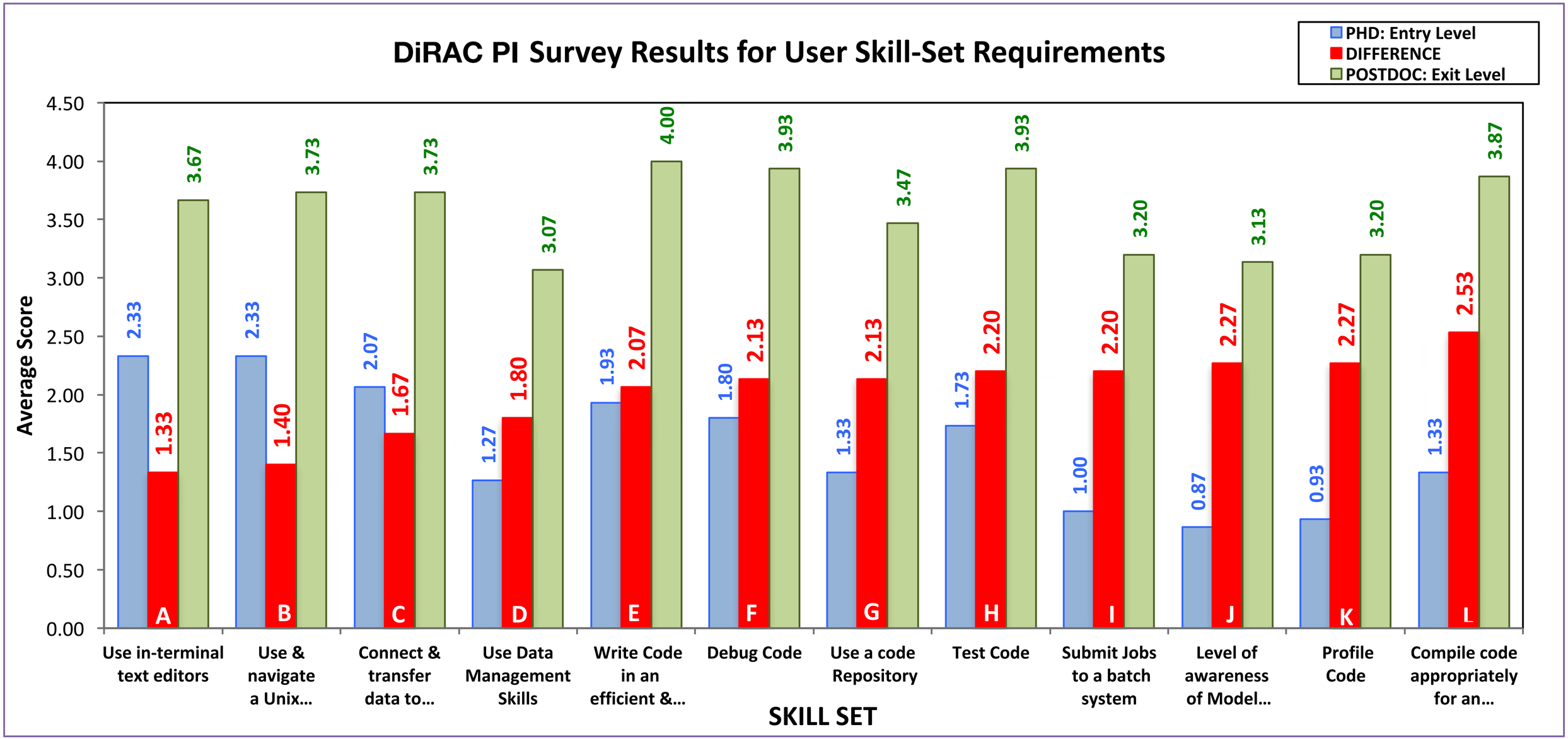
Heterogeneous architecture to support complex simulation and modelling workflows

Memory Intensive
"COSMA8"
(Durham)

DELL EMC

360 TB RAM to support largest cosmological simulations

Identifying training needs and skills gaps



DiRAC Essentials Programme

- DiRAC Essentials Level test ensures all users have base level of HPC skills
- Online (and compulsory!)
- DiRAC provides ***access*** to training from wide pool of providers
- Facility training goals:
  - maximise DiRAC science output through more efficient software
  - flexibility to adopt most cost-effective technologies
  - future-proofing our software and skills
  - contribute to increasing skills of wider UK economy

DiRAC Training Programme

- Training programme includes workshops: Software Design & Optimisation; MPI programming
- CodeCamps: Nvidia, Intel
- New in 2021: AI/ML workshops in collaboration with STFC SciML team
- Currently working with Software Sustainability Institute to develop bespoke training material in 2022

DiRAC Hackathon Programme

- Started in 2018 - success of first three events (Nvidia, Arm, Intel) has led to this becoming a core DiRAC innovation and training activity in support of the DiRAC science progamme.
- Benefits:
  - Provides insights into potential value of new technology for DiRAC services
  - Provides industry partners with feedback on new and future products
  - Enhances skills of users and technical staff through exposure to new technologies
  - Opportunity to improve performance of current codes on existing hardware

Innovation placements

- Key pillar of DiRAC training portfolio
- 6-month placements for PhD students and early-career PDRAs
- 12 placements supported to date

- 2021/22 opportunities:
  - DiRAC-3 partner funded: 3 placements
  - DiRAC Federation Project (UKRI Digital Research Infrastructure funding): ~10 placements
  - Ideas for partners and/or projects welcome
- Benefits include:
  - Bi-directional knowledge exchange
  - Enhanced employability
  - Experience of working in industrial environment
  - Skills development

Kickstart your HPC career

- The UKRI e-Infrastructure Roadmap highlighted the essential need for increased numbers of skilled people to support the successful exploitation of HPC and AI in UK science and industry
- CIUK 2020 featured the inaugural CIUK Student Cluster Challenge as a step towards addressing this need
  - Developing the next generation of HPC experts
  - Encouraging engagement with industry partners in innovation projects
- CIUK and committee recognised the importance of this initiative and were determined to ensure the competition proceeded in the virtual-only format for CIUK 2020.
- In 2022, we will be exploring how to expand the cluster challenge programme

Conclusions

- Training needs to be at the heart of the UKRI Digital Research Infrastructure
- DiRAC Innovation placements - closing date: 13th December, 4pm
- Watch out for details of the “Kickstart your HPC career” programme in 2022

DiRAC - HPC

@DiRAC_HPC

dirac.ac.uk

HPC-SIG Proposal

E^2SAS: Starting point – possible plan of action

November 2021

Globally acknowledged challenge

HPC-SIG proposal EPSRC Tier-2 call (Q3/2019) "E^{2}SAS"

- E^{2}SAS (Environment for the Enhancement of System Administration Skills) strong impact agenda encompasses knowledge, people, society and economy, with **outreach and training** central to the proposal.
- £Billions internationally invested in UK in technology start ups
- Increasing diverse research domains and complexities of workflows conducted on research computing services – main stream solutions.
- Acknowledgement: **skilled workforce to run these services is dwindling**, in stark contrast to the capital investment in technology.
- Globally academia and industry are continually struggling to recruit staff into these roles – often citing a **lack of talented individuals / continually recruiting from the known existing sources** (fishing in the same talent pool exhausting supplies!) – QUOTA EXCEEDED?

HPC-SIG Approach:

- Discussions within the HPC-SIG have identified this failure to **invest in fostering the next generation of system administrator** as a major risk given how mainstream research computing has become.
- E^{2}SAS proposed a foundation **framework** to act as a focal point for outreach to foster the community and align these activities with **existing initiatives to encourage the next generation** of technical support staff.
- Opportunity to review the skills and job descriptions (clear **role definition and consistent terminology**), establishing a core set of best practice training material and mentorship schemes (**industry recognized accredited engineers standards**)
- **promote** the diversity of opportunities associated with running research computing facilities.
- strong presence in related **public engagement and dissemination events.**

Four Cornerstone Foundations

1. **Foundation/focal point to address current under-investment** in system administration / service provision, foster & promote opportunities to new generation
2. Cannot address all issues simultaneously – intention to **establish foundation framework** that can align and complement existing initiative and activities to maximize impact and allow expansion as initiatives evolves.
3. Replicate success through EPSRC's leadership investment in RSEs to start to **address career pathways and removing perceived barriers** to ensure continue to evolve services to meet research challenges
4. **Partnership and co-development environment** for academia, suppliers, and public to collaborate to ensure best practices are adopted to address the lack of recognized skilled workforce impacting the entire community

BETTER SERVICES, BETTER RESEARCH

Service excellence as important as software development – without investment in ensuring optimal environment, software development not reach full potential!

arcca E²SAS "Endorsements"

Right direction, not optimal call – important to ensure issue escalated

Significant support – over 15 Letters of Support from European Centres, Industrial suppliers, and HPC-SIG committee institutions – genuine enthusiasm to jointly address the skills shortage challenge.

Opportunity to build on this initial engagement piece to start to develop a template proposal – requires **dedicated effort to drive forward**, cannot be achieved under a 'best endeavours' activity.

Community buy-in essential - to agree the core skills and processes – technology agnostic approach – basic competencies agnostic to solution – identify methodologies and principles core to building the next generation.

Proposed in 2019 – **progress since**?

arcca E²SAS Summary

Foundation underpinning UK e-Infrastructure given its strategic goal of knowledge transfer and upskilling the next generation of service providers - enhancing the **productivity of present and future HPC services**.

Recognised skills shortage at a national and international level by seeking to attract the next generation of system administrators to the opportunities associated with running research computing facilities**.**

Key opportunities identified by E²SAS include:

- **Raising awareness of opportunities within system administration**.
- **Co-development i**n evaluating software tools to meet the challenges
- **Adoption by HEIs and Suppliers** – apprenticeship schemes to provide evaluation opportunities prior to committing to future career paths.
- Provision of training into **return to work programmes and developing an environment to foster under-represented communities** to develop skills.
- Establish UK as recognized source of **talented and accredited** skilled workforce – **career pathways**.
- **Skills refresh** for both academia and technology suppliers

How to attract private sector IT talent

Deliver more innovation and better services to the UK

Attracting private IT talent into research, academia and public sector

- Today’s IT Job Market
- Your Secret Weapon
- Practical Steps

How to Attract Private Sector IT Talent

- Merger /Start-Up
- TUPE some IT staff
- Reliance on contractor staff
- Attract and hire approx. 40 staff

- **No site**
- **No identity or brand**
- **12 month Fixed Term Contracts**
- **Salaries below market average**

How to Attract Private Sector IT Talent

Results:

- Attracted private sector talent across all IT functions
  - Infrastructure Engineers
  - Software Developers
  - IT Support

- **Some took pay cuts to join**

Vision: A better society, enabled by technology

- CIOs
- Heads of Scientific Computing
- Technical Architects
- DevOps Engineers
- Software Developers
- IT Support

Working with organisations who make a difference

Vision: A better society, enabled by technology

Companies House

BANK OF ENGLAND

Intellectual Property Office

Department for International Trade

Working with organisations who make a difference

Today's IT Job Market

Demand:

- Advertised job numbers close to pre-pandemic levels
- Employer competition increasing salaries and counter offers
- More advertised jobs than active job seekers

Impact of COVID-19:

- Increased employee loyalty; less churn
- Impact of remote / hybrid working (job choices; disposable income)
- Reduced application numbers

Today's IT Job Market: Public Sector Perception

Perception:

- Outdated technology
- Bureaucratic
- Inflexible
- Non-inclusive
- Limited personal development
- No value to CV
- Stuffy, boring

More Like:

- Technology transformations
- Innovative
- Agile
- Diversity & Inclusive focused
- Plenty scope for personal development
- Make a difference
- Challenging, meaningful

Your Secret Weapon: Social impact

Social Impact:

- What's your **social impact?**
- How do promote **the difference** you make?
- How important is it that new hires **share** that passion?

Promotion:

- Inform and sell your **social value/impact**
- **Change** perceptions
- Go **social -** employee stories, articles, blogs, videos

Differentiate:

- **Culture** and **meaning** matter
- Emphasis **diversity & inclusion, learning** and **growth**
- Build your **employer brand** and **reputation**

Practical Steps:

Mindset

- Use your **secret weapon!**
- Be **proactive**
- **Sell** job opportunities (technology, growth, meaning)

Flexibility

- Make the candidate journey **simple and engaging**
- Be **flexible** where you can
- Hire on **values and cultural fit** as well as skills

Actions

- Review your **recruitment process**
- Define your **Employer Value Proposition**
- Develop **wider talent pools** and **relationships**

Advice:

Andrew Medhurst:
andrew@inspirepeople.net

Tel: 020 7871 8558
www.inspirepeople.net

Prof Viv Kendon (University of Strathclyde)

QEVEC: integrating quantum computing with HPC

Abstract: ExCALIBUR cross-cutting project QEVEC - quantum enhanced and verified exascale computing - contributes to the development of exascale computing by focusing on how to add quantum computers as co-processors to HPC. Early quantum computers will be much smaller -- in terms of the amount of classical data they can process in one go -- than current HPC.

But the processing power on that data can be much faster due to their quantum properties of superposition and coherence. The most promising way to use them is thus to accelerate those parts of the computations that are slow for HPC. This requires detailed study of the algorithms, both quantum and classical, which QEVEC will do for two specific applications areas, fluids simulations and materials simulations. I will explain why quantum computing is so promising for enhancing computational capability, while avoiding the current hype, and focusing on the hard work required to realise this potential for useful applications.

Bio: Prof Kendon joined Strathclyde in Nov 2021 as Professor of Quantum Technology, has been working on quantum computing for the past 20 years, including three fellowships, most recently EPSRC Established Career Fellow in Hybrid Quantum Computing (2014-19 at Durham University). She is PI of QEVEC, ExCALIBUR cross-cutting project Quantum Enhanced and Verified Exascale Computing incorporating quantum computing into future exascale high performance computing, and chairs CCP-QC, the Collaborative Computational Project in Quantum Computing, an EPSRC-STFC funded network linking quantum computing and scientific computing experts in applications suitable for quantum enhancement.

QEVEC: integrating quantum computing with HPC

Quantum
Enhanced
Verified
Exascale
Computing

Viv Kendon

CNQO
Physics
Strathclyde
viv.kendon@strath.ac.uk

*Manchester – **CIUK 2021** – 9+10 December 2021*

Funding: UKRI – EPSRC – ExCALIBUR

Partners: UK National Quantum Technologies Programme; QCS Hub; Quantum Computing & Simulation Hub; NQCC National Quantum Computing Centre; CCP-QC

quantum computing *HYPE*

'Digital marketing poised for a quantum leap'

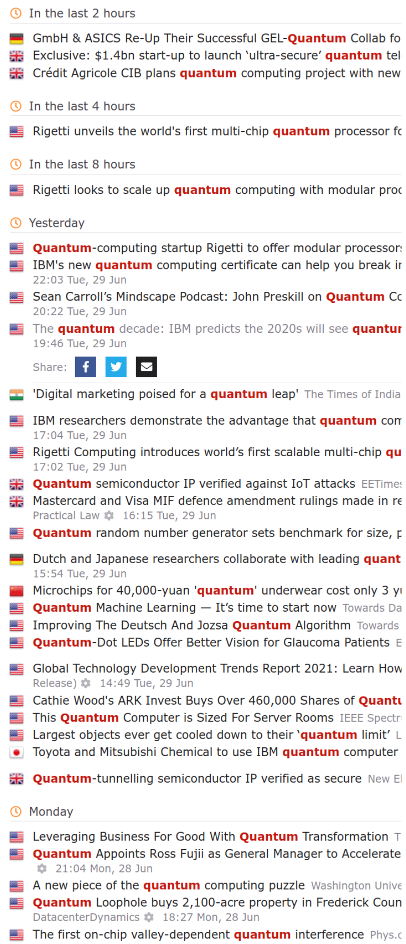
Quantum Machine Learning — It's time to start now

Microchips for 40,000-yuan 'quantum' underwear cost only 3 yuan . . .

- wild claims and promises
- talk of revolutions and paradigm shifts
- huge venture capital investment

$\Longrightarrow$ *remember:* venture capitalists expect >9/10 to fail

★ this talk: quantum computing without the hype . . .


In the last 2 hours
GmbH & ASICS Re-Up Their Successful GEL-Quantum Collab fo
Exclusive: $1.4bn start-up to launch 'ultra-secure' quantum tel
Crédit Agricole CIB plans quantum computing project with new
In the last 4 hours
Rigetti unveils the world's first multi-chip quantum processor fo
In the last 8 hours
Rigetti looks to scale up quantum computing with modular proc
Yesterday
Quantum-computing startup Rigetti to offer modular processor
IBM's new quantum computing certificate can help you break ir
22:03 Tue, 29 Jun
Sean Carroll's Mindscape Podcast: John Preskill on Quantum Co
20:22 Tue, 29 Jun
The quantum decade: IBM predicts the 2020s will see quantur
19:46 Tue, 29 Jun
Share:
'Digital marketing poised for a quantum leap' The Times of India
IBM researchers demonstrate the advantage that quantum com
17:04 Tue, 29 Jun
Rigetti Computing introduces world's first scalable multi-chip qu
17:02 Tue, 29 Jun
Quantum semiconductor IP verified against IoT attacks EETimes
Mastercard and Visa MIF defence amendment rulings made in re
Practical Law 16:15 Tue, 29 Jun
Quantum random number generator sets benchmark for size, p
Dutch and Japanese researchers collaborate with leading quant
15:54 Tue, 29 Jun
Microchips for 40,000-yuan 'quantum' underwear cost only 3 yu
Quantum Machine Learning — It's time to start now Towards Da
Improving The Deutsch And Jozsa Quantum Algorithm Towards
Quantum-Dot LEDs Offer Better Vision for Glaucoma Patients E
Global Technology Development Trends Report 2021: Learn How
Release) 14:49 Tue, 29 Jun
Cathie Wood's ARK Invest Buys Over 460,000 Shares of Quantu
This Quantum Computer is Sized For Server Rooms IEEE Spectr
Largest objects ever get cooled down to their 'quantum limit' L
Toyota and Mitsubishi Chemical to use IBM quantum computer
Quantum-tunnelling semiconductor IP verified as secure New El
Monday
Leveraging Business For Good With Quantum Transformation T
Quantum Appoints Ross Fujii as General Manager to Accelerate
21:04 Mon, 28 Jun
A new piece of the quantum computing puzzle Washington Unive
Quantum Loophole buys 2,100-acre property in Frederick Coun
DatacenterDynamics 18:27 Mon, 28 Jun
The first on-chip valley-dependent quantum interference Phys.c

headlines from newsnow.co.uk $\longrightarrow$

current quantum computers?

promises are for venture capitalists . . . what can they actually do?

- IBM and Google have quantum chips with about 100 qubits
  – *noisy/imperfect even when cooled to very low temperatures*
- running a particular random sampling task (not useful)
- needs **near exascale** classical HPC to verify output

LOTS of caveats and arguments about this, of course . . .

$\Longrightarrow$ take home message from the hype:

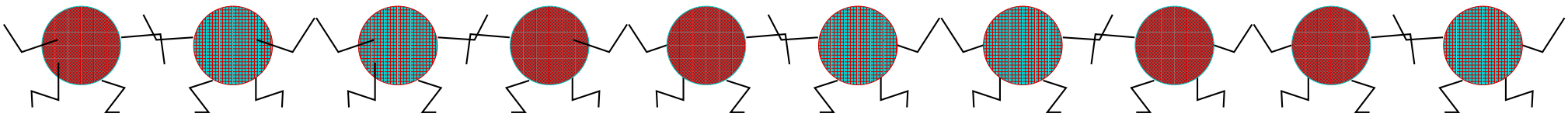
★ potential is real, even for early hardware ★

★**IF**★ we can use it for something we actually want to compute . . .

$\Longrightarrow$ time to think seriously about how to use a quantum computer $\Longleftarrow$

overview of talk

- hype (done)
- co-processors and context
- what is quantum computing?
- how does it work?
- how to make it useful?
- QEVEC and CCP-QC
- opportunities and jobs ...



hybrid computers . . .

practice: co-processors:

conventional:

- GPUs graphics cards
- ASIC application-specific integrated circuit
- FPGA field-programmable gate array

unconventional: control + substrate:

- quantum
- NMR
- reservoir
- slime mould

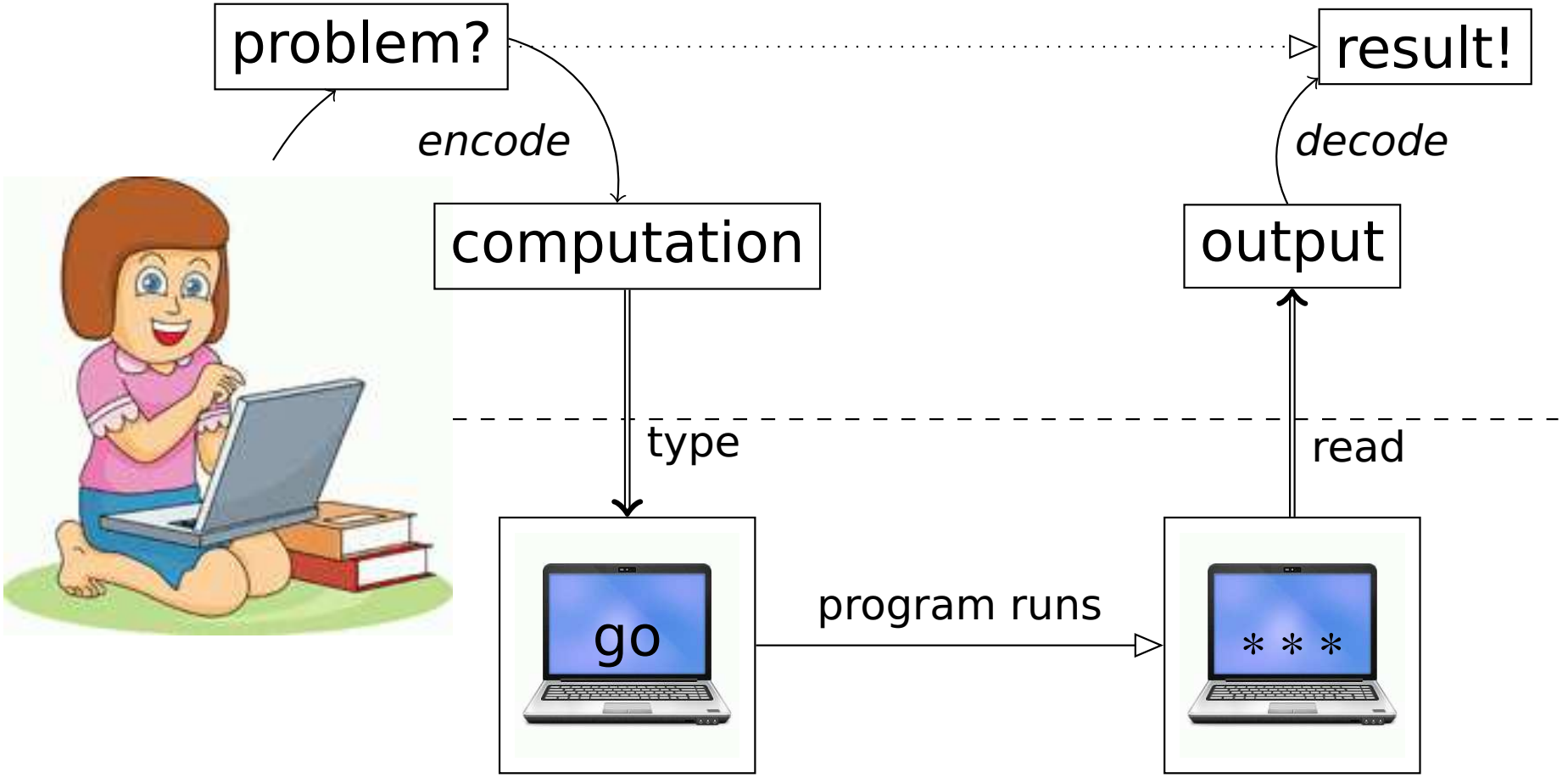
★ hybrid computational systems are the norm ★

theory: single paradigm:

- classical – Turing Machine
- analog – Shannon's GPAC
- quantum – gate model, QTM, CV, MBQC, QW, AQC, . . .
- linear optics (Bosons) [Aaronson/Arkhipov STOC 2011 ECCC TRI-10 170]

many different quantum models: quantum circuit/gate model roughly corresponds to digital classical

computing



article: “When does a physical system compute? Proc. Roy. Soc. A 2014 **470**, 20140182

`http://dx.doi.org/10.1098/rspa.2014.0182` (Horsman/Stepney/Wagner/VK)

quantum computing

input $\longrightarrow$ encode $\longrightarrow |\psi_{in}\rangle \longrightarrow \hat{U} \longrightarrow |\psi_{out}\rangle \longrightarrow$ decode $\longrightarrow$ result

$\hat{U}$ is unitary evolution (or more generally, open system/environment)
– can be gate sequence or engineer Hamiltonian $\hat{H}(t)$ such that

$$|\psi_{out}\rangle = \mathcal{T} \exp\{-i/\hbar \int dt\ \hat{H}(t)\}\, |\psi_{in}\rangle$$

★ covers most of quantum information processing . . .
. . . including communications, where aim is *result=input*

encode – arbitrary choices:

using spin-down $|\downarrow\rangle \equiv 0$ instead of spin-up $|\uparrow\rangle \equiv 0$ makes no difference
$\rightarrow$ provided encode and decode done consistently

quantum information processing:

quantum computing (& quantum comms) built on the idea that:

quantum logic allows greater **EFFICIENCY** than **classical logic**

classical	**quantum**
bits, 0 or 1	qubits, $\alpha\|0\rangle + \beta\|1\rangle$ *coherence*
yes or no, binary decisions	yes and no, *superposition*
HEADS or TAILS, random numbers	random measurement outcomes

⇒ quantum gives different computation from classical: *how different?*

- **computability** – what can be computed?
- **complexity** – how efficiently can it be computed?

★ quantum **computability** is the same as classical
complexity differs: some problems can be computed more **EFFICIENTLY**

quantum advantage?

how to translate quantum logic into better computing devices?

depends on definition of **EFFICIENCY**

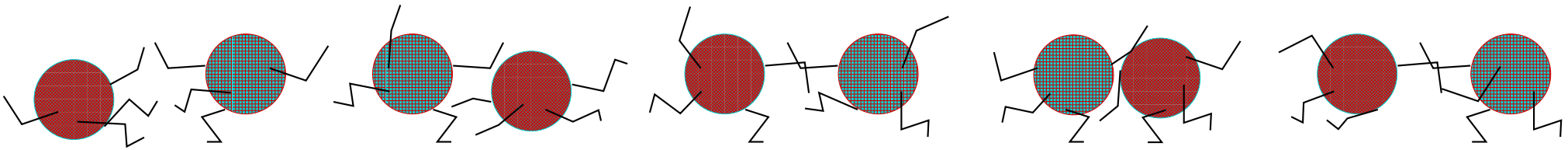
- *in theory: polynomial scaling with system size*
- *in practice: produces answers on human timescales*

roughly speaking:
quadratic speed up exploits quantum coherence, interference effects

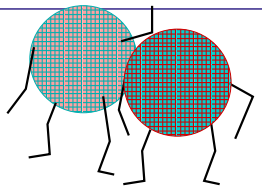
exponential speed up exploits parallelism in quantum superposition

★ comparison of real physical devices, not of mathematical theories

⇒ complexity theory alone won't tell you whether useful in practice



quantum gates



qubits: 2-state quantum systems: *examples:* electron spin, photon polarisation
localised: distinguishable – no Fermi or Bose statistics to deal with

- choose a basis: $|0\rangle$ and $|1\rangle$ superpositions $\alpha|0\rangle + \beta|1\rangle$

input $\longrightarrow$ encode $\longrightarrow |\psi_{in}\rangle \longrightarrow \hat{U} \longrightarrow |\psi_{out}\rangle \longrightarrow$ decode $\longrightarrow$ result

$\Rightarrow$ how to apply $\hat{U}$ to the quantum state? $\hat{U}$ is "program" – could be any unitary
– efficient program needs efficient description

n qubits: $\hat{U}$ is $2^n \times 2^n$ so efficient program = sparse matrix, $POLY(n)$ non-zero

decompose $\hat{U}$ into smaller operations applied to a few qubits at a time
★ important result: like classical, this is possible for quantum too ★

one entangling two-qubit gate + a few single qubit gates $\Rightarrow$ construct any $\hat{U}$

$\Rightarrow$ which universal gate set to choose?
– depends on physical system and error correction requirements...

quantum superposition

consider three bits:

0	1	2	3	4	5	6	7
000	001	010	011	100	101	110	111

suppose we want to compute $F(x)$ for 8 different values of x ?
run the program eight times, once for each value of x

$\Longrightarrow$ what if we could input all eight values at once?
quantum mechanics lets us do this: **superposition:**

$$|000\rangle + |001\rangle + |010\rangle + |011\rangle + |100\rangle + |101\rangle + |110\rangle + |111\rangle$$

no free lunch: get a superposition of all eight answers . . .

select one, find out which x it corresponds to: quantum advantage!

how? ↑ smart quantum algorithms ↑

useful quantum computing needs **quantum algorithms**

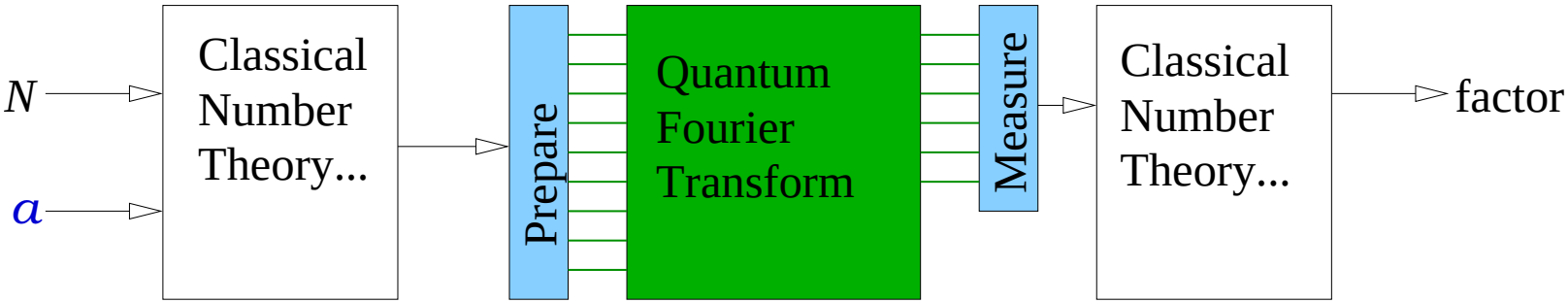
factoring large numbers – is hard!

basis of crypto schemes, use numbers too large to factor classically

large enough **quantum computer** will break this: **Shor's algorithm**

problem: find a factor of a number $N = pq$

algorithm: first choose a co-prime number a



classical *quantum . . quantum* *classical*

★ the hard bit is to find r, the periodicity of $x^a \pmod N$ for $0 < x < N$

→ Fourier transforms good at this...QFT exponentially faster than FT

$a^{r/2} \pm 1$ gives a factor of N (with high probability)

how good is Shor's algorithm?

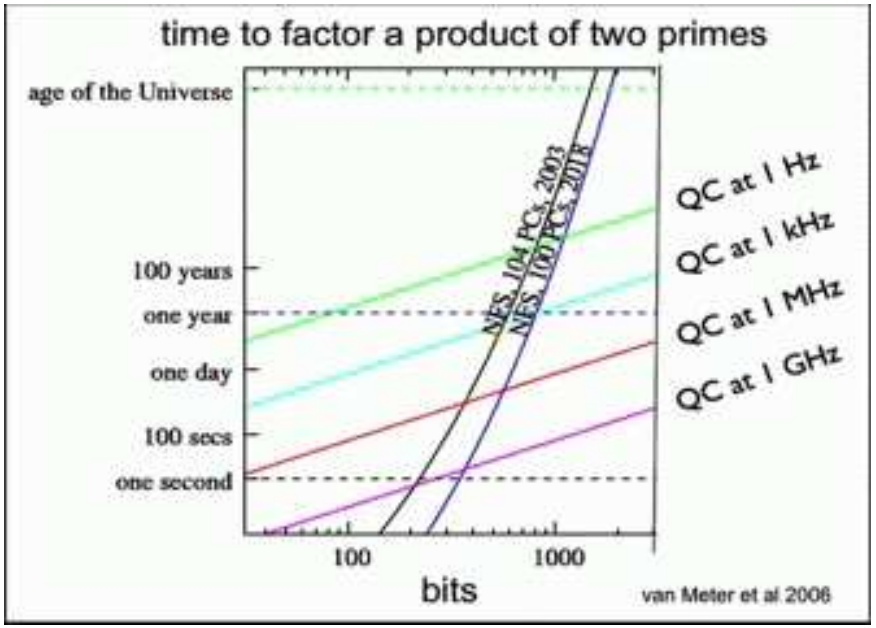
need to beat: best classical (2009): 232 digits (RSA-768) = 768 bits

Shor's algorithm for n-bits needs: $2n$ qubits in QFT register plus $5n$ qubits for modular exponentiation = $7n$ logical qubits → 768 bits need 5376 logical qubits

now add error correction: depends on error rates...
–if error rate close to threshold
– need more error correction

for low error rates, maybe 20–200 physical qubits per logical qubit; for high error rates, maybe 10^5

suggests we may need Teraqubit quantum computers to break factoring – scaling favours quantum, but the crossover point is high
current h/w: factor 35 i.e., $n = 6$ bits



Quantum annealing (D-Wave) may challenge sooner: different algorithm – scaling $\sim n^2$ qubits; current h/w: 18 bit numbers [SciRep (2018) 8:17667]
→ *~1500 times larger RSA-768* – embedding o/h uncertainty for 147456 qubits

encoding problems in qubit Hamiltonians

+ computational basis state $|j\rangle = |q_0 q_1 \dots q_k \dots q_{n-1}\rangle$ with $q_k \in \{0, 1\}$
+ superposition of all basis states:

$$|\psi_0\rangle = 2^{-n/2} \sum_{j=0}^{2^n-1} |j\rangle = 2^{-n/2} \left(|0\rangle + |1\rangle\right)^{\otimes n}$$

encode problem into n-qubit Hamiltonian $\hat{H}_p$

such that **solution** is lowest energy state (ground state)

example: find state $|m\rangle$ then $\hat{H}_p = \mathbb{1} - |m\rangle\langle m|$

example: three qubits, exactly one must be $|1\rangle$

$$\hat{H}_p = (\mathbb{1} - \hat{Z}_1 - \hat{Z}_2 - \hat{Z}_3)^2$$

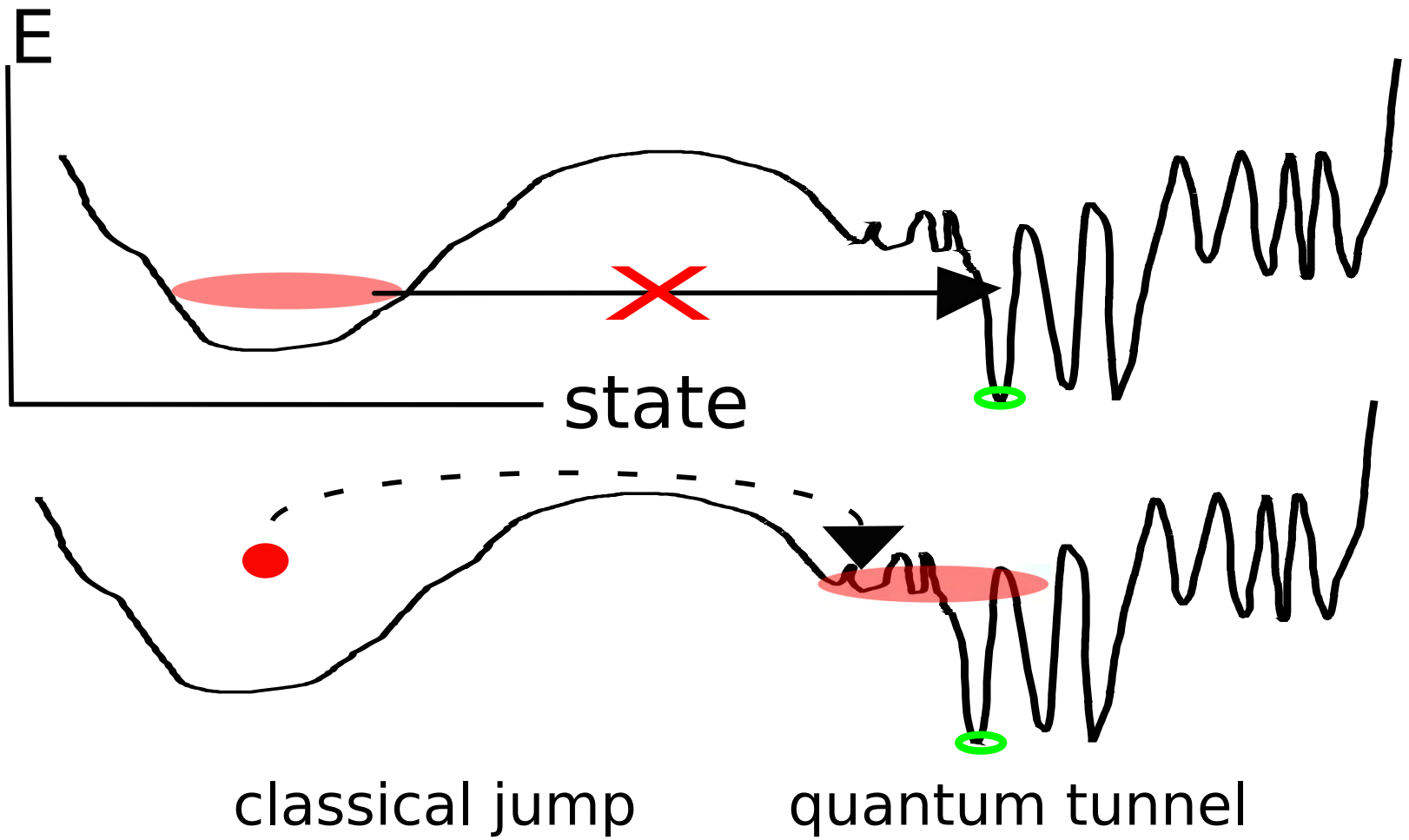
Pauli-Z operator: $\hat{Z}\,|0\rangle = |0\rangle$ and $\hat{Z}\,|1\rangle = -\,|1\rangle$

$$\hat{Z} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

$$\hat{X} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

quantum annealing energy landscape

find the lowest point:



graphic credit: Dr Nick Chancellor, Durham University

beyond classical?

simulating a quantum system: example – $N \times$ 2-state particles
$\longrightarrow 2^N$ possible states – could be in superposition of all of them

classical requires:
one complex number per state: $2^{N+1}\times$size-of-double $\longrightarrow$ 1Terabyte holds $N = 36$
– each additional particle doubles memory required
compression methods can squeeze a little more [Chen et al arXiv:1802.06952, 64 qubits]
– *methods being developed for verification of early quantum computers*

100 qubits = well beyond exascale

if don't need to track all superpositions, larger classical simulations possible
(e.g., matrix product states, tensor networks . . .)

classical is limited to subspace size < 50 *qubits however large system is*

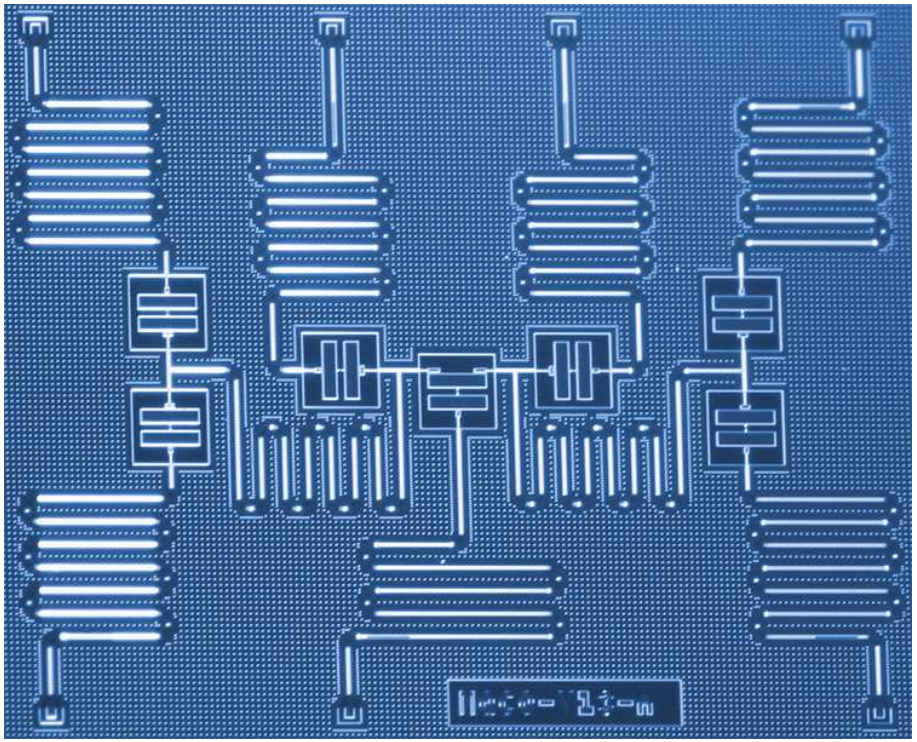
who is building quantum computers?

→ universities and research centres are developing hardware and software . . .

also *many companies* – notably:

have hardware they are talking about now:

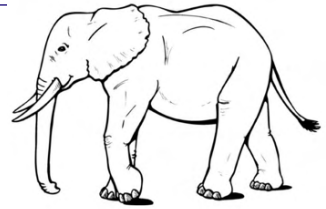
- IBM (Yorktown Heights, US)
- Google (Santa Barbara, US)
- D-Wave Inc, Vancouver, Canada)
- Rigetti (San Francisco, US)
  + *Rigetti UK with ISCF funding...*
- IonQ (Maryland, US) ions
- PsiQuantum (Palo Alto, US) photons

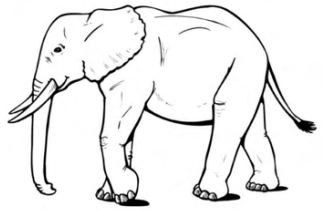


IBM 7 qubit chip

how big? IBM: 127 qubits like the above; D-Wave: 5,000 qubits *(noisy and not as powerful)*; ionQ: 32 low noise qubits, in 19" rack

what I didn't cover:





there are some elephants in the room ...

- **quantum error correction** - not user level; resource intensive
- **quantum verification** - we know how (w/o exascale HPC)
- low level issues for integrating quantum computing with HPC
  = **timescales**, syncronisation
  = **data formats**, data type conversion

→ **we need**

- quantum computing experts
- HPC experts
- applications experts

all working together

⇒ ***to develop hybrid classical quantum algorithms for applications***

CCP-QC

Collaborative Computational Project – Quantum Computing

– facilitating quantum computing applications by networking between computational scientists and the quantum computing community –

CCP-QC activities:

- joint meetings with other CCPs and the quantum computing community
- training days for computational scientists about quantum computing
- run small projects to develop proof-of-principle code and demonstrations on early quantum computing hardware – with STFC CoSeC support
- online information resource on scientific applications of quantum computing

https://ccp-qc.ac.uk @QC_CCP

contact: info@ccp-qc.ac.uk

UKRI Engineering and Physical Sciences Research Council

UKRI Science and Technology Facilities Council

Quantum Enhanced Verified Exascale Computing

★ **Strathclyde:**
Viv Kendon (PI)
★ **Durham:** Richard Bower,
Alastair Basden,
Stewart Clark,
Nicholas Chancellor,
Halim Kusumaatmaja
★ **London Southbank:**
John Buckeridge (KE)
★ **UCL:** Scott Woodley,
Richard Catlow,
Paul Warburton
★ **Warwick:** Animesh Datta

★ ExCALIBUR Cross-Cutting project:
potential disruptor: quantum computing

current – NISQ* era – quantum computers
need near exascale classical to verify

⇒ challenge is to make this potential useful ⇐

[*NISQ = noisy intermediate-scale quantum]

★ *two use cases:* fluids sim and materials sim

systematic evaluation, identification, and development of relevant quantum algorithms for exascale subroutines

★ quantum VVUQ

★ methodology to apply to other use cases

★ *funders and partners:*

UK National Quantum Technologies Programme
Engineering and Physical Sciences Research Council
Quantum Computing & Simulation Hub
National Quantum Computing Centre
CCP-QC

Quantum Enhanced Verified Exascale Computing

★ *two use cases:* fluids sim and materials sim

systematic evaluation, identification, and development of relevant quantum algorithms for exascale subroutines

★ quantum VVUQ

★ methodology to apply to other use cases

★ opportunities and jobs

- *postdocs in:*
  - quantum verification (Warwick)
  - quantum computing (Strathclyde – live!)
  - fluids simulations (Durham – soon …)
  - materials simulations (UCL – soon …)
- *PhDs in:*
  - Reliable quantum simulations of plasma and fusion physics (Warwick)
  - Hybrid quantum algorithms (Strathclyde)
  - UCL CDT …

★ *funders and partners:*

UK National Quantum Technologies Programme · UKRI Engineering and Physical Sciences Research Council · Quantum Computing & Simulation Hub · UKRI National Quantum Computing Centre · CCP-QC

Nick Brown (EPCC, University of Edinburgh)

FPGAs for scientific workloads: The why and the how

Abstract: Field Programmable Gate Arrays (FPGAs) enable the electronics of the chip to be configured to represent a specific kernel or application.

Such tailoring of the electronics to the code means that we bypass the general purpose micro-architecture of CPUs and GPUs, thus being able to organise aspects such as the logic and cache memory to entirely suit what is being executed. Whilst FPGAs have been popular in embedded computing for years, they are yet to gain wide acceptance for HPC workloads. There are a variety of reasons for this, but in the past couple of years there has been massive investments made by vendors in FPGA hardware and software ecosystems, making them a much more attractive choice than ever before.

In this talk I will use real-world applications and kernels to describe the role we see for FPGAs complimenting other hardware technologies at the exascale. I will describe that, whilst high performance is possible, the devil is in the detail and the programmer must recast their algorithm into a dataflow algorithmic style. Lastly I will provide access details for the ExCALIBUR H&ES FPGA testbed, where audience members can sign-up to for free and experiment with FPGAs for their workloads.

Bio: Dr Nick Brown is a Research Fellow at EPCC the University of Edinburgh with interests in HPC application development, novel heterogeneous architectures, data science, programming language design, and compilers.

He is involved with running the UK's FPGA testbed system, which aims to encourage HPC developers to experiment with exploring FPGAs for their scientific and engineering workloads. Nick is a course organizer on EPCC's MSc in HPC and data science courses, as well as supervising MSc and PhD students.

Welcome to CIUK 2021

UKRI Science and Technology Facilities Council

Scientific Computing

CIUK

CIUK 2021

FPGAs for scientific workloads: The why and the how

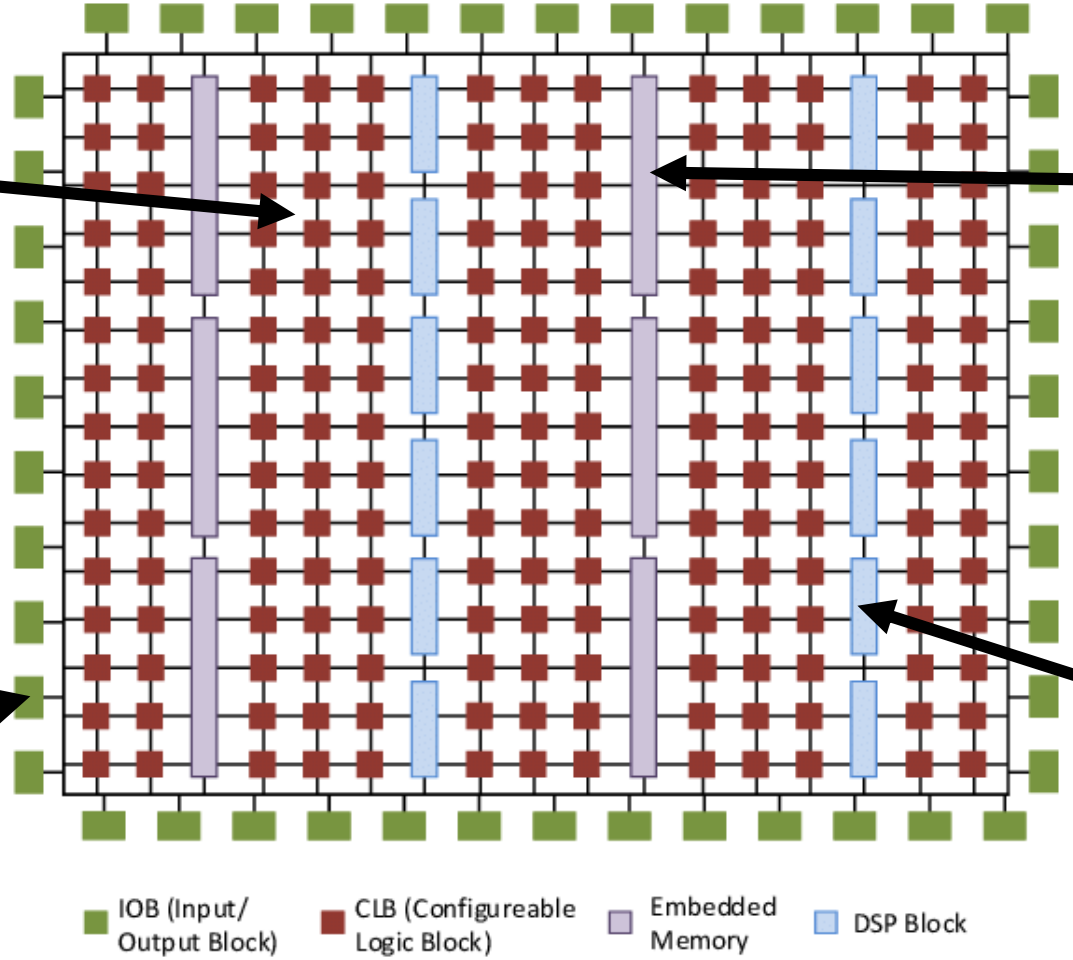
Dr Nick Brown, EPCC

n.brown@epcc.ed.ac.uk

What are Field Programmable Gate Arrays (FPGAs)?

Configurable Logic Block (CLB) contain look up tables which are configured with the application logic. These are sitting within a sea of configurable interconnect

Lots of I/O connections to the outside world, such as PCIe, HBM2, DDR-DRAM, QSFP28 networking



Very fast on-chip memory, known as Block RAM (BRAM) and approx. 40 TB/s, similar to L1 cache and accessible in approx. 1 cycle. Typically a few MB on top end FPGAs

ASIC style components to perform arithmetic, used as the building blocks for floating point arithmetic as this saves a large amount of configurable logic

Worth a fresh look for HPC

- Over 10 years ago we had an FPGA cluster in EPCC
  - But immaturity of the hardware (struggling to match CPU performance) and software ecosystem (difficult to program and lack of tooling) ultimately meant that this was not continued into large scale HPC adoption

- But a decade is a long time, and times change!
  - Much more capable hardware and exciting new technologies on the horizon
  - (Very) significantly enhanced software ecosystem allowing the programming of these via C or C++

But isn't programming these things still hard?

- High Level Synthesis (HLS) allows us to write code in C or C++ for FPGAs – no need to write code in VHDL or Verilog anymore!
  - Xilinx HLS and OpenCL for Xilinx FPGAs, OpenCL for Intel FPGAs
  - OpenCL on the host to drive kernels and transfer data
  - Recent developments make this more a question of software development rather than hardware design, but there are still some challenges!

```
extern "C" {
void sum_kernel(float * input, float * result, float add_val, int num_its) {
    #pragma HLS INTERFACE m_axi port=input offset=slave
    #pragma HLS INTERFACE m_axi port=result offset=slave
    #pragma HLS INTERFACE s_axilite port=add_val bundle=control
    #pragma HLS INTERFACE s_axilite port=num_its bundle=control
    #pragma HLS INTERFACE s_axilite port=return bundle=control

    float sum=0;
    for (unsigned int i=0;i<num_its;i++) {
        float d=input[i] + add_val;
        sum+=d;
    }
    *result=sum;
}
}
```

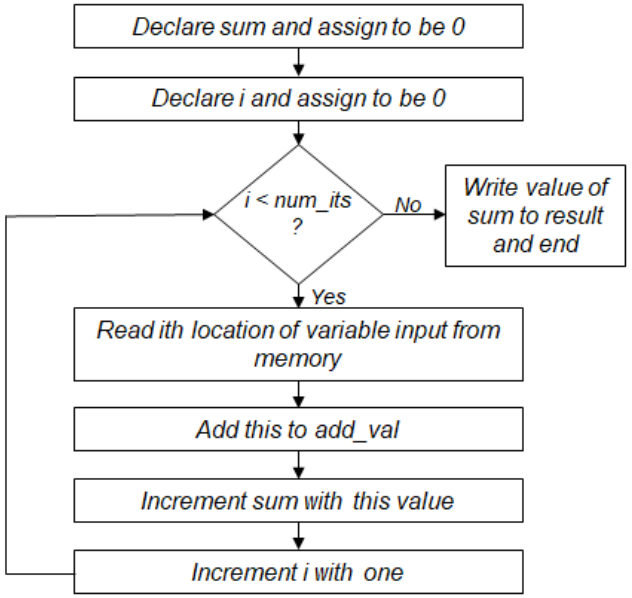
```
> v++ -t hw --config design.cfg -O3 -c -k sum_kernel -o'sum.hw.xo' device.cpp
```

- Some profiling and debugging tools also provided (e.g. Intel integrated with Vtune)
- But building takes a long time (many hours!)
  - Therefore these toolchains provide emulation capabilities where one can build their code in minutes and emulate on the CPU how it would run

What's reconfigurable/spatial/dataflow computing?

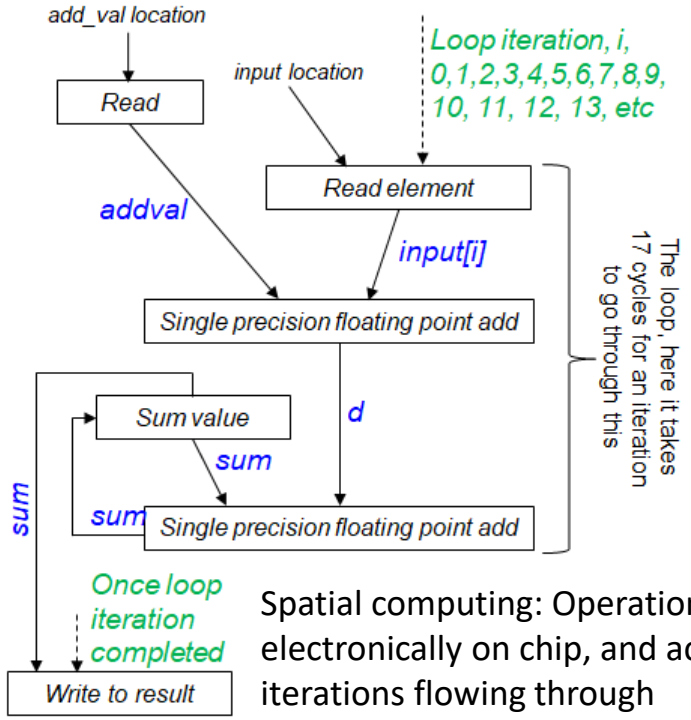
Temporal computing (CPU or GPU)

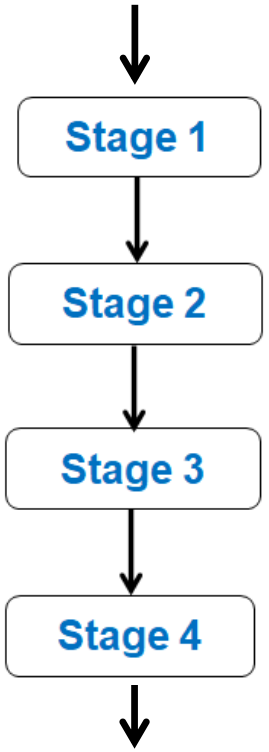
```
float sum=0;
for (unsigned int i=0;i<num_its;i++) {
  float d=input[i] + add_val;
  sum+=d;
}
*result=sum;
```



Temporal computing: Can be thought of like a flowchart, with the PE (e.g. CPU or GPU) executing one stage after another

Reconfigurable architecture (dataflow)

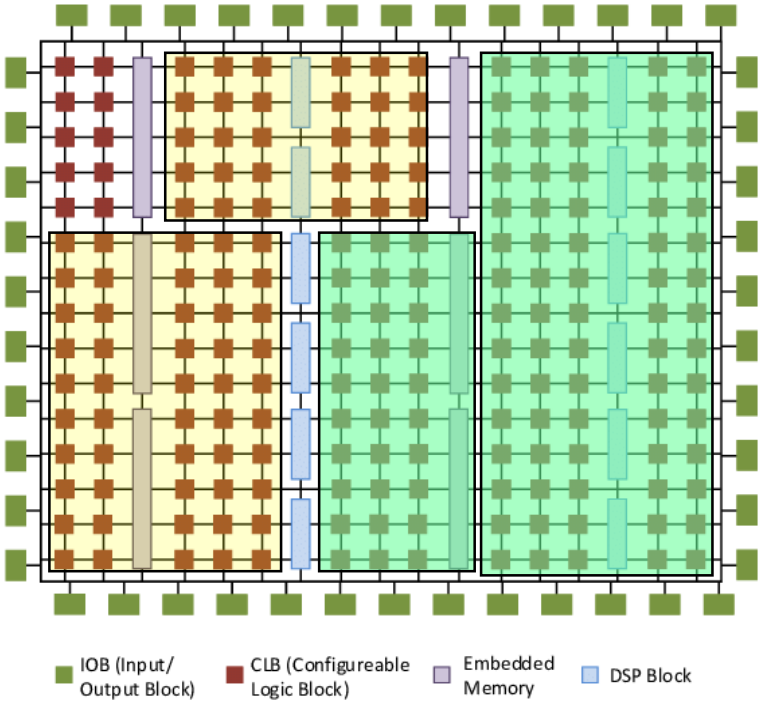


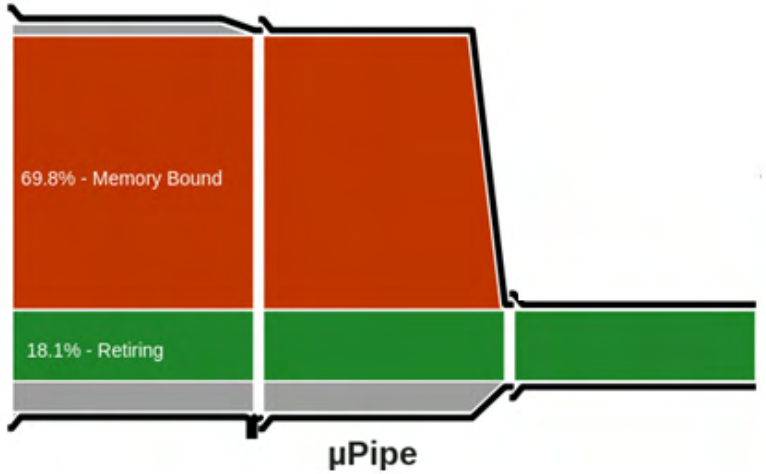


Spatial computing: Operations implemented electronically on chip, and acts as a pipeline, loop iterations flowing through

But what HPC workloads are best suited?

- It's important to pick your battles, and with CPUs and GPUs we have a good idea what performance properties are best suited to the technologies

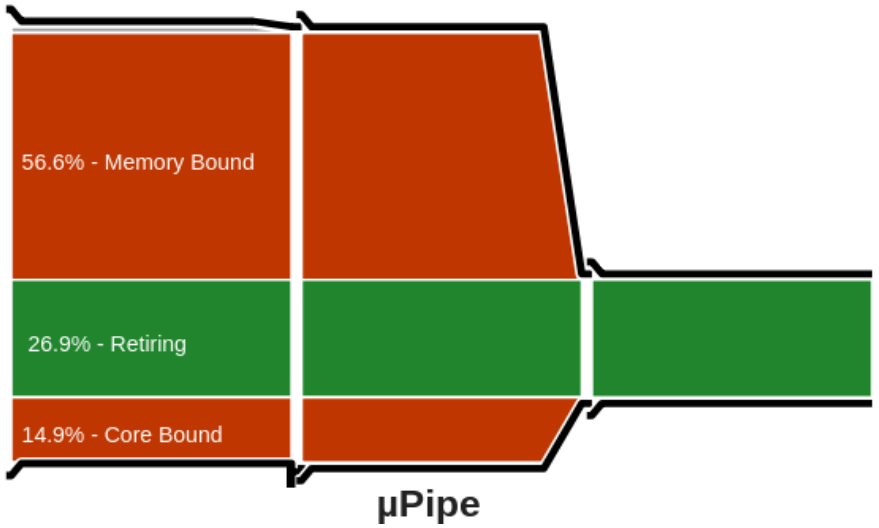




- For FPGAs, if your workload is compute-bound then a GPU is probably a better option!
- However if your code is memory bound or bound by other core issues, this is when you could get benefits
  - Tailoring how you use that fast L1-style BRAM memory
  - Exploiting high-bandwidth off-chip connections and HBM2

Example: AX kernel of Nekbone proxy-app

- Nek5000 is used for high fidelity simulation of rotating parts.
  - Nekbone is a proxy-app that captures the basic structure of Nek5000



On 24 CPU cores: 65.74 GFLOPs

- ➢ Only 11.7 times faster than one CPU core

On a Xeon Platinum Cascade Lake with N=16, 800 elements

```
subroutine ax(n, nelt, w, u, g, dxm1, dxtm1)
  integer, intent(in) :: n, nelt
  real(n,n,n,nelt), intent(in) :: u, g, dxm1, dxtm1
  real(n,n,n,nelt), intent(out) :: w

  do e=1, nelt
    ax_e(n, nelt, w(:,:,:,e), u(:,:,:,e), ...)
  enddo
end subroutine ax

subroutine ax_e(n, w, u, g, dxm1, dxtm1)
  integer, intent(in) :: n
  real(n,n,n), intent(in) :: u, g, dxm1, dxtm1
  real(n,n,n), intent(out) :: w

  real(n*n*n) :: ur, us, ut
  real :: wr, ws, wt

  call local_grad3(ur, us, ut, u, n, dxm1, dxtm1)

  do i=1,n*n*n
    wr = g(1,i)*ur(i) + g(2,i)*us(i) + g(3,i)*ut(i)
    ws = g(2,i)*ur(i) + g(4,i)*us(i) + g(5,i)*ut(i)
    wt = g(3,i)*ur(i) + g(5,i)*us(i) + g(6,i)*ut(i)
    ur(i) = wr
    us(i) = ws
    ut(i) = wt
  enddo

  call local_grad3_t(w, ur, us, ut, n, dxm1, dxtm1)
end subroutine ax e
```

Iterate over elements

Multiply and add values calculated in local_grad3

```
subroutine local_grad3(ur, us, ut, u, n, dxm1, dxm2)
  integer, intent(in) :: n
  real(n,n,n), intent(in) :: u, dxm1, dxm2
  real(n,n,n), intent(out) ::ur, us, ut

  call mxm(dxm1, n, u, n, ur, n*n)
  do k=0,n
    call mxm(u(:,:,k), n, dxtm1, n, us(:,:,k), n)
  enddo
  call mxm(u, n*n, dxtm1, n, ut, n)
end subroutine local grad3
```

Matrix multiplications

- All double precision floating point
- AX kernel applies the Poisson operator of the CG solver accounts for approx. 75% of overall runtime of Nekbone
- 800 elements, and a size of N, with the number of grid points equal to N^3
  - For instance with N=16 then there are 831488 double precision floating point operations per element

Use FPGA to ameliorate overhead of memory access and keep compute fed with data

Overview of single kernel performance

```
void ax_kernel(double * w, double * u, double * gxyz, double * dxm1, double * dxtm1, double * ur, double * us,
    double * ut, double * wk, int nx1, int ny1, int nz1, int nelt, int ldim) {
  #pragma HLS INTERFACE m_axi port=w offset=slave
  #pragma HLS INTERFACE m_axi port=u offset=slave
  #pragma HLS INTERFACE m_axi port=gxyz offset=slave
  #pragma HLS INTERFACE m_axi port=dxm1 offset=slave
  #pragma HLS INTERFACE m_axi port=dxtm1 offset=slave
  #pragma HLS INTERFACE m_axi port=ur offset=slave
  #pragma HLS INTERFACE m_axi port=us offset=slave
  #pragma HLS INTERFACE m_axi port=ut offset=slave
  #pragma HLS INTERFACE m_axi port=wk offset=slave
  #pragma HLS INTERFACE s_axilite port=nx1 bundle=control
  #pragma HLS INTERFACE s_axilite port=ny1 bundle=control
  #pragma HLS INTERFACE s_axilite port=nz1 bundle=control
  #pragma HLS INTERFACE s_axilite port=nelt bundle=control
  #pragma HLS INTERFACE s_axilite port=ldim bundle=control
  #pragma HLS INTERFACE s_axilite port=return bundle=control

  for (int e=0;e<nelt;e++) {
    ax_e(&w[nx1*ny1*nz1*e], &u[nx1*ny1*nz1*e], &gxyz[nx1*ny1*nz1*2*ldim*e], dxm1, dxtm1, ur, us, ut, wk, nx1, ny1, nz1);
  }
}
```

	Negative Slack	BRAM	DSP	FF	LUT
ax_kernel	0.39	4	182	65210	42524
ax_e	0.39	0	173	62095	39963
local_grad3_t	0.39	0	78	28760	18725
mxm16_1	0.39	0	24	9080	5964
mxm16_4	0.39	0	24	9046	5194
mxm16_3	0.39	0	21	8645	5406
add2	0.01	0	3	1282	1069
local_grad3	0.39	0	72	28311	18239
mxm16	0.39	0	24	9495	5723
mxm16_2	0.39	0	21	9250	5657
mxm16_1	0.39	0	24	9080	5964

Description	Performance GFLOPs	% CPU performance
24 cores of Xeon (Cascade Lake) CPU	65.74	-
Initial FPGA port	0.020	0.03%
Optimised top down for dataflow	0.28	0.43%
Optimise bottom up	27.78	42.26%
Ping-pong buffering	59.14	89.96%
Increase clock frequency to 400 Mhz	77.73	118%

Von-Neumann

Approx. 4000 times difference in performance

Dataflow

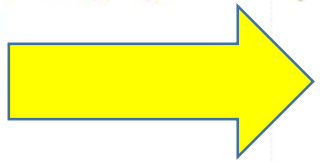
For N=16, Runs performed on a Xilinx Alveo U280

	Pipelined	Latency	Iteration Latency	Initiation Interval	Trip
mxm16_1	-	-	-	-	-
Loop 1	yes	-	108	102	-

Detailed information available at https://arxiv.org/pdf/2011.04981.pdf

Bottom up optimisations

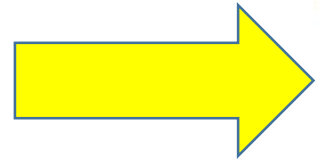
```
for (int i=0; i<N; i++) {
    double d=x*y;
    double j=d*z;
    double p=d*j
    result=p;
}
```



```
for (int i=0; i<N; i++) {
#pragma HLS pipeline II=1
    double d=x*y;
    double j=d*z;
    double p=d*j
    result=p;
}
```

Loop pipelining

```
for (int i=0; i<N; i++) {
#pragma HLS pipeline II=1
    double d=x*y;
    double j=d*z;
    double p=d*j
    result=p;
}
```



```
for (int i=0; i<N; i++) {
#pragma HLS pipeline II=1
#pragma UNROLL FACTOR=4
    double d=x*y;
    double j=d*z;
    double p=d*j
    result=p;
}
```

Loop unrolling

```
double val=0;
for (int i=0; i<N; i++) {
#pragma HLS pipeline II=1
    val=val+external[i];
}
```

Spatial dependency

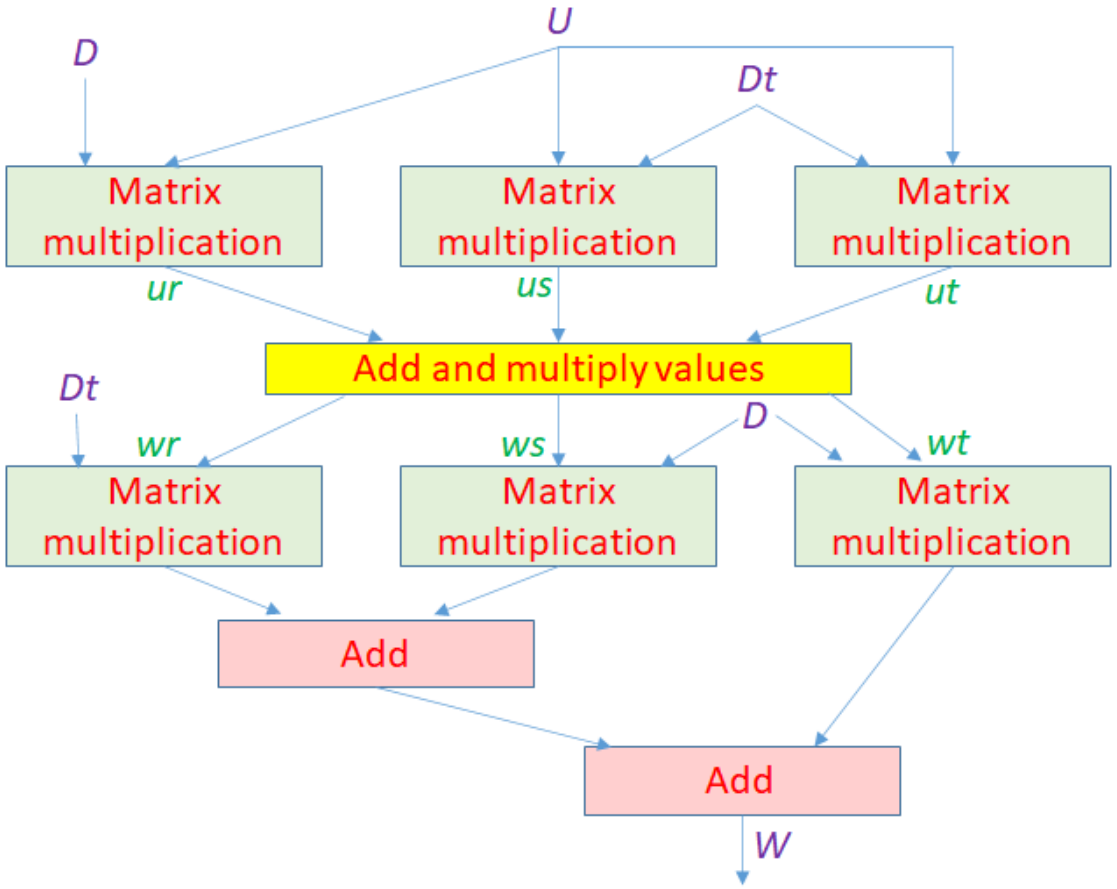
```
for (int i=0; i<N; i++) {
#pragma HLS pipeline II=1
    double d=external_data[i];
    double j=external_data[i+1];
    ...
}
```

Conflict on external port to HBM2/DDR memory

```
double local_data[M];
for (int i=0; i<N; i++) {
#pragma HLS pipeline II=1
    local_data[i-2]=a;
    local_data[i-1]=b;
    double v=local_data[i];
}
```

Conflict on (dual-ported) on-chip BRAM memory

Working top down: Adopting a dataflow design

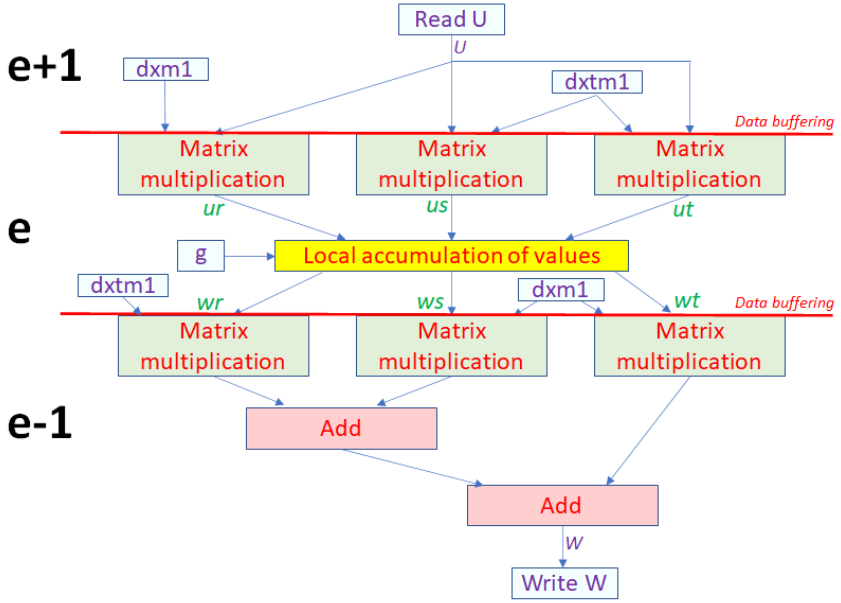


- Each stage is an independent function running concurrently and connected via streams
  - Idea is to keep each part continually fed with data and processing
- Golden rule: Keep the data flowing, each cycle generating a result

Description	Performance GFLOPs	% CPU performance
24 cores of Xeon (Cascade Lake) CPU	65.74	-
Initial FPGA port	0.020	0.03%
Optimised top down for dataflow	0.28	0.43%
Optimise bottom up	27.78	42.26%
Ping-pong buffering	59.14	89.96%
Increase clock frequency to 400 Mhz	77.73	118%

Buffering data between stages

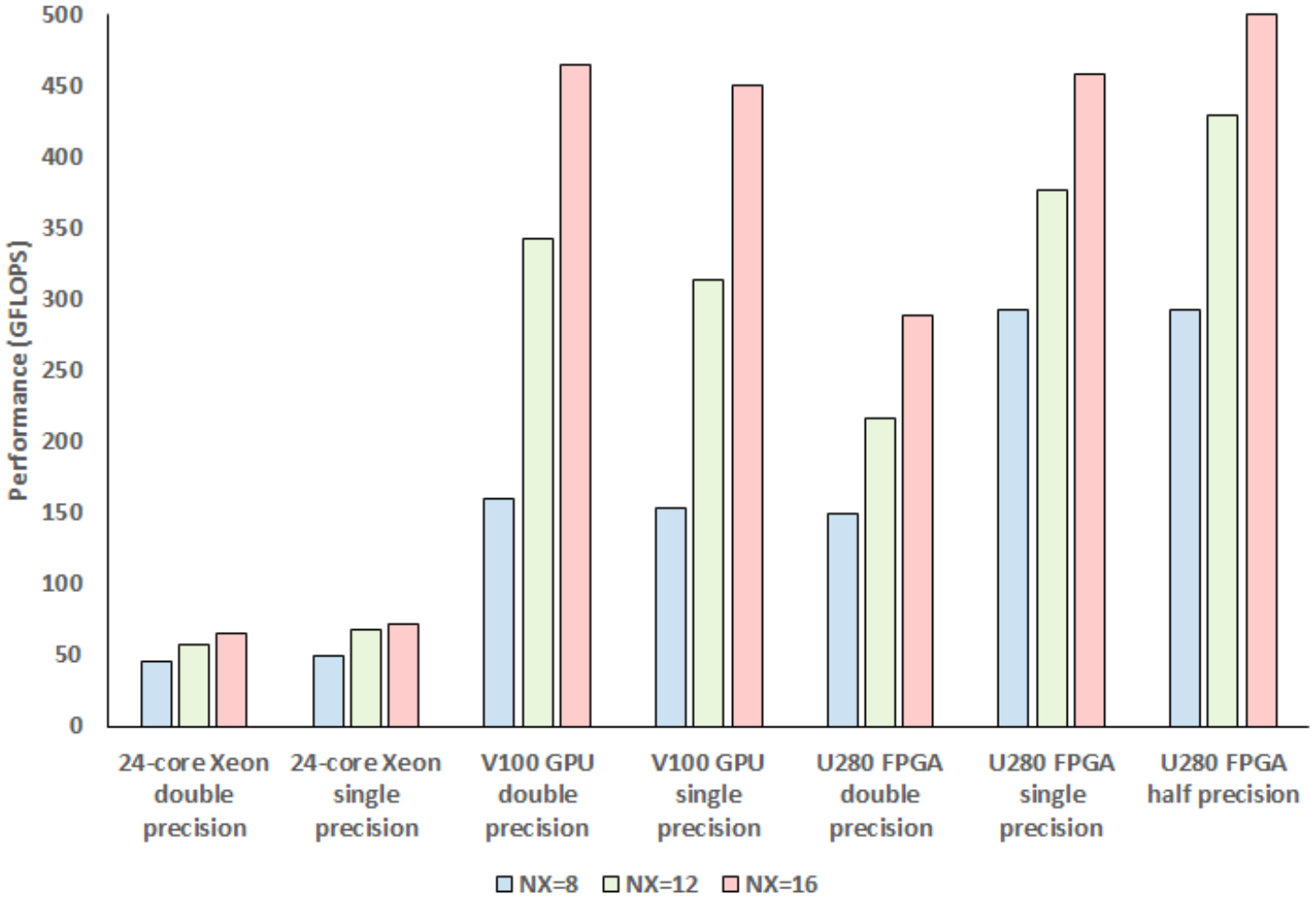
- Adopt ping-pong (double) buffering
  - Works in three phases, loading data for the next element, processing the first three MM for the current element, and the last three MM for the previous element
  - Keeps all parts running concurrently



- Also possible to accurately predict the realistic theoretical best performance our algorithm can deliver
  - Each MM is 31 FLOP/cycle and accumulation is 17/cycle, equals 203 FLOP/cycle. Multiply this by clock frequency for theoretical FLOPS
  - If not achieving this then not keeping data flowing!

Description	Performance GFLOPs	% CPU performance	Theoretical performance	% Theoretical performance
24 cores of Xeon (Cascade Lake) CPU	65.74	-	-	-
Initial FPGA port	0.020	0.03%	6.9 GFLOPs	0.29%
Optimised top down for dataflow	0.28	0.43%	6.9 GFLOPs	4.06%
Optimise bottom up	27.78	42.26%	61 GFLOPs	45.54%
Ping-pong buffering	59.14	89.96%	61 GFLOPs	96.95%
Increase clock frequency to 400 Mhz	77.73	118%	81.2 GFLOPs	95.73%

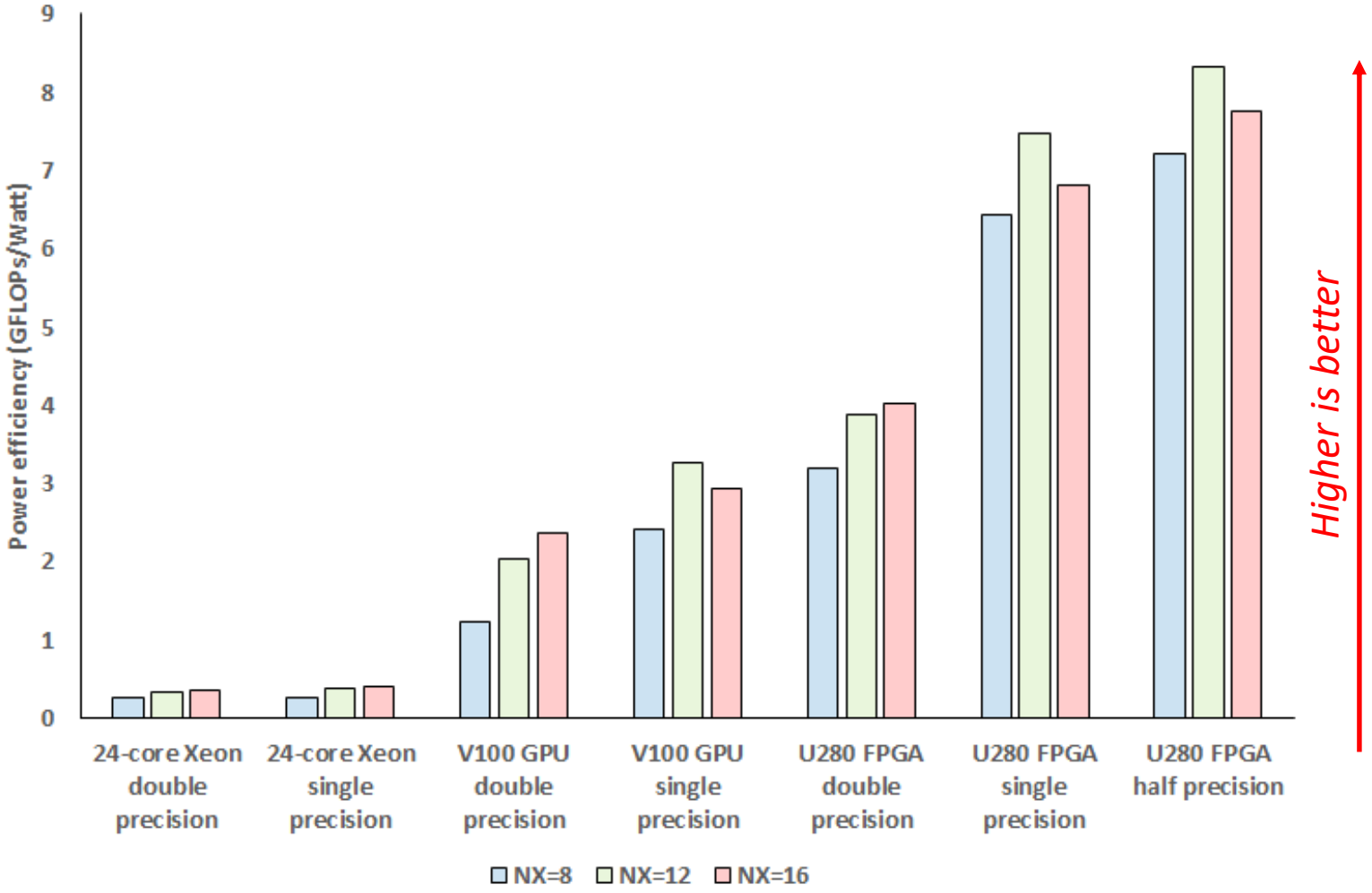
Performance against CPU and GPU



Higher is better

- FPGA runs on an Alveo U280
- CPU on 24-core Cascade Lake Xeon Platinum
- GPU on NVIDIA V100
- 4 FPGA kernels for double precision
  - Depends on exact problem size (NX)
- 7 FPGA kernels for single & half precision
  - Depends on exact problem size (NX)

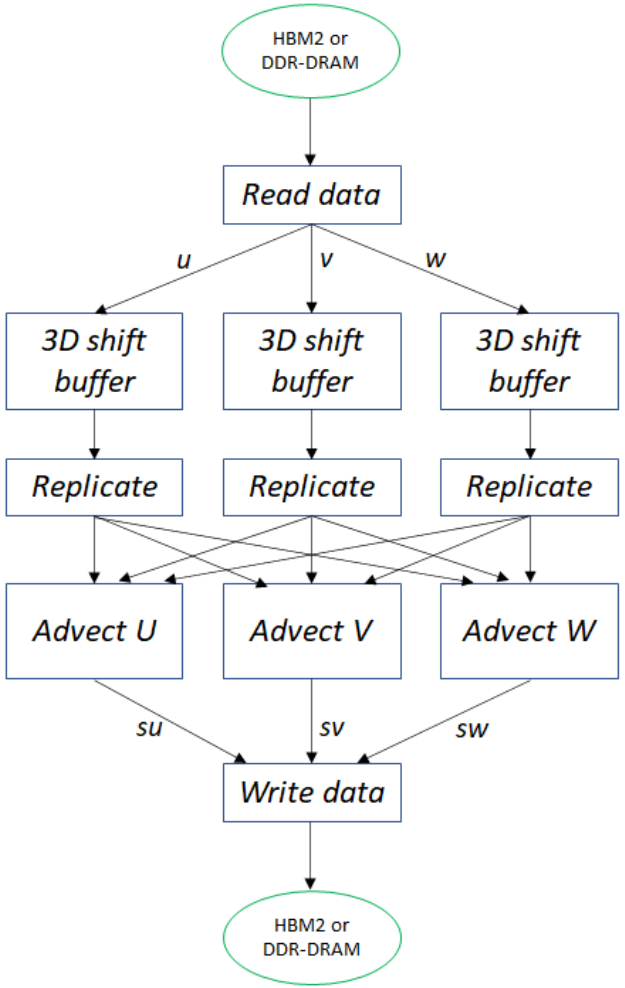
Power efficiency against CPU and GPU



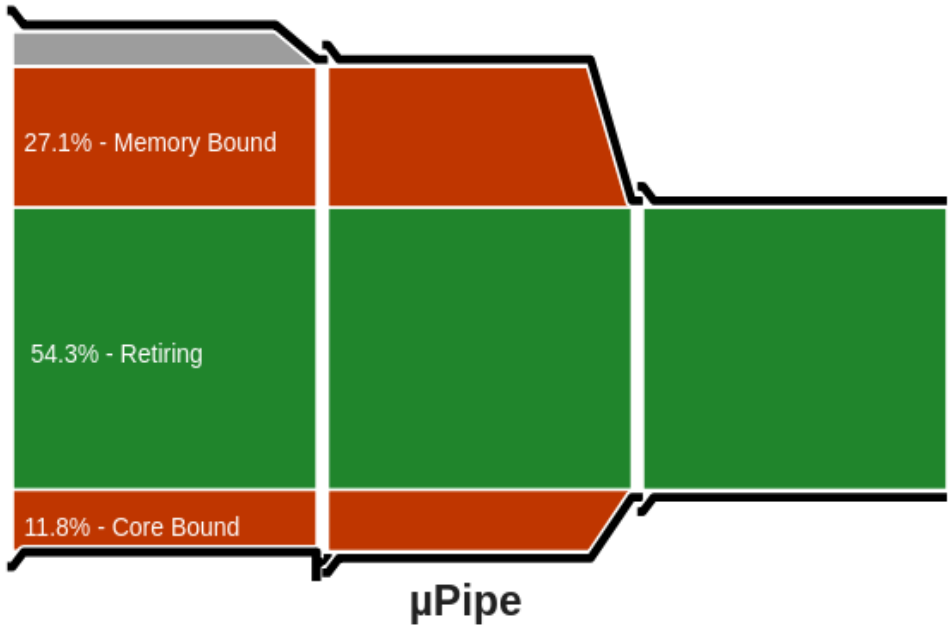
Higher is better

- FPGA runs on an Alveo U280
- CPU on 24-core Cascade Lake Xeon Platinum
- GPU on NVIDIA V100
- 4 FPGA kernels for double precision
  - Depends on exact problem size (NX)
- 7 FPGA kernels for single & half precision
  - Depends on exact problem size (NX)

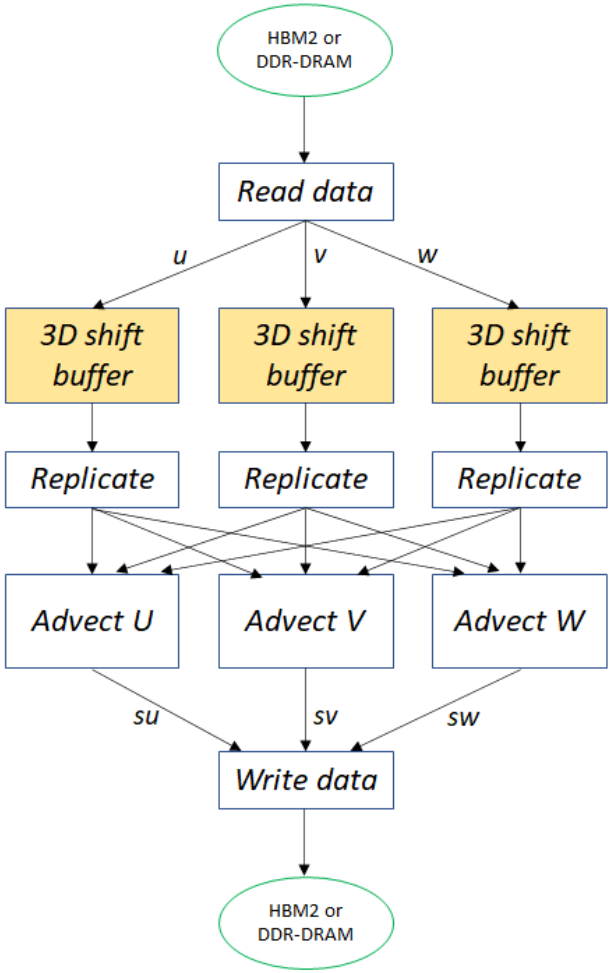
Another example: Atmospheric advection



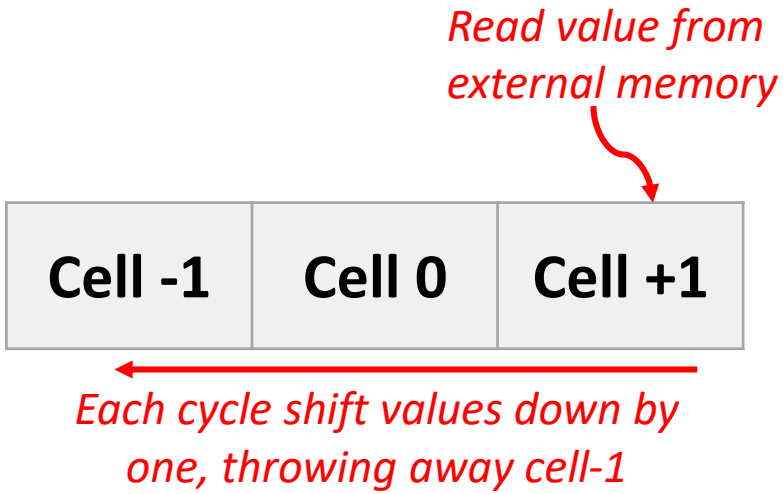
- Part of Met Office NERC Cloud (MONC) model
  - Accounts for around 40% of the model runtime
  - Stencil code working on three fields (U, V, W which is wind in x, y, z dimensions)
- Same ideas as previously
  - Kernel is fairly memory bound
  - Focussed on bottom up and top down dataflow algorithmic techniques
  - Running on both Xilinx Alveo and Intel Stratix-10



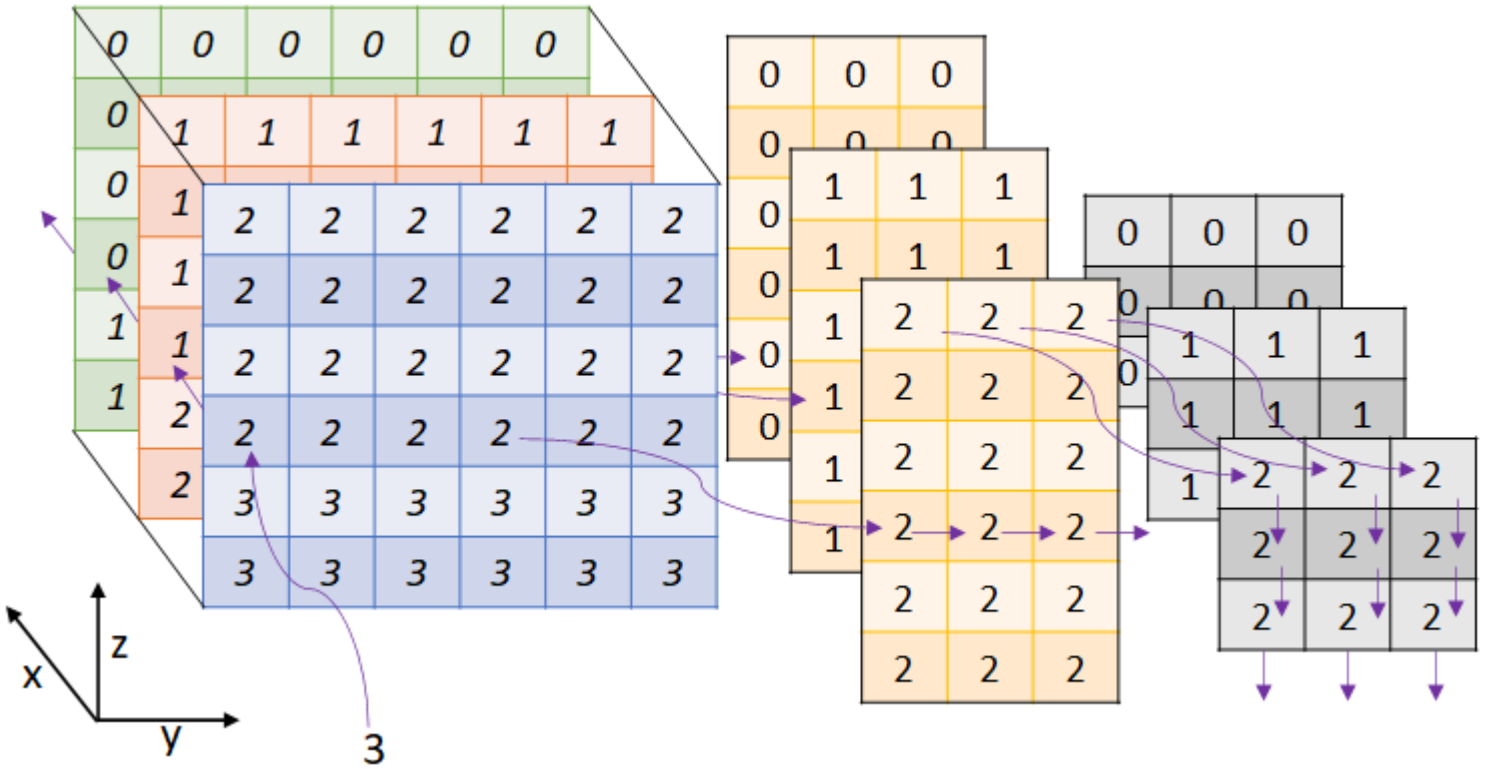
Tailoring caching of data via shift buffer



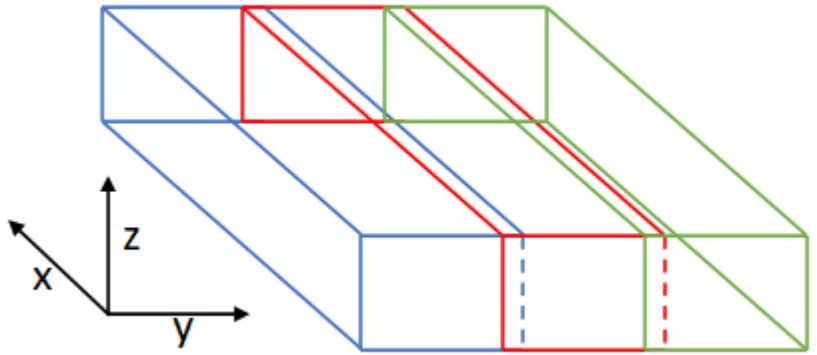
- 3D domain, where stencil computations require up to 27-points to calculate value for each grid cell
- Want to read only one new value from external memory for each field per cycle as otherwise get conflicts on the memory port
  - But need to provide 27 values per cycle to the advect routine in order to achieve a result for each field for each clock cycle
- Use a shift buffer - 1D example



Tailoring caching of data via shift buffer



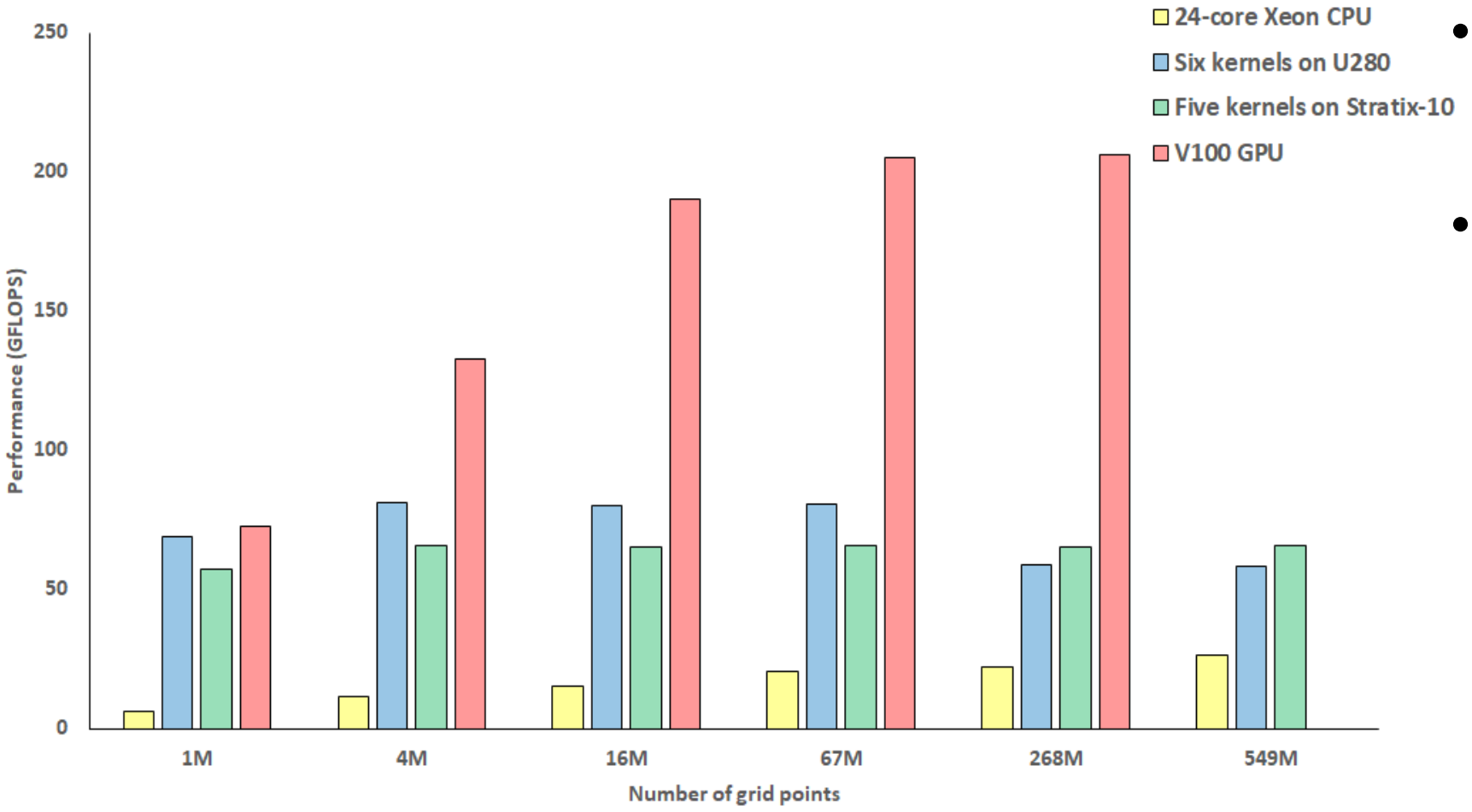
- Have these windows running across the 3D domain
- Generates 27-point stencil each cycle
- Memory on FPGA limits size in Y dimension so work in chunks



- The golden rule of keeping the data flowing and generating a result per cycle necessitated this shift buffer, which then impacted how the code is running and having to chunk in Y

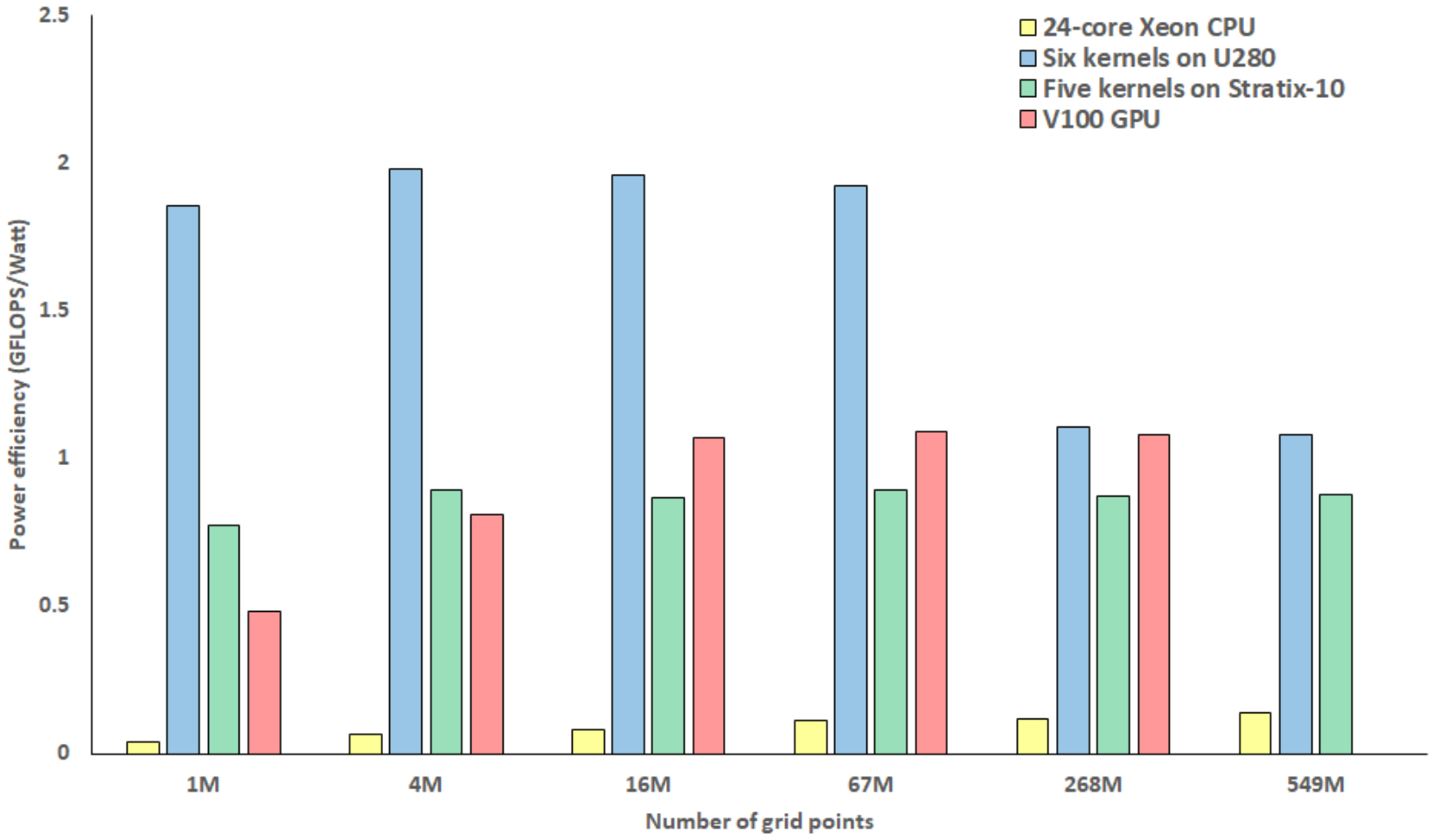
Detailed information available at https://arxiv.org/pdf/2107.13500.pdf

Performance comparison



- FPGAs outperform the CPU by a long way, but GPU is a tough test!
- The Xilinx Alveo U280 tends to outperform the Intel Stratix 10
  - 6 kernels on Alveo vs 5 on the Stratix 10
  - 5 kernels on Stratix 10 are running at 250 MHz, whereas Alveo at 300MHz
  - At largest problem sizes Alveo must use DDR-DRAM rather than 8GB HBM2

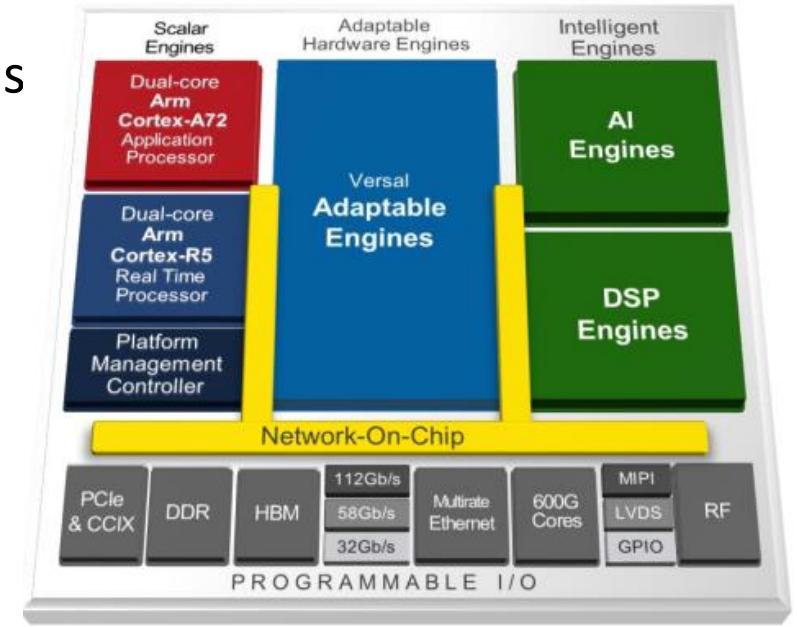
Power efficiency



- Alveo U280 has excellent power efficiency until it has to switch from HBM2 to DDR-DRAM
  - Combination of reduced performance and (slightly) increased power draw
- Higher power draw of Stratix 10 means it is competitive against the GPU, especially for smaller problem sizes

Have a play with FPGAs for your code!

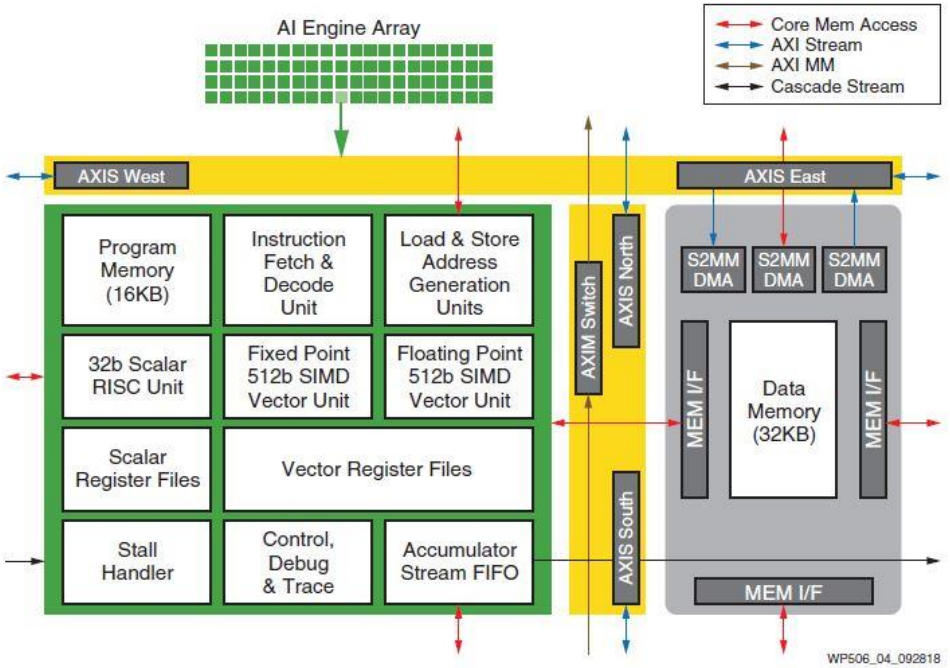
- In EPCC we host the ExCALIBUR H&ES FPGA testbed
  - Give HPC code developers access to FPGAs for their workloads
  - Provides a range of FPGAs to see what works best
  - All tooling preinstalled and provide resource for code development (e.g. building and emulation)
  - In collaboration with UCL and Warwick who are developing enabling software



Access is free, visit https://fpga.epcc.ed.ac.uk for more details

Summary

- It’s an exciting time in HPC, and FPGAs have a potential role to play
  - But important to pick your battles and focus on solving code level challenges suited to FPGAs, they certainly won’t replace GPUs or CPUs!
- FPGAs have become far more capable and next-generation technologies are very exciting
  - Programming FPGAs has become significantly more productive, but need to rethink our algorithms from the perspective of dataflow to achieve good performance
    - Technologies like SYCL on the horizon which look to be very interesting
  - Software experts have a significant contribution to make here around algorithmic techniques and successes stories at the application level.



Gihan Mudalige (Reader (Associate Professor), University of Warwick,
Department of Computer Science)

Multi-Layered Abstractions for Performance Portability - Lessons Learnt and Challenges

Abstract: The rapid evolution of High Performance Computing architectures has resulted in highly complex systems with massively parallel heterogeneous processors, deep and multiple memory hierarchies and interconnects. As a result, maintaining performance as platforms change has becomes increasingly difficult. On the one hand, open standards have been slow to catch up with supporting new hardware, and for many real applications have not provided the best performance achievable from these systems. On the other hand, proprietary solutions have only targeted narrow vendor-specific devices resulting in a proliferation of parallel programming models and technologies. The only practically viable approach to addressing the above issue, is through the development of appropriate application-oriented, high-level programming abstractions such as Domain Specific Languages (DSLs). In this talk I will present how two of the earliest DSLs developed in the UK, OP2 and OPS, have been able to build on multi-layered abstractions techniques to address these challenges. I will detail how simple source-to-source and automatic code-generation techniques using maintainable software technologies have enabled us to develop frameworks that has remained agile in delivering performance portability in the face of nearly a decade of hardware and parallel programming innovations. I will also discuss lessons learnt and my outlook for our DSLs in the run-up to deploying exa-scale systems.

Bio: Dr. Gihan Mudalige is an Reader (Associate Professor) in High Performance Computing at the University of Warwick's Department of Computer Science. His research focuses on the development of next-generation high performance computing numerical simulation software libraries through the utilization of domain–specific languages and high-level abstraction frameworks. As part of this work Dr. Mudalige acts as one of the main developers of the OP2 and OPS embedded domain specific languages (eDSLs), two of the earliest high-level frameworks to demonstrate the utility of these techniques for developing production-grade HPC applications. In 2018, he was awarded a four-year Royal Society Industry Fellowship with Rolls-Royce plc., focusing on developing future-ready massively-parallel CFD simulations for Exascale HPC systems. Previously Dr. Mudalige worked as a Research Associate and Senior Researcher at the University of Oxford’s eResearch Centre before joining the Warwick Computer Science faculty in 2016. He has also worked as a research intern at the University of Wisconsin-Madison’s (US) Department of Computer Science and holds a PhD. in Computer Science from the University of Warwick. Dr. Mudalige is a member of the AC.

Multi-Layered Abstractions for Performance Portability - Lessons Learnt and Challenges

Gihan Mudalige
Royal Society Industry Fellow
Reader (Associate Professor) in High Performance Computing
Department of Computer Science, University of Warwick
g.mudalige@warwick.ac.uk

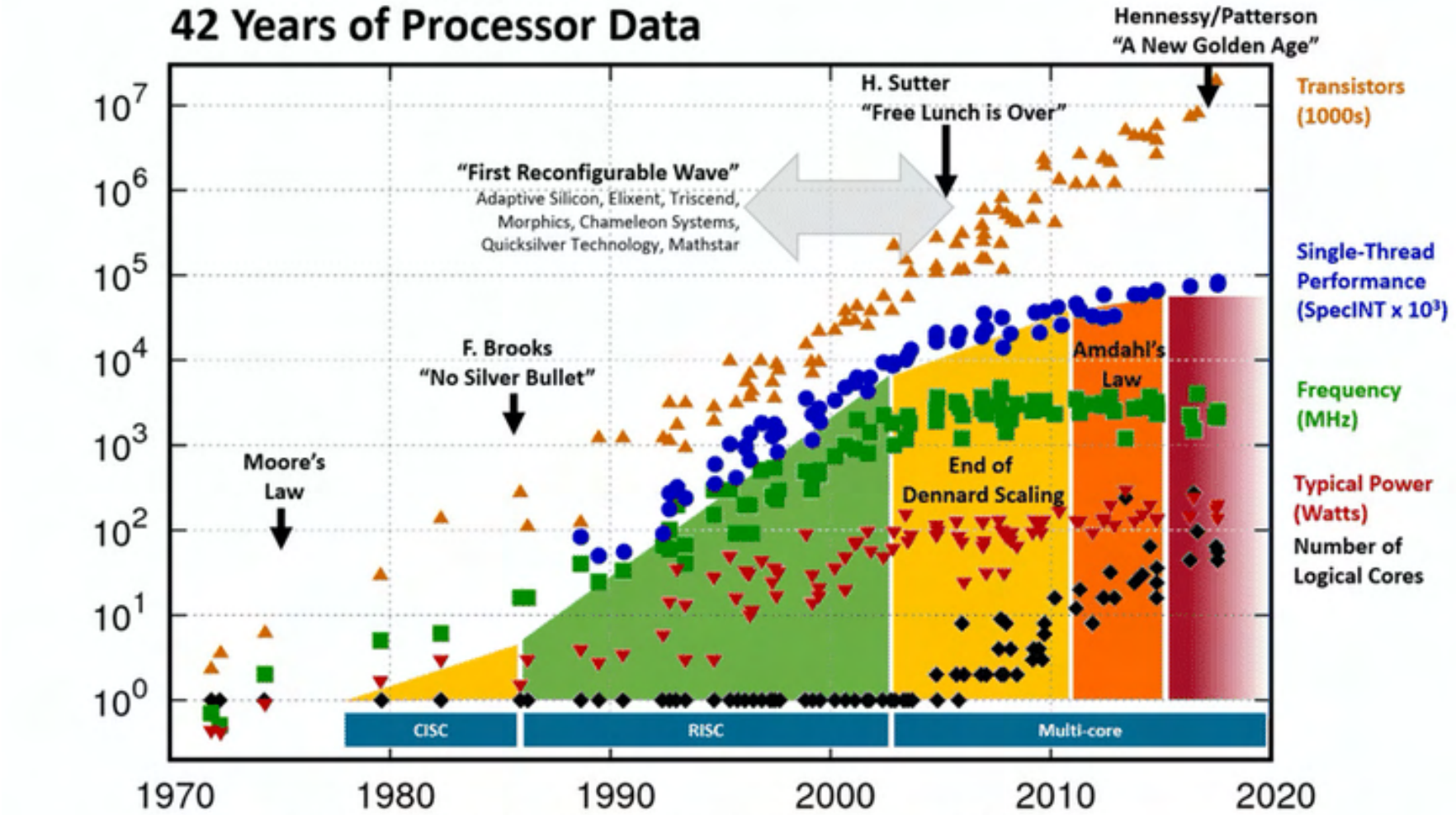
Joint work with:

Istvan Reguly @ PPCU,
Kamalavasan Kamalakannan, Arun Prabhakar and others at the HPSC group @ Warwick
Neil Sandham and team @ Southampton, Dario Amirante @ Surrey,
Mike Giles @ Oxford, Paul Kelly and many more @ Imperial,
Rolls-Royce plc., NAG, UCL, STFC, IBM and many more industrial and academic collaborators.

10th December 2021 – Computing Insight UK 2021

WARWICK
THE UNIVERSITY OF WARWICK

Single thread speedup is dead – Must exploit parallelism



Hennessy and Patterson, Turing Lecture 2018, overlaid over "42 Years of Processors Data"
https://www.karlrupp.net/2018/02/42-years-of-microprocessor-trend-data/;"First Wave" added by Les Wilson, Frank Schirrmeister
Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2017 by K. Rupp

The Hail Mary pass !

http://www.theemike.com/mikes-free-football-comic-book-hail-mary-pass/

"The semiconductor industry threw the equivalent of a Hail Mary pass when it switched from making microprocessors run faster to putting more of them on a chip - doing so without any clear notion of how such devices would in general be programmed."

David Patterson, University of California - Berkeley 2010

Diverse Hardware Landscape – Compounded by the Race to Exascale !

- Traditional CPUs
  - Intel, AMD, ARM, IBM
  - multi-core (> 20 currently)
  - Deep memory hierarchy (cache levels and RAM)
  - longer vector units (e.g. AVX-512)
- GPUs
  - NVIDIA (A100), AMD (MI200) , Intel (Xe GPUs)
  - Many-core (> 1024 simpler SIMT cores)
  - CUDA cores, Tensor cores
  - Cache, Shared memory, HBM (3D stacked DRAM)
- Heterogeneous Processors
  - Different core architectures over the past few years
  - ARM big.LITTLE
  - NVIDIA Grace.Hopper
- XeonPhi (discontinued)
  - Many-core – based on simpler x86 cores
  - MCDRAM (3D stacked DRAM)
- FPGAs
  - Dominated by Xilinx and Intel
  - Various configurations
  - Low-level language / HLS tools for programming
  - Significant energy savings
- DSP Processors
  - Phytium / The Chinese Matrix2000 GPDSP accelerator (Yet to be announced Chinese Exascale system ?)
- TPUs, IPUs ….
- Quantum ?

But .. Even more diverse ways to programming them !

OpenMP,
SIMD,
CUDA, OpenCL,
OpenMP4.0, OpenACC,
SYCL/OneAPI,
HIP/ROCm,
MPI, PGAS
Task-based (e.g Legion)
and others ….

- Open standards (e.g OpenMP, SYCL) – so far have not been agile to catch up with changing architectures
- Proprietary models (CUDA, OpenACC, ROCm, OneAPI) – restricted to narrow vendor specific hardware
- Need different code-paths/parallelization schemes to get the best performance
  - E.g. Coloring vs atomics vs SIMD vs MPI vs Cache-blocking tiling for unstructured mesh class of applications
- What about legacy codes ? There is a lot of FORTRAN code out there !

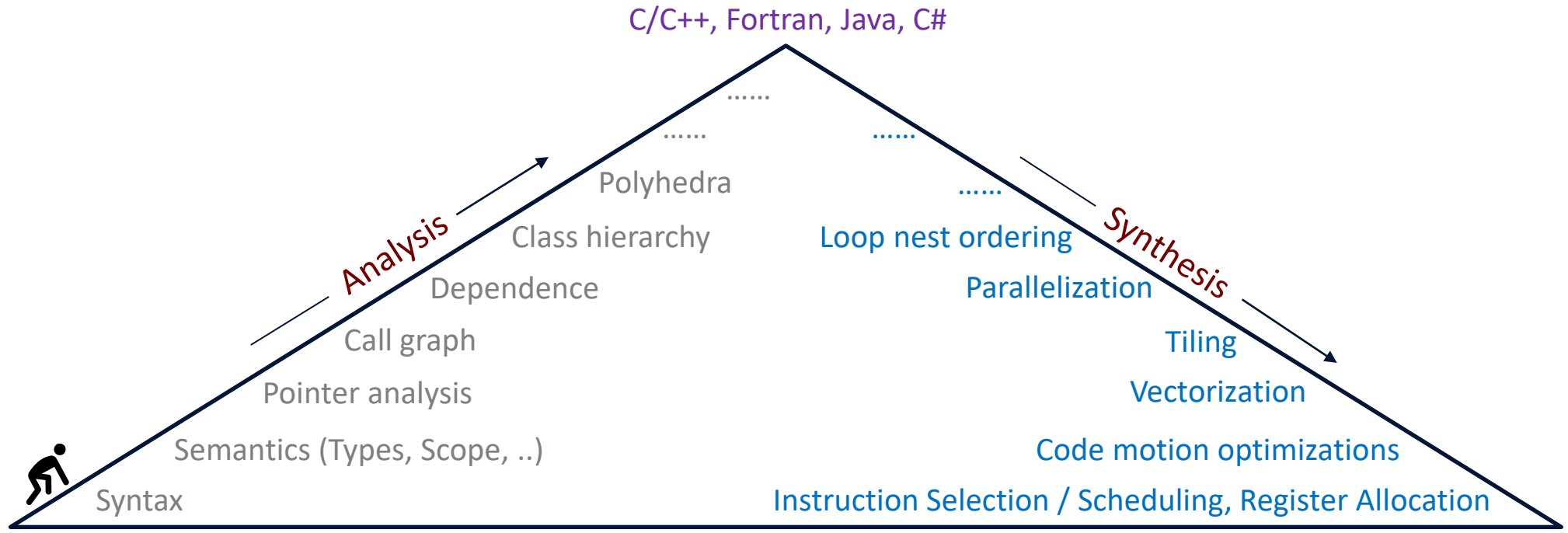
Software Challenge – A Moving Target

- What would an Exa-scale machine architecturally look like ?
- Each new platform requires new performance tuning effort
  - Deeper memory/cache hierarchies and/or shared-memory (non-coherent)
  - Multiple (heterogeneous) memory spaces (device memory/host memory)
  - Complex programming skills set needed to extract best performance on the newest architectures
- Not clear which architectural approach is likely to *win* in the long-term
  - Cannot be re-coding applications for each *new* type of architecture or parallel system
  - Nearly impossible for re-writing legacy codes
- Need to future-proof applications for their continued performance and portability
  - If not – significant loss of investment : applications will not be able to make use of emerging architectures

Outline

- ❑ Motivation ✓
- ❑ Raising the Level of Abstraction
- ❑ OP-DSLs
- ❑ Codes and Projects using OP2/OPS
- ❑ Future Plans – ExCALIBUR
- ❑ Challenges and Lessons Learnt
- ❑ Conclusions

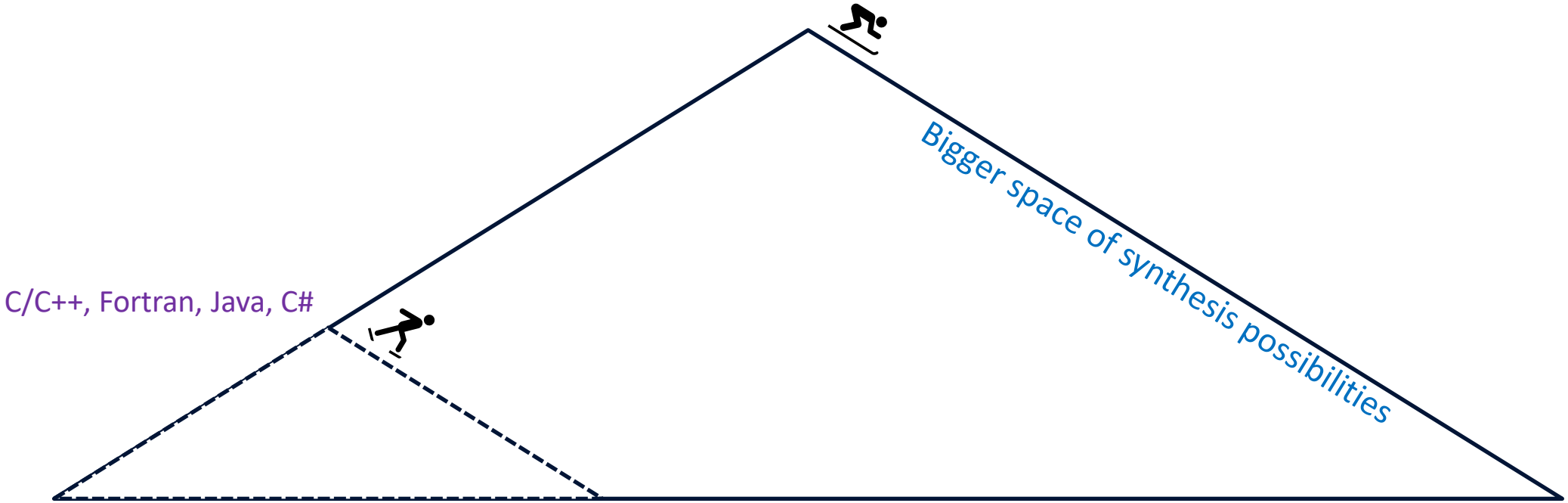
The Level of Abstraction - Climbing the Analysis Hill and Generating Code



Adapted from: *Synthesis versus Analysis: What Do We Actually Gain from Domain-Specificity?* Keynote talk at the LCPC 2015. Paul H. J. Kelly (Imperial College London)

- Classical compiler have two halves : Analysis and Synthesis
- The higher you can get to (in analysis) the bigger the space of code synthesis possibilities
- If you start at a lower level – climbing higher is a struggle
  - Difficult to ensure optimizations are safe (e.g. data races, pointer aliasing)
  - Sometimes, impossible to extract richer information (e.g. data partitioning/layouts, memory spaces)
  - Limits the optimizations possible
- Compounding the issue - the way code is written by (most) people will not be easy to analyse !

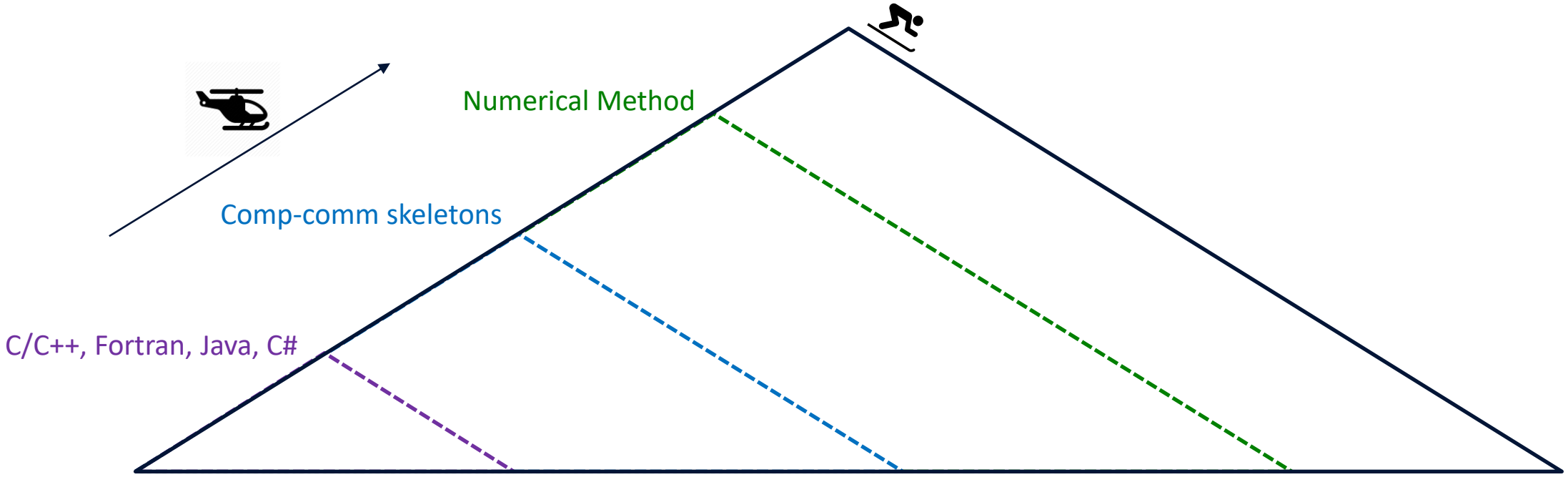
The Level of Abstraction



Adapted from: *Synthesis versus Analysis: What Do We Actually Gain from Domain-Specificity?* Keynote talk at the LCPC 2015. Paul H. J. Kelly (Imperial College London)

- If you can start higher
  - Results in a bigger space of code synthesis possibilities
  - Could they give the same (or better) performance as code written by hand ?
  - Could these possibilities include targeting different (parallel) architectures ?
- How can you start higher ?

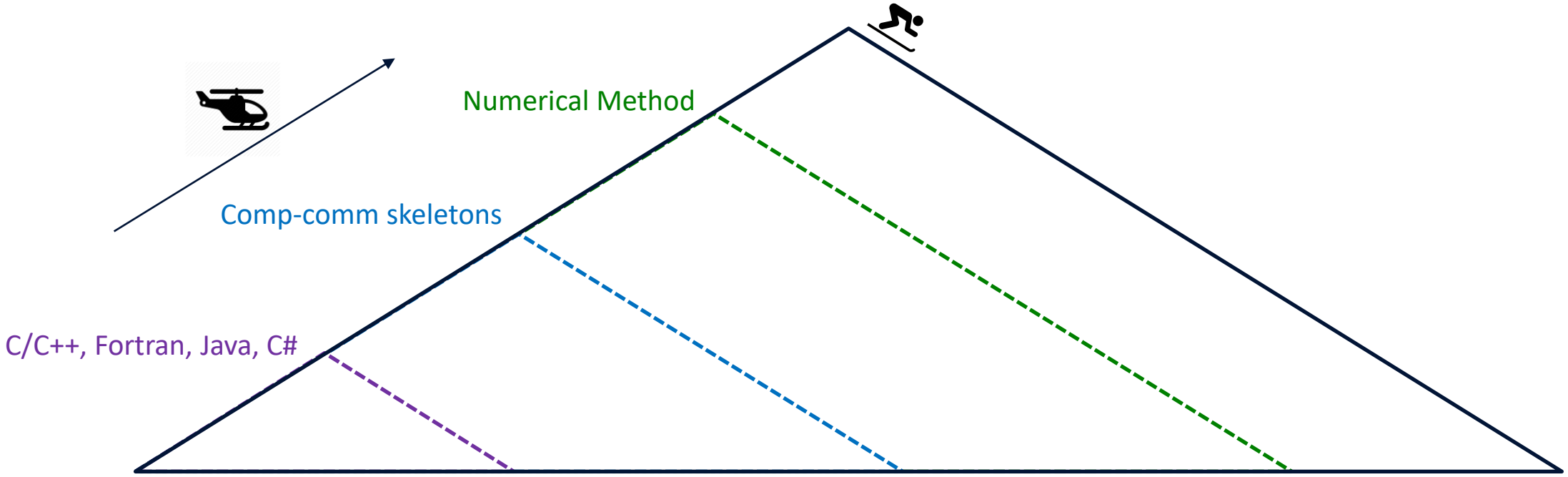
Domain Specific Abstractions



Adapted from: *Synthesis versus Analysis: What Do We Actually Gain from Domain-Specificity?* Keynote talk at the LCPC 2015. Paul H. J. Kelly (Imperial College London)

- Rise the abstraction to a specific domain of variability
- Concentrate on a narrower range (class) of computations
  - Computation-Communications skeletons - Structured-mesh, Unstructured-mesh, … 7 Dwarfs [Colella 2004] ?
  - (higher) Numerical Method - PDEs, FFTs, Monte Carlo …
  - (even higher) Specify application requirements, leaving implementation to select radically different solution approaches

Domain Specific Abstractions



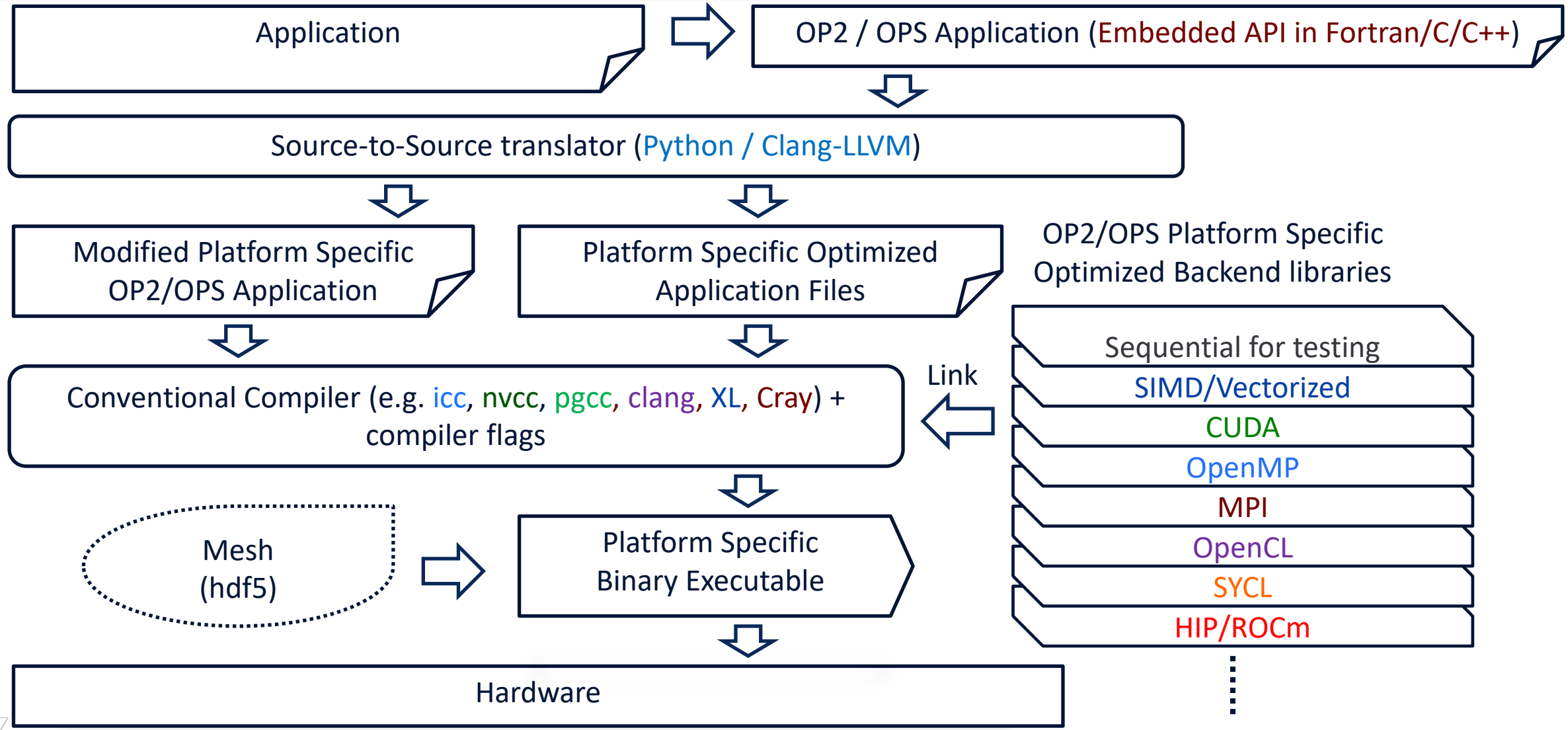
Adapted from: *Synthesis versus Analysis: What Do We Actually Gain from Domain-Specificity?* Keynote talk at the LCPC 2015. Paul H. J. Kelly (Imperial College London)

- If you get the abstraction right, then:
  - Can isolate numerical methods from mapping to hardware
  - Can reuse a body of optimizations/code generation expertise/techniques for this class (or numerical method) to match target hardware

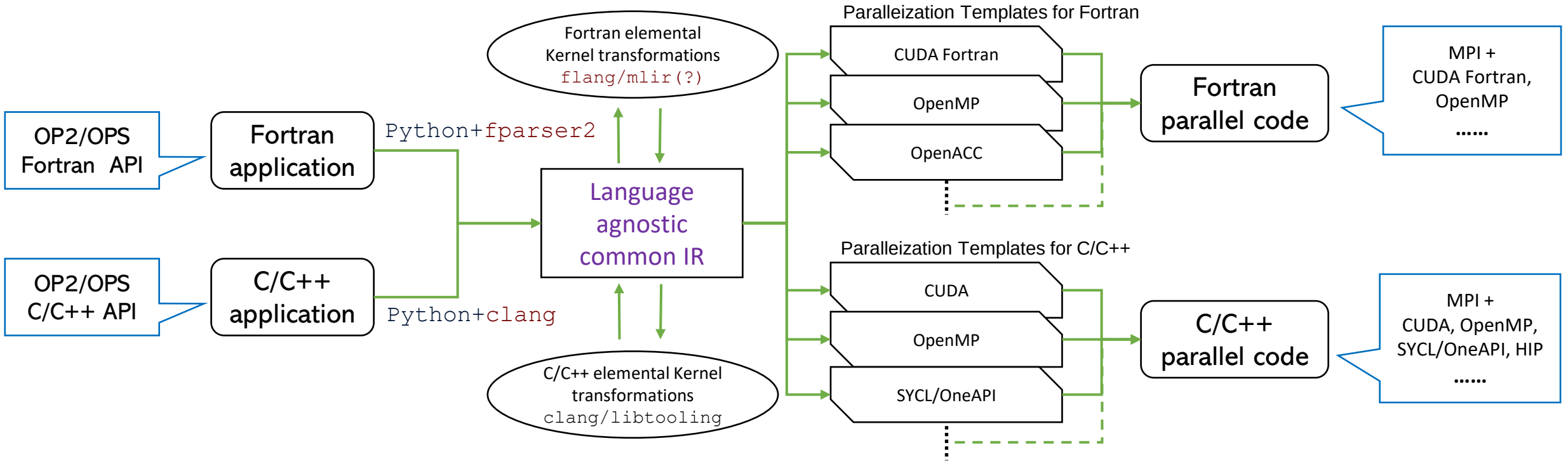
OP-DSL

- Separation of Concerns (… back in 2010 !)
  - Specify the problem – not the implementation
  - Leverage the best implementation for the target context
  - Can be many contexts - hardware, programming model, parameters etc.
- Domain Specific API
  - Get application scientists to pose the solution using domain specific constructs – provided by the API
  - Handling data done *only* using API – contract with the user
- Restrict writing code that is difficult (for the compiler) to reason about and optimize
  - *"OP2 and OPS are a straightjacket"* – Mike Giles
  - Build in safe guards so that user cannot write bad code !
- Implementation of the API left to a lower level
  - Target implementation to hardware – can use best optimizations
  - Automatically generate implementation from specification for the context
  - Exploit domain knowledge for better optimisations - reuse what we know is best for each context

Application Development



OP2/OPS Code Generation



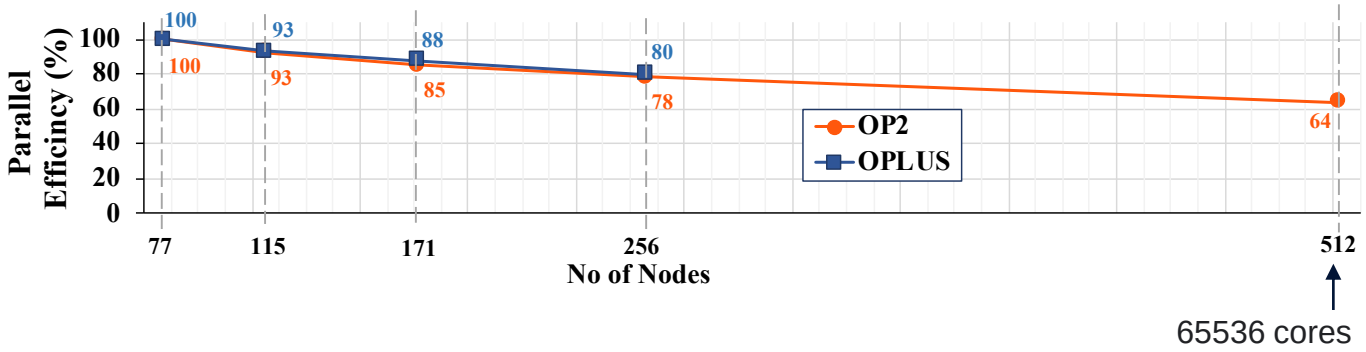
- Simplest Code generation / translation
  - Intermediate representation is simply the loop descriptions + elemental kernels
  - Generated parallel code can be viewed and understood by a human !
- Multi-layered – no opaque / black box layers
- Built with well supported / long-term technologies - Python, Clang/libtooling, [flang, mlir]

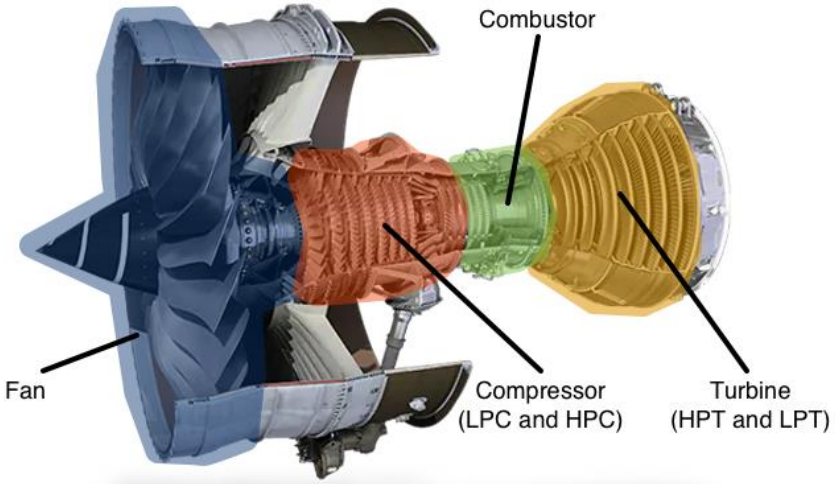
Code Synthesis possibilities

- Auto-parallelization
  - Target different hardware and programming models (SIMD, SIMT, SPMD, Task parallelism?)
  - Sophisticated orchestration of parallelizations – handle data races to match the context
- Full responsibility for data layout and movement
  - Data Layout – SoA - AoS , distributed memory partitioning, local block partitioning
  - Data movement – MPI halo creation and exchange, host/device data movement (memory spaces)
  - Communication avoidance – computation vs communication balance, cache-blocking tiling
- Load-balancing
  - Across heterogeneous processor architectures
- More ..
  - Automatic checkpointing
  - Runtime compilation (JIT)
  - Insitu visualization ?
  - Uncertainty quantification ?

PRODUCTION APPS - EPSRC PROSPERITY PARTNERSHIP – ASIMOV PROJECT

- Virtual certification of Gas Turbine Engines
- Main consortium with partners – EPCC, Warwick, Oxford, Cambridge, Bristol and Rolls-Royce plc.





Grand Challenge : Rig250 – Compressor from DLR

- 1-10 passage full annulus
- Aim: 1 Rev in 24 hours (ARCHER2)
- Achieved : 1 Rev in 11 hours (ARCHER2 32k cores / 256 nodes)
- Predicted : 1 Rev in less than 5 hours (408 V100 GPUs / ~100 nodes)
  1 Rev in less than 6 hours (168 A100 GPUs / ~ 64 nodes)

  less than ½ or ¼ th of the ARCHER2 machine size

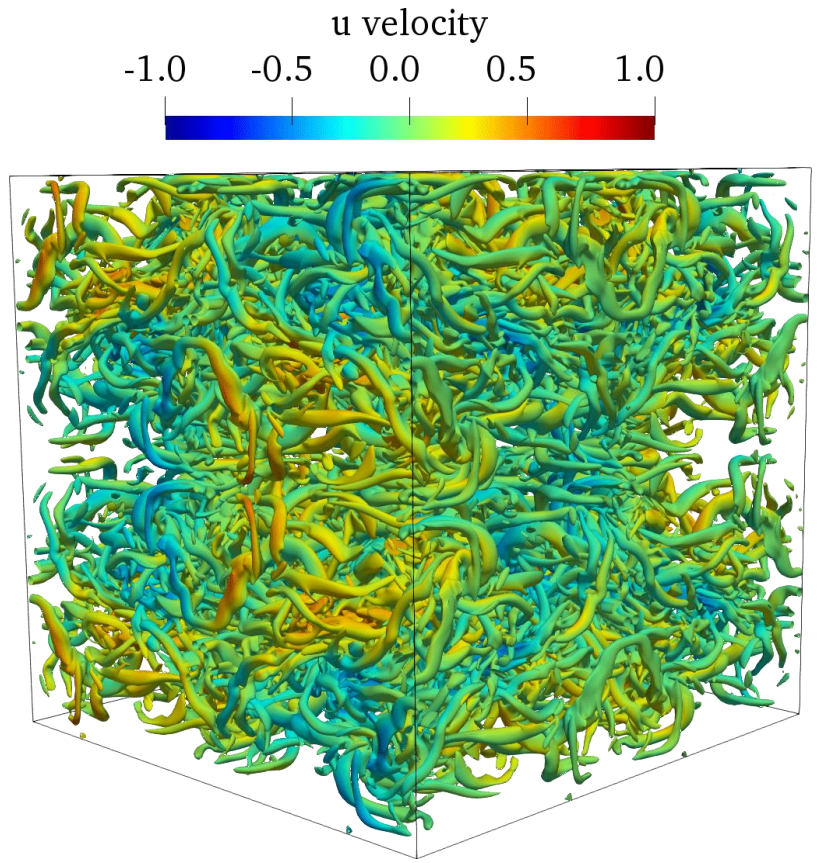
Production Apps – OpenSBLI (Uni. Of Southampton)

- Compressible Navier-Stockes solver
  - With shock capturing WENO/TENO
  - 4th order Finite Difference
  - Single/double precision
- OpenSBLI is a Python framework
  - Write equations in SymPy expressions
  - OPS code generated

```
ndim = 3
sc1 = "**{\'scheme\':\'Teno\'}"
# Define the compresible Navier-Stokes equations in Einstein notation.
mass = "Eq(Der(rho,t), - Conservative(rhou_j,x_j,%s))" % sc1
momentum = "Eq(Der(rhou_i,t) , -Conservative(rhou_i*u_j + KD(_i,_j)*p,x_j , %s) + Der(tau_i_j,x_j) )" % sc1
energy = "Eq(Der(rhoE,t), - Conservative((p+rhoE)*u_j,x_j, %s) - Der(q_j,x_j) + Der(u_i*tau_i_j ,x_j) )" % sc1
stress_tensor = "Eq(tau_i_j, (mu/Re)*(Der(u_i,x_j)+ Der(u_j,x_i) - (2/3)* KD(_i,_j)* Der(u_k,x_k)))"
heat_flux = "Eq(q_j, (-mu/((gama-1)*Minf*Minf*Pr*Re))*Der(T,x_j))"
# Numerical scheme selection
Avg = RoeAverage([0, 1])
LLF = LLFTeno(teno_order, averaging=Avg)
cent = Central(4)
rk = RungeKuttaLS(3, formulation='SSP')
# Specifying boundary conditions
boundaries[direction][side] = IsothermalWallBC(direction, 0, wall_eqns)
# Generate a C code
alg = TraditionalAlgorithmRK(block)
OPSC(alg)
```

OpenSBLI

https://opensbli.github.io/

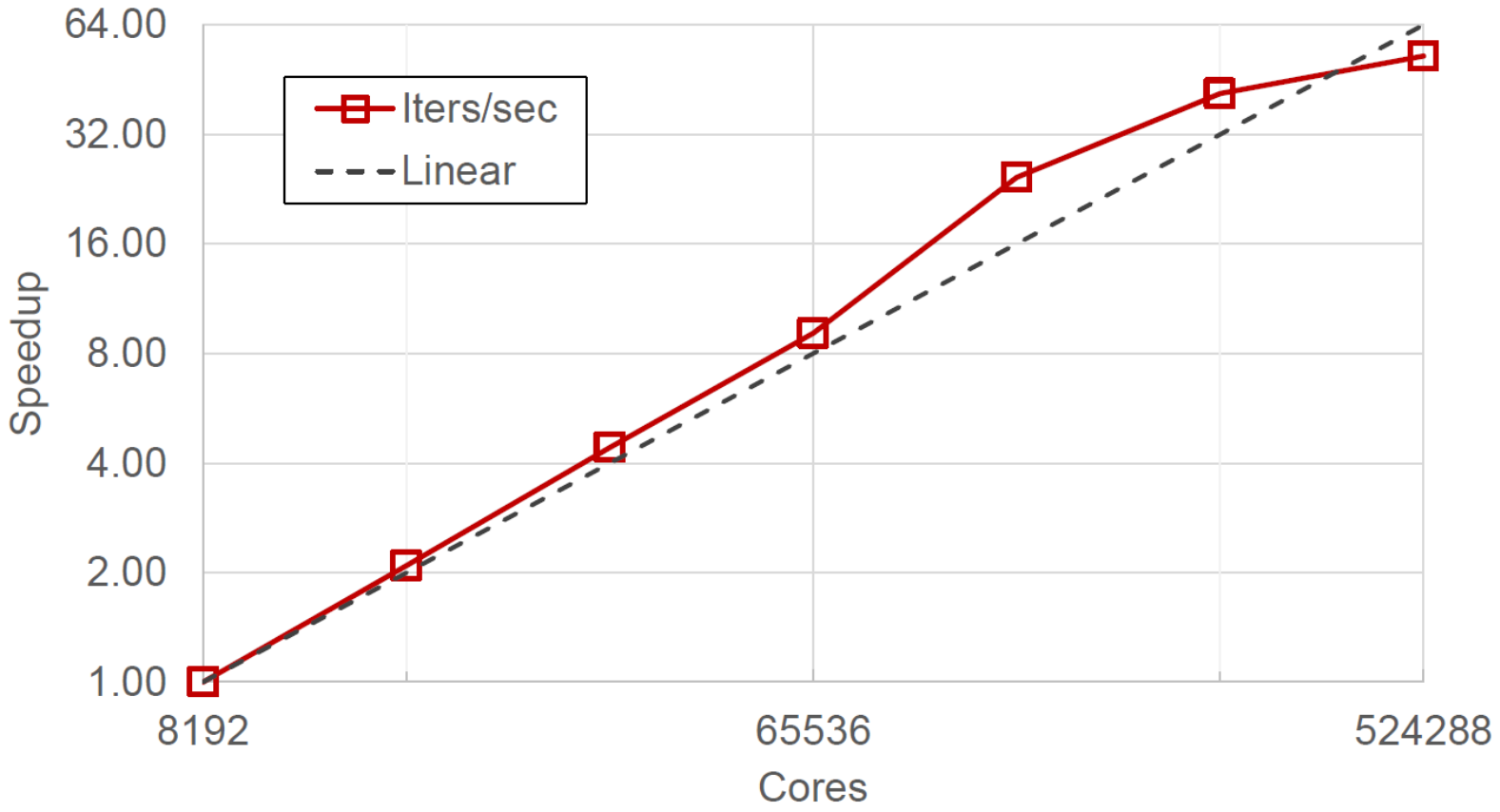


Jacobs, C. T., Jammy, S. P., Sandham N. D. (2017). OpenSBLI: A framework for the automated derivation and parallel execution of finite difference solvers on a range of computer architectures. *Journal of Computational Science*, 18:12-23, DOI: 10.1016/j.jocs.2016.11.001

OpenSBLI on ARHCER2

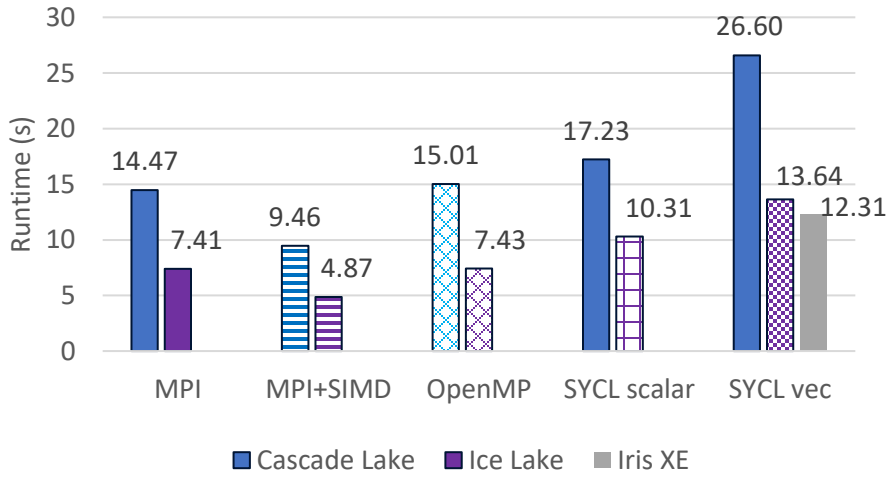
- Taylor – Green Vortex Problem – ARCHER2 benchmark
  - Strong Scaling - 1024^3 Mesh
  - Double precision
  - Speedup calculated from 1000 iterations – includes start up time.

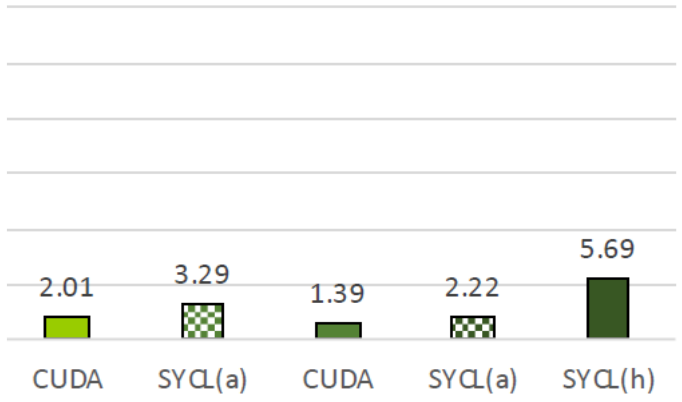
From recent benchmarking runs done by Andrew Turner and the ExCALIBUR Benchmarking team (Oct 2021)



New Parallel Programming Models / Languages - OP2 Generating SYCL

- MG-CFD – Multigrid CFG MiniAPP:
  - NASA Rotor37, 4 multigrid levels, 8M edges
  - Generate Parallelization using OP2
  - Intel compilers - from oneAPI
  - Intel MPI - for MPI, SIMD, OpenMP, MPI+OpenMP
- GPUs – NVIDIA P100 and V100, AMS Radion VII, Intel Iris XE MAX
- CPUs – single socket only to avoid NUMA issues:
  - Intel(R) Xeon(R) Gold 6226R CPU @ 2.90GHz, 16 cores
  - Intel(R) Xeon(R) Platinum 8360Y @ 2.40 GHz, 36 cores
- SYCL compilers - Intel OneAPI 2021.4 and HipSYCL





I.Z. Reguly, A.M.B. Owenson, A. Powell, S.A. Jarvis, and G.R. Mudalige, *Under the Hood of SYCL – An Initial Performance Analysis With an Unstructured-mesh CFD Application*, International Supercomputing Conference (ISC 2021), June 2021

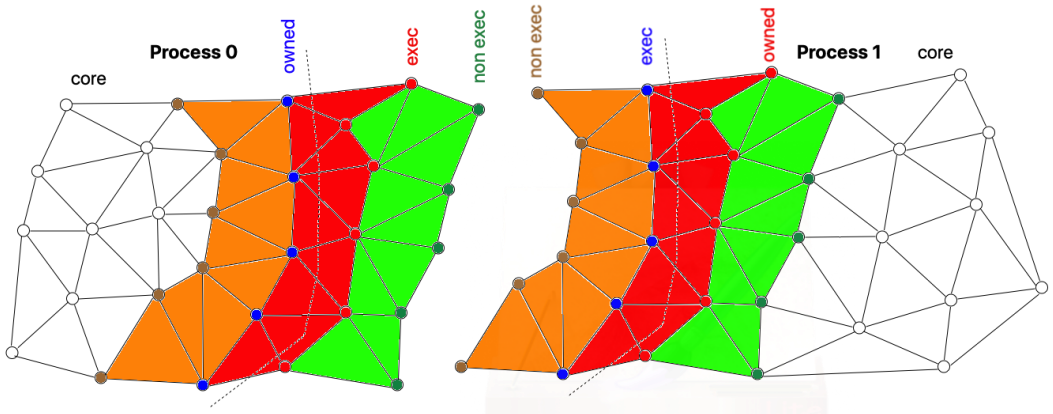
I.Z. Reguly. *Performance of DPC++ on Representative Structured/Unstructured Mesh Applications.* Intel DevSummit at SC21

OneAPI/SYCL

- OP2 and OPS can generate SYCL paralleizations
  - Structured-mesh / Regular applications have good performance portability
  - But various execution strategies needed for unstructured-mesh (irregular) applications to avoid data races
- Key challenge: understanding mapping from SYCL code (SIMT abstraction) to the hardware
  - Reasonably trivial for GPU architectures, where the hardware is a good fit for SIMT
  - Still problematic for SIMD architectures (such as CPUs)
  - OneAPI is quite aggressive about vectorization, and the sub-group API really helps with mapping to SIMD.
  - Performance improving.
- SYCL a much more productive alternative to OpenCL, and performance is improving rapidly
  - But the challenges in terms of performance productivity remain
  - Need multiple code paths for different architectures – e.g. Coloring vs Atomics

Exotic Optimizations - Communication Avoiding Algorithms

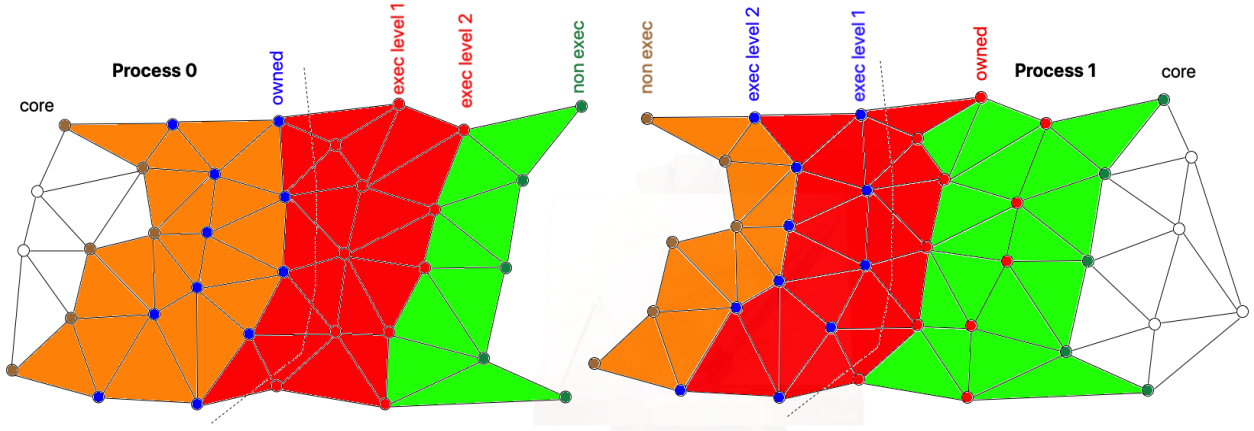
With MPI communication



- Standard OP2 redundant execution over one halo level
- Less computations per node

```
op_par_loop(adt_calc, "adt_calc", cells, …) // MPI Comm
op_par_loop(res_calc, "res_calc", edges, …) // MPI Comm
op_par_loop(bres_calc,"bres_calc", bedges,… ) // MPI Comm
op_par_loop(update, "update", cells,… ) // MPI Comm
```

Without MPI Communication



- Extend halo by one further level
- Redundant compute over both levels
- MPI Comm now avoided – but more computations per node

```
loop_chain_start {
// do all the MPI comms here – with 1 large message per neighbour
op_par_loop(adt_calc, "adt_calc", cells, …);
op_par_loop(res_calc, "res_calc", edges, …);
op_par_loop(bres_calc,"bres_calc", bedges,… );
op_par_loop(update, "update", cells,… );
} loop_chain_end
```

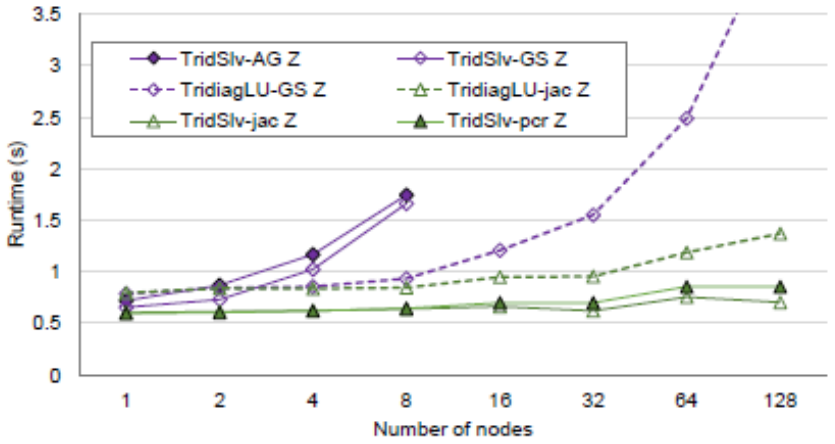
Direct Solvers - Multi-Dim Tridiagonal Solvers on Clusters of GPUs

$$a_i u_{i-1} + b_i u_i + c_i u_{i+1} = d, \quad i = 0, 1, ..., N-1$$

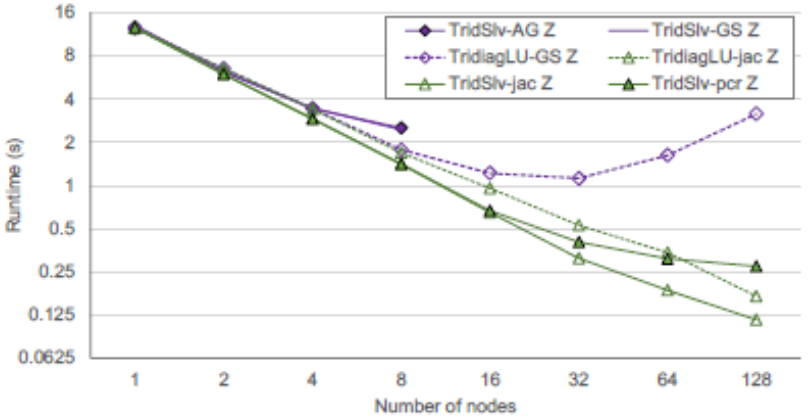
$$\begin{bmatrix} b_0 & c_0 & 0 & \dots & 0 \\ a_1 & b_1 & c_1 & \dots & 0 \\ 0 & a_2 & b_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a_{N-1} & b_{N-1} \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \\ u_2 \\ \vdots \\ u_{N-1} \end{bmatrix} = \begin{bmatrix} d_0 \\ d_1 \\ d_2 \\ \vdots \\ d_{N-1} \end{bmatrix}$$

pcr – parallel cyclic reduction (PCR) solver
Jac – Jacobi iterative solver
AG – All Gather
GS – Gather Scatter
HC – host copy
GD – GPU direct

ARCHER2



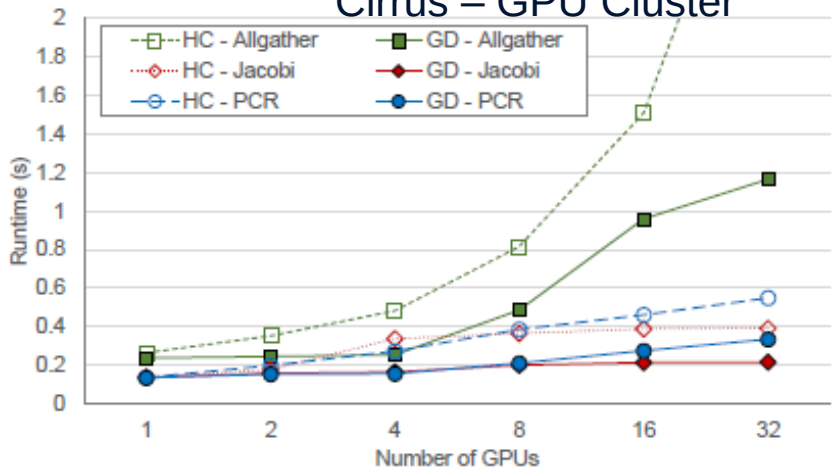
(b) TridSlv vs TridiagLU weak scaling



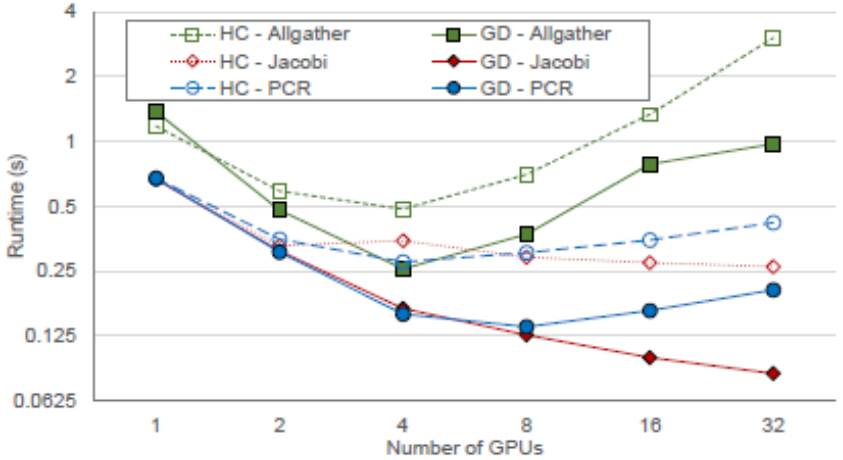
(d) TridSlv vs TridiagLU strong scaling

Weak Scaling : 512^3 per Node
Strong Scaling : 512× 512 × 8192

Cirrus – GPU Cluster



(b) TridSlv, weak scaling, HC vs GD

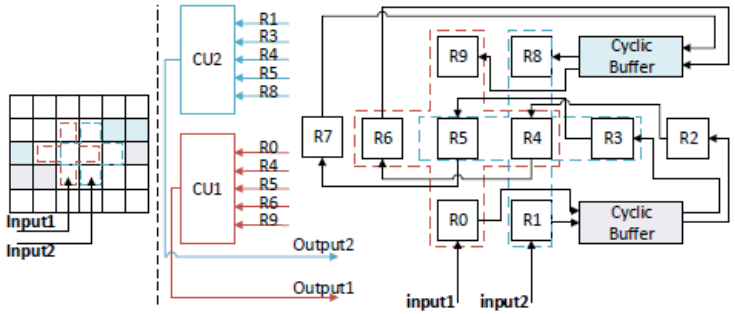


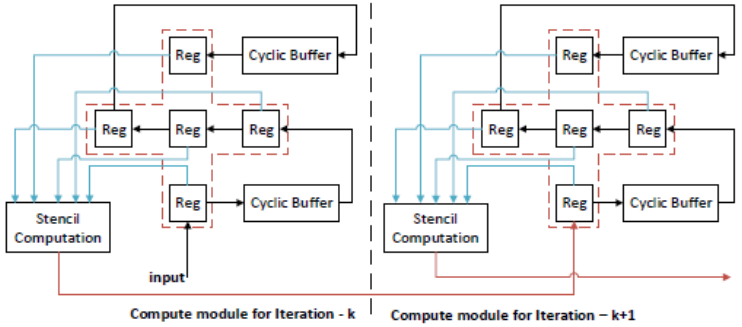
(d) TridSlv, strong scaling, HC vs GD

Weak Scaling : 512^3 per GPU
Strong Scaling : 512× 512 × 2048

G.D. Balogh, T. Flynn, S. Laizet, G.R. Mudalige, I.Z. Reguly. *Scalable Many-core Algorithms for Tridiagonal Solvers,* in 2021 Computing in Science & Engineering, vol. , no. 01, pp. 1-1, 5555. doi: 10.1109/MCSE.2021.3130544

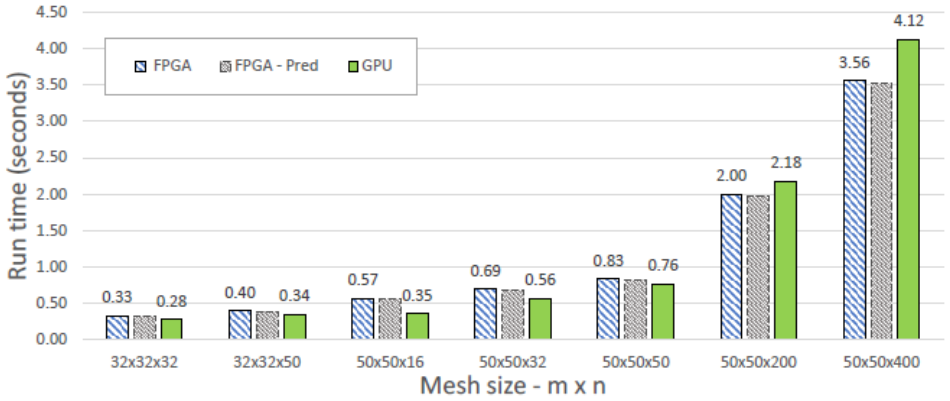
STENCILS ON FPGAS



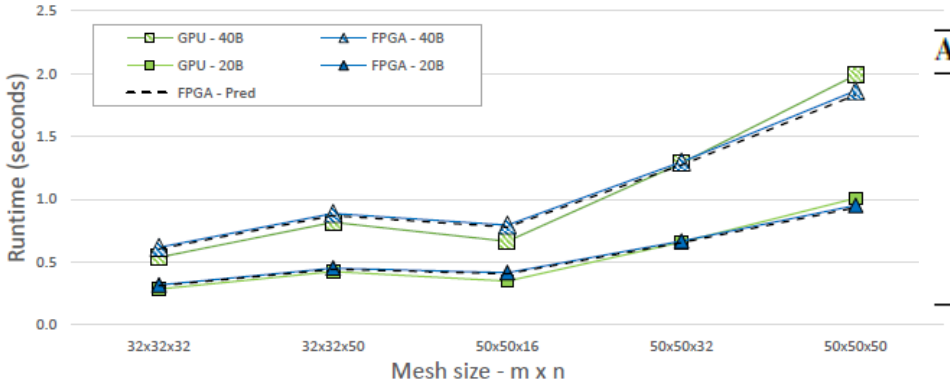


FPGA	Xilinx Alveo U280 [28]
DSP blocks	8490
BRAM / URAM	6.6MB (1487 blocks) / 34.5MB (960 blocks)
HBM	8GB, 460GB/s, 32 channels
DDR4	32GB, 38.4GB/s, in 2 banks (1 channel/bank)
Host	Intel Xeon Silver 4116 @2.10GHz (48 cores) 256GB RAM, Ubuntu 18.04.3 LTS
Design SW	Vivado HLS, Vitis-2019.2
GPU	Nvidia Tesla V100 PCIe [28]
Global Mem.	16GB HBM2, 900GB/s
Host	Intel Xeon Gold 6252 @2.10GHz (48 cores) 256GB RAM, Ubuntu 18.04.3 LTS
Compilers, OS	nvcc CUDA 9.1.85, Debian 9.11

3D - Reverse Time Migration (RTM) – Forward Pass



(a) Baseline - 1800 iterations



(b) Batching - 180 iterations

Algorithm 1 RTM - Forward Pass

for $i = 0, i < n_{iter}, i++$ **do**
$K_1 = f_{pml}(Y_{25pt}, \rho, \mu) \times dt;\ T = Y + K_1/2$
$K_2 = f_{pml}(T_{25pt}, \rho, \mu) \times dt;\ T = Y + K_2/2$
$K_3 = f_{pml}(T_{25pt}, \rho, \mu) \times dt;\ T = Y + K_3$
$K_4 = f_{pml}(T_{25pt}, \rho, \mu) \times dt$
$Y = Y + K_1/6 + K_2/3 + K_3/3 + K_4/6$
end for

- Competitive runtimes with GPUs
- Even when runtime is inferior to the GPU we get significant energy savings (e.g. over 2x for for the RTM app)

K. Kamalakkannan, **G.R. Mudalige**, I.Z. Reguly, S.A. Fahmy, *High-Level FPGA Accelerator Design for Structured-Mesh-Based Explicit Numerical Solvers*, in 2021 IEEE International Parallel and Distributed Processing Symposium (IPDPS 2021), Portland Oregon, USA, 2021 pp. 1087-1096. doi: 10.1109/IPDPS49936.2021.00117

Multi-Dim Tridiagonal Solvers on FPGAs

$$Ax = d$$

$$a_i u_{i-1} + b_i u_i + c_i u_{i+1} = d, \quad i = 0, 1, ..., N-1 \qquad (1)$$

$$\begin{bmatrix} b_0 & c_0 & 0 & \dots & 0 \\ a_1 & b_1 & c_1 & \dots & 0 \\ 0 & a_2 & b_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a_{N-1} & b_{N-1} \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \\ u_2 \\ \vdots \\ u_{N-1} \end{bmatrix} = \begin{bmatrix} d_0 \\ d_1 \\ d_2 \\ \vdots \\ d_{N-1} \end{bmatrix} \qquad (2)$$

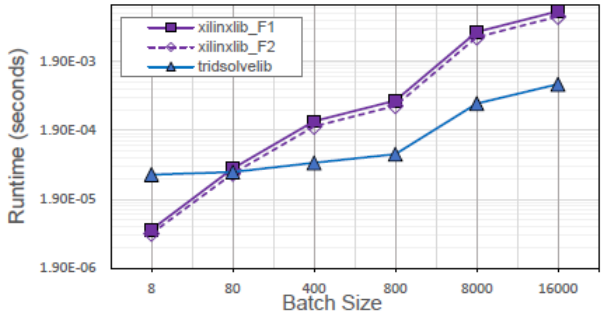
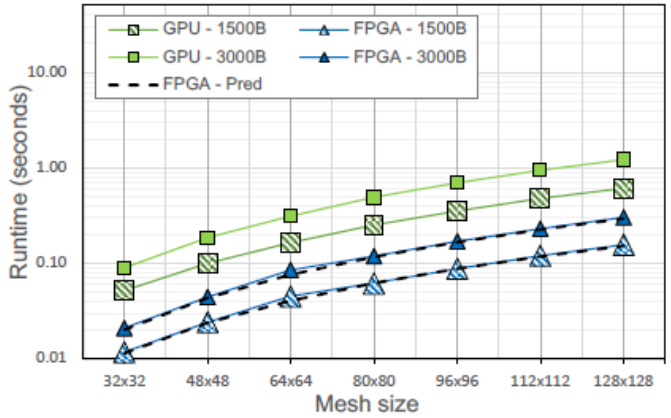


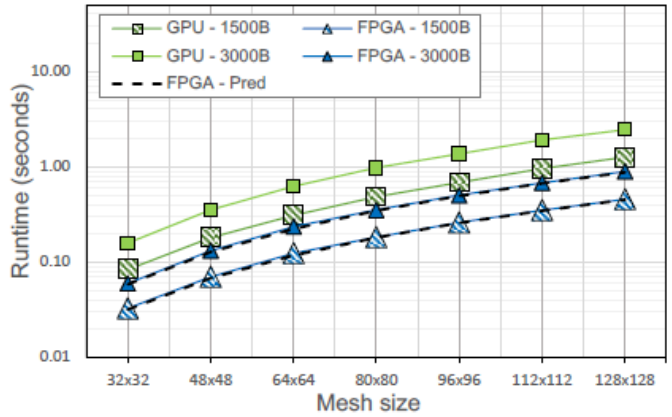
Fig. 4: `tridsolvlib` vs `xilinxlib`, System size-128

Algorithm 3: `3D ADI Heat Application`

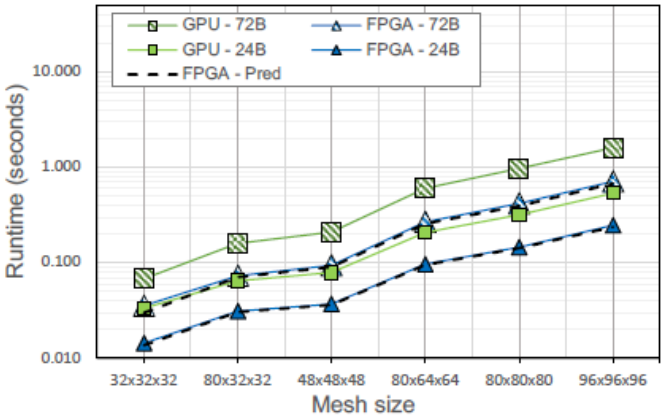
1: **for** $i = 0, i < n_{iter}, i++$ **do**
2: `Calculate RHS :` $d = f_{7pt}(u), a = \frac{-1}{2}\gamma, b = \gamma, c = \frac{-1}{2}\gamma$
3: `Tridslv(x-dim), update` d
4: `Tridslv(y-dim), update` d
5: `Tridslv(z-dim), update` d
6: $u = u + d$
7: **end for**



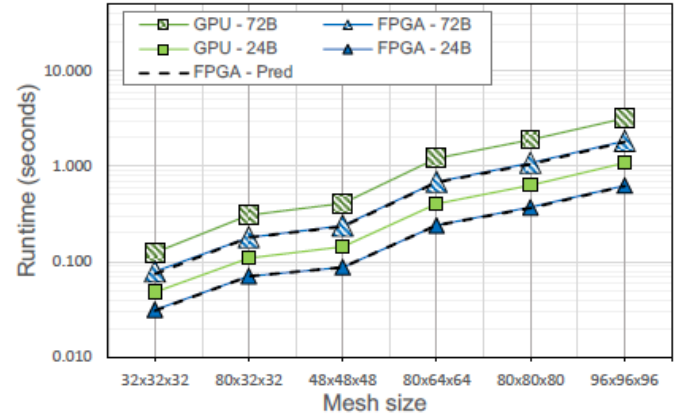
(a) 2D FP32 - 120 iter.$v = 8, f_U = 3, N_{CU} = 3$



(b) 2D FP64 - 120 iter.$v = 8, f_U = 2, N_{CU} = 3$



(c) 3D FP32 - 100 iter.$v = 8, N_{CU} = 6$



(d) 3D FP64 - 100 iter. $v = 8, N_{CU} = 3$

Fig. 5: ADI Heat Diffusion Application Performance

Kamalavasan Kamalakkannan, Istvan Z. Reguly, Suhaib A. Fahmy, and Gihan R. Mudalige. *High Throughput Multidimensional Tridiagonal Systems Solvers on FPGAs* (2021) – Under Review

NON-TRADITIONAL ARCHITECTURES - MULTI-DIM TRIDIAGONAL SOLVERS ON FPGAS

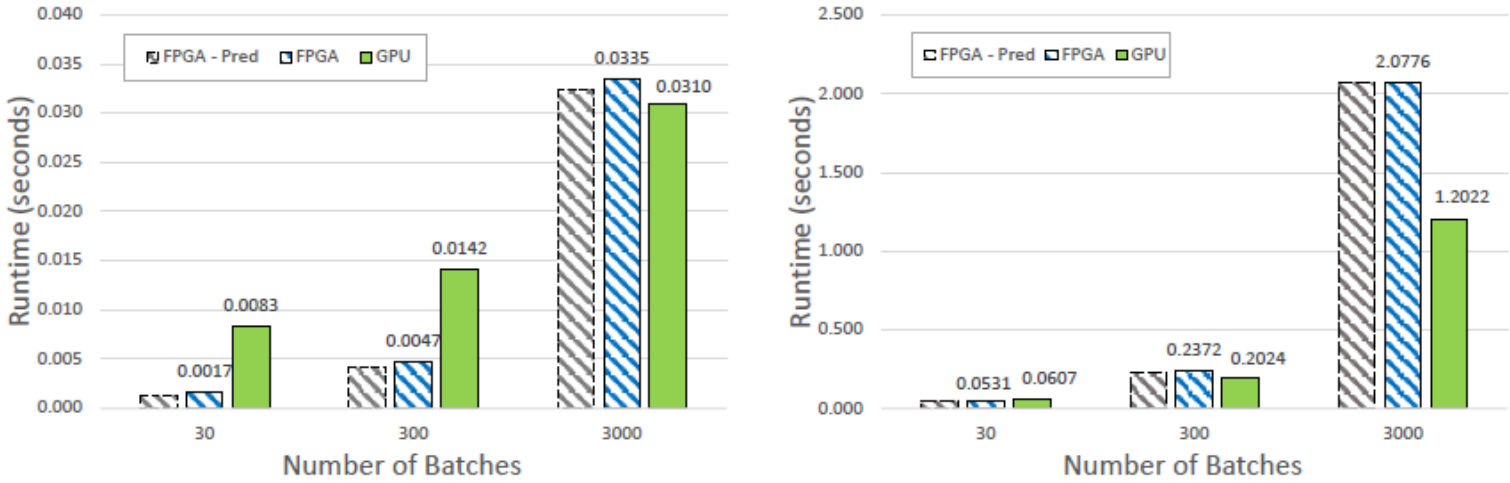
- Stochastic Local Volatility (SLV) model application - high throughput batched implementation on Xilinx FPGAs

Algorithm 4: 2D Heston SLV Backward

```
for i = 0, i < n_iter, i++ do
    hv_pred0(), hv_matrices()
    Tridslv(x-dim)
    hv_pred1(), Tridslv(y-dim)
    hv_pred2(), Tridslv(x-dim)
    hv_pred3(), Tridslv(y-dim)
end for
```

TABLE 5: SLV : Bandwidth, BW (GB/s) and Energy, E, (J)

Batch	40×20 mesh BW			E	
	F	Gx	Gy	F	G
30	55.24	3.04	28.01	0.13	0.45
300	202.31	16.48	176.51	0.35	1.02
3000	281.06	123.84	327.65	2.51	4.75
Batch	100×50 mesh BW			E	
	F	Gx	Gy	F	G
30	124.63	51.28	109.65	3.98	3.76
300	278.87	235.22	238.34	17.79	22.26
3000	318.36	421.77	429.21	155.82	216.40



(a) 40×20 mesh, $v = 1, f_U = 1, N_{CU} = 3$ (b) 100×50, $v = 1, f_U = 1, N_{CU} = 3$

Fig. 7: SLV application performance.

- Competitive runtimes with GPUs
- FPGA solution is over 30% more energy efficient for large batch solves over the GPU

Kamalavasan Kamalakkannan, Istvan Z. Reguly, Suhaib A. Fahmy, and Gihan R. Mudalige. *High Throughput Multidimensional Tridiagonal Systems Solvers on FPGAs* (2021) – Under Review

Other Projects using OP2/OPS

- ETH Zurich – BASEMENT code (Basic Simulation Environment for Computation of Environmental Flows and Natural Hazard Simulations)
  - Flood forecast and mitigation, River morphodynamics, Design of hydraulic structures
  - Finite volume discretisation, cell centred
  - Targeting OP2 for GPU and multi-core parallelisation
- STFC – HiLeMMS project (High-Level Mesoscale Modelling System):
  - high-level abstraction layer over OPS for the solution of the Lattice Boltzmann method
  - Adaptive mesh refinement - Chombo (Lawrence Berkeley National Labs)
- University of Nottingham – CFD code development with OPS
  - Simulation of Turbomachinery flows
  - Implicit solvers using OPS's Tridiagonal Solver API

Current work and Future work - ExCALIBUR Project

- CCP – Turbulence
  - Direct solver libraries – Tri-, penta-, 7-, 9-, 11 diagonal, multi-dimensional solvers
  - Integrate directsolver libraries to be called within OPS
  - OpenSBLI type high-level (Python) framework for XCompact3D – High Order FD framework

- ExCALIBUR Phase 1B – Turbulence at the Exascale
  - Imperial, Warwick, Newcastle, Southampton, Cambridge, STFC collaboration | UKTC and UKCTRF Communities
  - Xcompact3D and Wind Energy, OpenSBLI and Green Aviation, uDALES and Air Quality, SENGA+ and Net-Zero Combustion
  - Extending OPS capability – robust code-gen tools and parallel transformations | support future-proof code development
  - UQ, I/O, Coupling and Visualization
  - Machine Learning Algorithms for Turbulent Flow

- UK AEA Mini-Apps Project
  - Collaboration with University of York
  - Developing Prototype miniApps for UKAEA workload
  - Investigate / advise on performance portability techniques and current state-of-the-art.

UK TURBULENCE CONSORTIUM

UK CTRF UK CONSORTIUM ON TURBULENT REACTING FLOWS

ExCALIBUR

UK Atomic Energy Authority

Challenges – Cost / Effort of Conversion

- Converting legacy code is time consuming
  - Large code base,
  - Defunct 3rd party libs,
  - Fortran 77 or older !
- Difficult to validate code
  - New code giving the same accurate scientific output ?
  - What code should I certify ? High-level code/generated code ?
  - Difficult to convince users to use new code - fear of an opaque compiler / intermediate representation / black box !
- Incremental conversion – loop by loop
  - Simpler than CUDA, but more difficult than OpenACC/OpenMP
  - Automated conversion ?
- Changing user requirements
  - Wanting to use a DSL for doing things beyond what it was intended for !
  - Asking for “back-doors” / “escape hatches” -- leads to poor performance

Challenges – Code-generation

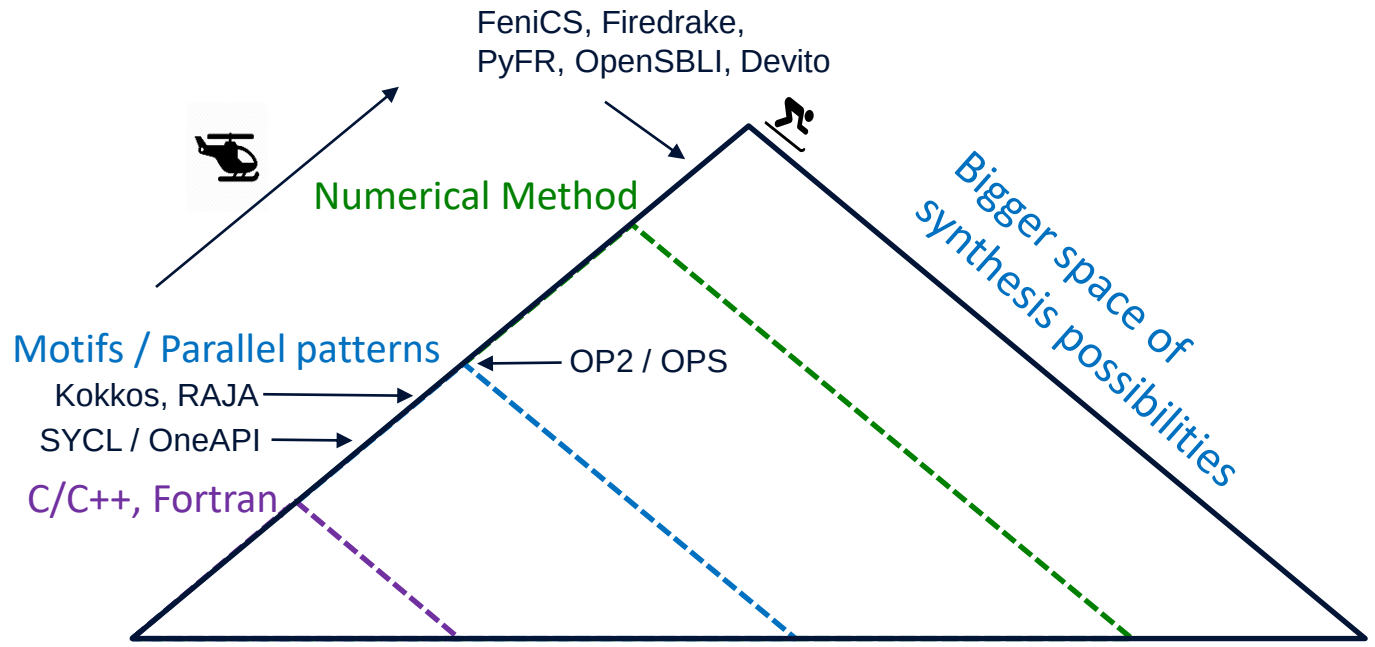
- Tools not entirely mature
  - Currently source-to-source with Python
  - Pushing clang/LLVM source-to-source to do what we want
  - What about Fortran - may be F18/Flang ?
  - MLIR appearing to give some advance capabilities – see ExCALIBUR xDSL project (Tobias Grosser, Paul Kelly et al.)
- Code-generation for more exotic architectures – e.g. FPGAs
  - Large design space
  - Complex source transformations
- Maintainable/long term source-to-source technologies
  - Domain Scientists not having expertise to understand / maintain DSLs

Challenges – Who Maintains the DSL, What DSL to choose ?

- Currently purely done via academic and (small/short term) industrial funding
- Long term funding and maintenance
  - Once established probably will not be different to any other classical library
  - Will require compiler expertise to maintain code generation tools
- What DSL to choose ?
  - Re-use technologies / DSLs – especially code-gen tools (best not to reinvent !)
- Skills Gap
  - Programme in C/C++/Fortran (at a minimum)
  - Knowledge of compilers / code-generation
  - Compete for applicants – Communicate what we do better | impact of HPC / Computational Sciences
  - Salary ☹
  - Contracts ☹

DSLs / High-level Abstractions Gaining Traction !

- FEniCS - PDE solver package - https://fenicsproject.org/
- Firedrake - automated system for the portable solution of PDEs using the finite element method https://www.firedrakeproject.org/
- PyFR - Python based framework for solving advection-diffusion type problems on streaming architectures using the Flux Reconstruction approach - http://www.pyfr.org/
- Devito - prototype DSL and code generation framework based on SymPy for the design of highly optimised finite difference kernels for use in inversion methods - http://www.opesci.org/devito-public
- GungHO project - Weather modelling codes (MetOffice)
- STELLA – DSL for stencil codes, for solving PDEs (Metro Swiss)
- Liszt – Stanford University : DSL for solving mesh-based PDEs - http://graphics.stanford.edu/hackliszt/
- Kokkos – C++ template library – SNL
- RAJA - C++ template libraries - LLNL

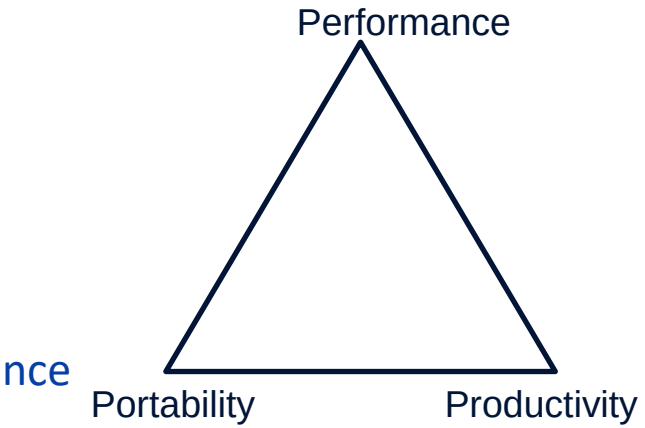


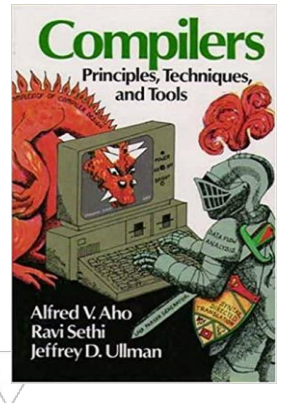
Adapted from: *Synthesis versus Analysis: What Do We Actually Gain from Domain-Specificity?* Keynote talk at the LCPC 2015. Paul H. J. Kelly (Imperial College London)

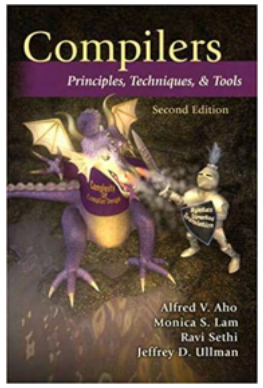
Separation of Concerns – One of the four pillars of ExCALIBUR

Lessons Learnt and Conclusions

- Utilizing domain knowledge will expose things that the compiler does not know
  - Iterating over the same mesh many times without change
  - Mesh is partitioned and colourable
- Compilers are conservative
  - Force it to do what you know is right for your code !
- Let go of the conventional wisdom that higher abstraction will not deliver higher performance
  - Higher abstraction leads to a bigger space of code synthesis possibilities
  - We can automatically generate significantly better code than what (most) people can (reasonably) write
  - Do not destroy performance portability by (hand-) tuning at a very low level to a specific platform







"Fundamentals and abstractions have more staying power than the technology of the moment"
Alfred Aho and Jeffrey Ullman (Turing Award Recipients 2020)

Downloads and More Information

GitHub Repositories

- OP2 – https://github.com/OP-DSL/OP2-Common
- OPS – https://github.com/OP-DSL/OPS

- OP-DSL Webpage - https://op-dsl.github.io/

Contact

Gihan Mudalige (Warwick) - g.mudalige@warwick.ac.uk
Istvan Reguly (PPCU – Hungary) - reguly.istvan@itk.ppke.hu

ACKNOWLEDGEMENTS

- OP2 was part-funded by the UK Technology Strategy Board and Rolls-Royce plc. through the SILOET project, and the UK EPSRC projects EP/I006079/1, EP/I00677X/1 on Multi-layered Abstractions for PDEs.

- OPS was part-funded by the UK Engineering and Physical Sciences Research Council projects EP/K038494/1, EP/K038486/1, EP/K038451/1 and EP/K038567/1 on "Future-proof massively-parallel execution of multi-block applications" and EP/J010553/1 "Software for Emerging Architectures" (ASEArch) project.

- This research is supported by Rolls-Royce plc., and by the UK EPSRC (EP/S005072/1) Strategic Partnership in ComputationalScience for Advanced Simulation and Modelling of Engineering Systems (ASiMoV).

- Gihan Mudalige was supported by the Royal Society Industrial Fellowship Scheme (INF/R1/180012)
- Research was part-supported by the Janos Bolyai Research Scholarship of the Hungarian Academy of Sciences.

- The research has been carried out within the project Thematic Research Cooperation Establishing Innovative Informatic and Info-communication Solutions, which has been supported by the European Union and co-financed by the European Social Fund under grant number EFOP-3.6.2-16-2017-00013.

- OpenSBLI was part-funded by EPSRC grants EP/K038567/1 and EP/L000261/1, and European Commission H2020 grant 671571 "ExaFLOW: Enabling Exascale Fluid Dynamics Simulations

CIUK 2021 Cluster Challenge

Following the success of the CIUK Cluster Challenge in 2020 we were delighted to welcome back our defending champions from Durham University to defend their title in 2021. The challengers were a combined team from Bristol University and Bath University.

As CIUK 2021 returned to a physical conference - as opposed to the online version in 2020 - the format of the competition changed slightly this year, with four online challenges in the months leading up to the conference and four challenges during the conference in Manchester.

After eight exciting, and very closely fought, challenges the team from Bristol / Bath emerged victorious by a narrow margin to take the title and go forward to represent CIUK in the ISC'22 Cluster Challenge competition in the summer.

As always we would like to thank the team members for embracing the spirit of the competition and for giving it their full attention. The result of each individual challenge was very close and in many cases the challenge mentors from our partner companies commented on the quality of the teams work.

We would also like to sincerely thank our partner companies for providing access to their systems for our teams and mentoring them as they worked through the challenges...

Boston, OCF, Lenovo, Alces Flight, Graphcore, Intel, SambaNova and ORock Technologies.

CIUK 2021 Poster Competition

Adam Tuft – Durham University

Otter: An OMPT Tool for Tracing and Visualising Task Creation and Synchronisation in OpenMP Programs

Existing tracing and visualisation tools, which provide a thread-timeline view of task execution as mapped onto threads at runtime, struggle to reveal the irregular and nested parallelism which can be expressed through task-based programming. This obscures the range of available scheduling choices. In contrast, graph-based analysis and visualisation methods can show the irregular and recursive structure which can be expressed in task-based programs. This work presents Otter, a prototype tool which uses the OpenMP Tools (OMPT) interface to trace and visualise OpenMP task creation and synchronisation constructs as a directed acyclic graph without direct instrumentation. Otter captures complex execution flow such as nested parallelism and recursive task-creation and synchronisation. A case study using a 2D finite volume Euler equations solver produced by Exahype-2 and Peano-4 demonstrates Otter's ability to visualise the target's task-based structure, giving a task-centric perspective independent of a particular runtime scheduling of the tasks and providing the first observation of the structure of this target application. Key limitations are that events occurring outside the OpenMP runtime such as inter-node communication phases are unavailable to an OMPT tool, the lack of selective tracing and no current mapping of OpenMP constructs to source representation. Otter helps to reduce the gap between the additional concurrency of task-based codes and the ability of established tools to represent the structure of such programs. Progress has been made towards answering further questions related to task-based parallelism, such as how to model the effect of different scheduling policies, how to predict optimal execution times and how to quantify the parallel efficiency of a particular task scheduling arrangement. Future work targets the present limitations, the extension of Otter to other tasking runtimes and the exploration of the effect of different scheduling policies on observed parallelism in task-based codes. This work is supported by ExCALIBUR's cross-cutting tasking theme (grant ESA 10 CDEL).

Otter: An OMPT Tool for Tracing and Visualising OpenMP Tasks

Adam Tuft *(MSc Scientific Computing and Data Analysis programme, Department of Computer Science, Durham University)* · **adam.s.tuft@durham.ac.uk** · **github.com/adamtuft/otter**

Introduction

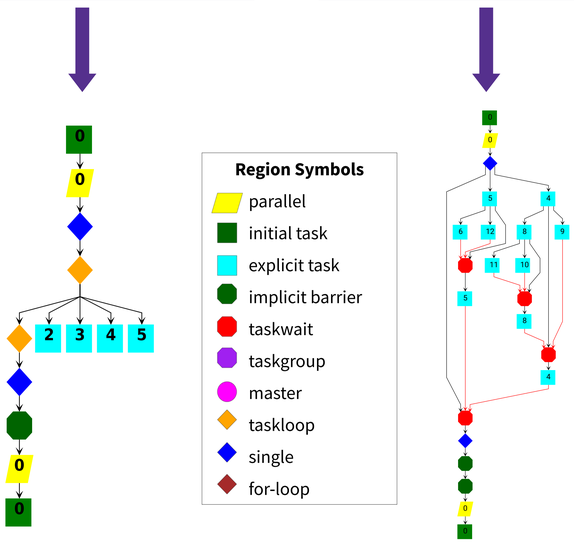
- Understanding performance of task-based code difficult due to additional concurrency of tasks & myriad scheduling possibilities.
- Thread-centric analysis tools obscure underlying task-graph structure by showing particular scheduling of tasks onto threads.
- Opportunity: portable performance analysis tool for measuring & visualising task graph structure of task-based code.
- Case study reveals the task-based structure of a PDE solver produced with ExaHyPE & illustrates performance bottlenecks identified in LLVM's OpenMP implementation.

Solution Overview

- **Otter:** event-driven, callback-based tool for tracing and visualising structure of *parallel-for*- & *task*-based OpenMP programs.
- OpenMP Tools (OMPT) interface is a non-invasive, portable alternative to direct instrumentation.
- Otter traces runtime event data from OMPT interface in OTF2 format.
- Trace data transformed into a directed acyclic graph (DAG) visualising *parallel-for*- and *task*-based structure.

```
#pragma omp parallel
{
  #pragma omp single nowait
  #pragma omp taskloop nogroup
  for (int j=0; j<4; j++)
  {
  // do work
  }
}
```

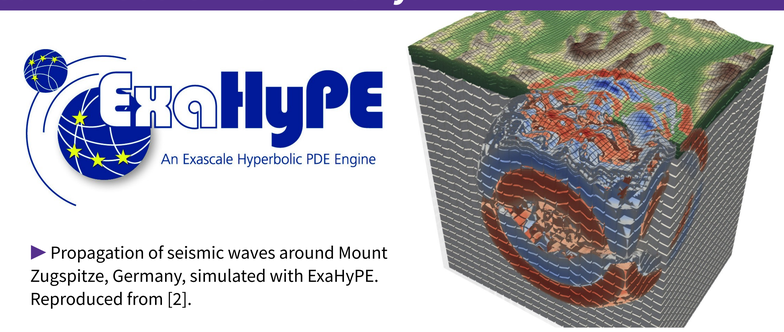
```
int fib(int n) {
  int i, j;
  if (n<2) return n;
  #pragma omp task
    i = fib(n-1);
  #pragma omp task
    j = fib(n-2);
  #pragma omp taskwait
  return i+j;
}
```



▲ OpenMP constructs are represented as nodes. Edges represent execution flow and task creation & synchronisation.

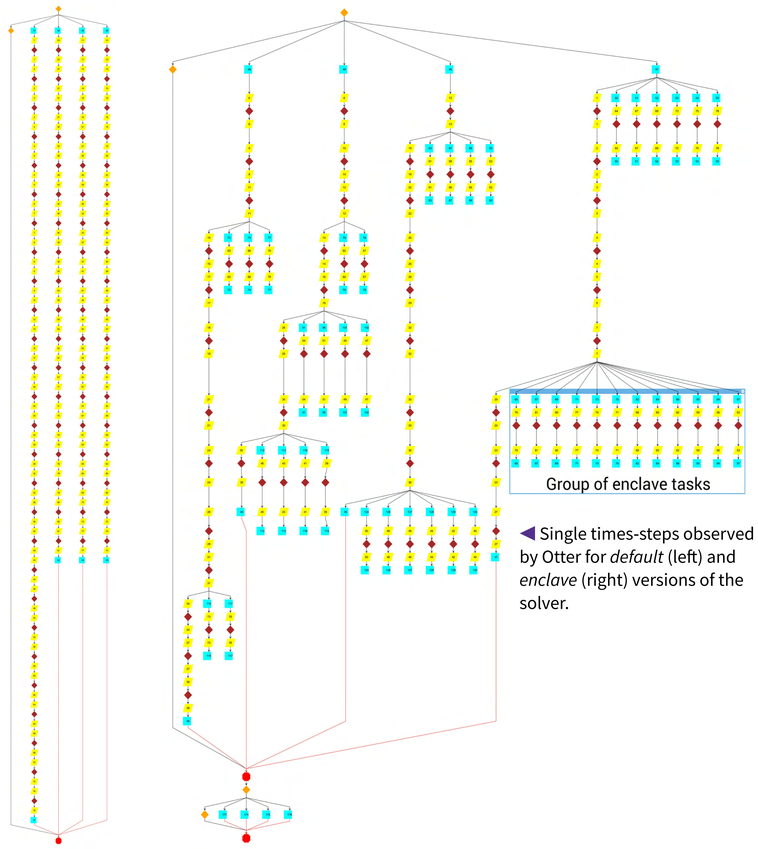
Case Study: Visualising the Task Graph of a Task-Based PDE Solver from ExaHyPE

- **ExaHyPE**: engine for solving first-order hyperbolic PDEs.
- Uses adaptive spatial grids of **Peano-4** to serialise domain cells along space-filling curve (SFC). May spawn OpenMP task on each cell.
- Case study target: Solver from [1], using **default** and **enclave** task generation modes.
- In a single time-step threads traverse partitions of the SFC to update the cells of the grid.
- **Default:** per-thread grid traversals mapped onto a set of synchronised OpenMP tasks.
- **Enclave**: non-critical cell updates packaged in *enclave* tasks to allow overlap with communication & reduce time-to-solution.



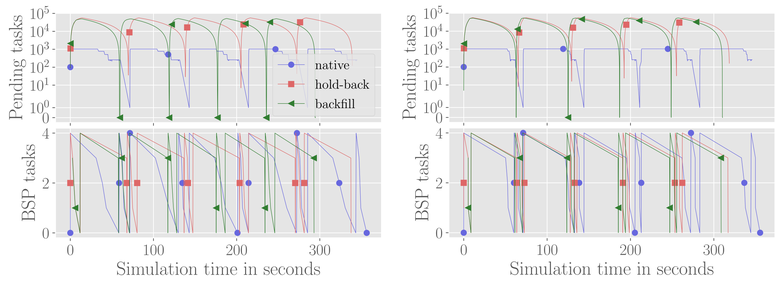
▶ Propagation of seismic waves around Mount Zugspitze, Germany, simulated with ExaHyPE. Reproduced from [2].

Results



◀ Single times-steps observed by Otter for *default* (left) and *enclave* (right) versions of the solver.

- Single time-steps observed in **default** and **enclave** tasking modes shown above.
- **Default** (left): cells updated sequentially in *parallel-for* blocks during synchronised domain traversal tasks.
- **Enclave** (right): synchronised tasks spawn unsynchronised non-critical enclave tasks which may be overlapped with later communication phase (not shown).
- Both time-steps show 81 parallel-for regions corresponding to cells of 9x9 grid, with graph structure determined by tasking mode.
- Otter illustrates inefficiencies of LLVM OpenMP implementation observed in [1] – native scheduler not able to take advantage of concurrency exposed by enclave tasks.



▲ Tasking inefficiencies observed in LLVM OpenMP implementation for ill-balanced (left) and well-balanced (right) loads. Native task limit and greedy consumption of ready tasks negates intended benefit of enclave tasks. Reproduced from [1].

Limitations & Future Work

Limitations:

- Insensitive to non-OpenMP events.
- Doesn't support `depend` clause or `distribute` & `workshare` constructs.
- Can't attribute nodes to target source.
- No means of filtering events.

Focus of future work:

- API & analysis workflow for data-driven taskification of serial code.
- Quantitative performance measurements.
- Support for other tasking runtimes e.g. Intel oneAPI toolchain.

Acknowledgements

The ExCALIBUR programme is supported by the UKRI Strategic Priorities Fund. The programme is led by the Met Office and the Engineering and Physical Sciences Research Council (EPSRC) along with the Public Sector Research Establishment, the UK Atomic Energy Authority (UKAEA) and UK Research and Innovation (UKRI) research councils, including the Natural Environment Research Council (NERC), the Medical Research Council (MRC) and the Science and Technologies Facilities Council (STFC).

This work is supported by ExCALIBUR's cross-cutting tasking theme (grant ESA 10 CDEL).

References

[1] Schulz, Holger, et al. (2021) *Task inefficiency patterns for a wave equation solver.* arXiv preprint arXiv:2105.12739

[2] Reinarz, Anne, et al. (2020) *ExaHyPE: An engine for parallel dynamically adaptive simulations of wave problems*. Computer Physics Communications 254

CIUK 2021 Poster Competition Winner

An Auto-Meshing Pipeline for Biosimulation at the Exascale

Megan Ratcli e *, James Gebbie-Rayet and Charles Moulinec

STFC Daresbury Laboratory, Sci-Tech Daresbury, Keckwick Lane, Daresbury, Warrington WA4 4AD

Biosimulations are becoming increasingly important in understanding the functions of biological systems and from this developing new therapeutics and healthcare technologies. Molecular dynamics are an example of such simulations. They have been applied to understand protein-protein interactions and drug and target interactions. Over recent decades, this field of biosimulation has developed a close relationship with experimental data from x-ray crystallography as this methodology lends itself well to such data. However, the rapid emergence of cryo-electron microscopy (cryo-EM) has presented new opportunities to study biological systems with simulations. Currently submissions to the Electron Microscopy Data Bank (EMDB)[1] is on the rise, and rapid experimental development in this field has led to the imaging of structures many orders of magnitude higher than can be observed using crystallography. Thus posing challenges for particle-based simulations such as molecular dynamics, whose simulations, at this scale, are highly computationally expensive. A multidisciplinary team of scientists at STFC and Leeds University academics are collaborating on a project aiming to utilise new continuum-based physics to simulate extremely large biological structures suited for the emerging generation of exascale supercomputers. The Fluctuating Finite Element Analysis (FFEA)[2] approach utilises meshes over atomic coordinates used in particle-based methods, therefore lending itself to the data produced by cryo-EM.

Alongside other projects in the collaboration, which are re-implementing the physics of FFEA into Code Saturne[3] to create a new biosimulation tool set (Bio Saturne), this project aims to solve usability issues in mesh generation from experimental data sources. Such errors include holes in the mesh, overlapping faces on the surface and widely heterogeneous tetrahedra mainly arising from noisy experimental data. Here we present progress on our automated pipeline which links existing software tools across the project's subject domains. The pipeline functionality will enable users to access mesh quality scoring, automated adjustments for common errors and intuitive messages for finer modifications.The development of this pipeline will increase accessibility of Bio Saturne which enables the biosimulation field to access the wealth of experimental data from new microscopy sources such as cryo-EM and using new facilities such as those at Diamond's Electro-Bioimaging centre (e-BIC)[4].

[1] https://www.ebi.ac.uk/emdb/
[2] https://ffea.readthedocs.io/en/stable/index.html
[3] https://www.code-saturne.org/cms/web/
[4] https://www.diamond.ac.uk/Instruments/Biological-Cryo-Imaging/eBIC.html
* megan.ratcliffe@stfc.ac.uk

An Auto-Meshing Pipeline for Biosimulation at the Exascale

Megan Ratcliffe, James Gebbie-Rayet, Charles Moulinec
STFC Daresbury Laboratory, Sci-Tech Daresbury, Keckwick Lane, Daresbury, Warrington, WA4 4AD, UK

code_saturne

Science and Technology Facilities Council

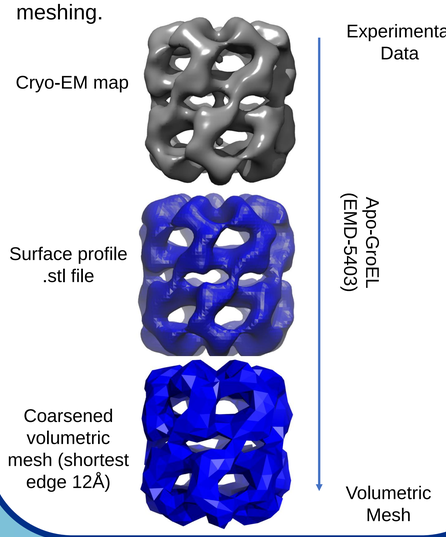
UNIVERSITY OF LEEDS

Current Biosimulation Methodology

Biosimulations are pivotal in understanding protein-protein interactions and developing new healthcare technologies. Molecular dynamics are a particle-based example of such simulations and have developed a close relationship with X-Ray crystallography. However, the rapid emergence of cryo-electron microscopy (cryo-EM) has led to imaging of structures many orders of magnitude higher than can be viewed using X-ray crystallography. Thus posing challenges for particle-based simulations which are highly computationally expensive at this scale.

FFEA Approach

The Fluctuating Finite Element Analysis (FFEA) approach uses meshes over atomic co-ordinates, therefore lending itself to data produced by cryo-EM. The original software was developed by a team at Leeds University and utilises continuum physics to model biomolecules as a finite element tetrahedral mesh and thus the volumetric data is ideal for meshing.

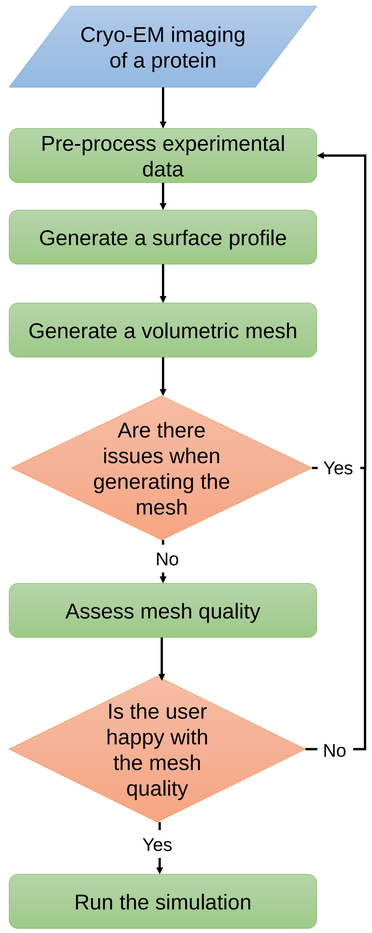


The Project

The automated pipeline is a part of a wider collaboration project between multi-disciplinary scientists at STFC and Leeds University researchers to re-implement the physics of FFEA into Code_Saturne to create a new biosimulation tool called Bio_Saturne. The pipeline aims to solve usability issues with the initial software surrounding mesh generation from noisy experimental data. These issues include:

- Holes in the mesh
- Widely heterogenous tetrahedra
- Overlapping surface faces

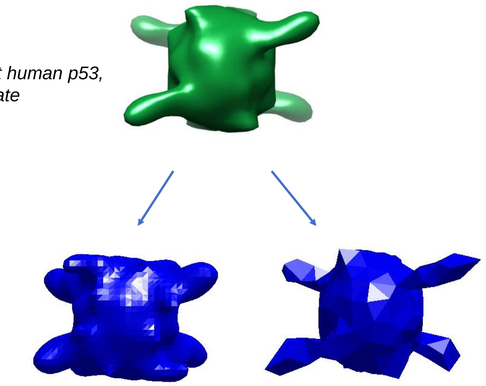
Software Pipeline



Automation

The software pipeline will be automated to enhance usability of the Bio_Saturne toolkit. The program will interact with the command line interface to run the required tools to build a mesh and extract any error messages which may arise.

EMD-11858
Recombinant human p53, tetrameric state

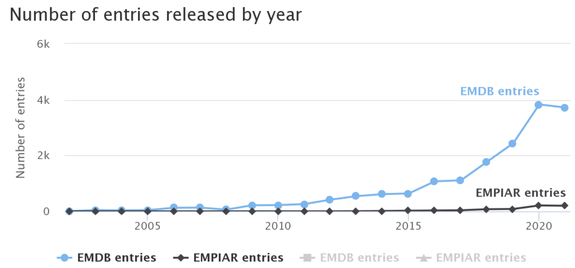


Coarsening Level: 0.5 1.15

The functionality of the pipeline will allow users to access mesh quality scoring and provide both automated and user-defined adjustments. Intuitive messages will be displayed to the user from processed command line messages to allow them to apply finer modifications.

Future Application

The automated pipeline will make Bio_Saturne more accessible to the wider biomolecular simulation and experimental structural biology communities, this will lower the barrier to access the wealth of experimental data emerging from new microscopy sources.



Data from: https://www.ebi.ac.uk/emdb/

This includes both current and next generation cryo-EM microscopes, such as those at Diamond's Electro-Bioimaging Centre (e-BIC).

Running at the Exascale

Combined with the pre-exascale readiness of Code_Saturne this framework is capable of tackling extremely large biological problems using next generation exascale computing in a way currently unfeasible with other methodologies.

Example of CFD for 57B Tetra Cell Mesh (HAWK – AMD ROME EPYC)

MPI Tasks	Time in Solver	Efficiency
131,072	68.959 s	100%
262,144	34.769 s	99%
524,288	**18.677 s**	**92%**

Biosimulations at the end of the pipeline could be ran on an exascale machine in several ways

- Complex protein simulations, requiring 100 billion elements.
- Several complex proteins interacting with each other.

In which cases the pipeline would be used once and several times respectively.

References and Acknowledgements

FFEA

Solernou A., Hanson B. S., Richardson R. A., Welch R., Harris S. A., Read D. J., Harlen O. G. "Fluctuating Finite Element Analysis (FFEA): A continuum mechanics software tool for mesoscale simulation of biomolecules" (2018), PLoS Comput. Biol. 14(3): e1005897.

Project Collaborators

- Leeds University
- Science Technology and Facilities Council
- Code Saturne Developers Team

Jemma Bennett – Durham University

Error Suppression in Continuous-time Quantum Computing

We build on [Young et al., PRA 88,062314, 2013] where logical qubits in multiple copies constitute an Ising spin system. Copies are linked together to increase the logical system's robustness to error. We introduce refinements that improve the scheme significantly. First we note that only one copy needs to be correct by the end of computation, since solution quality can be checked efficiently. Second, we find that ferromagnetic links do not help in the "one correct copy" situation, but anti-ferromagnetic links do help sometimes. Third, we developed a protocol based on local field and coupling strengths in the problem Hamiltonian, to decide whether logical qubits should be connected anti-ferromagnetically, or left disconnected. Numerical simulations show that three qubit copies connected in a loop (triangle) perform better than two or more copies connected in a chain. In logical systems which contain frustration, the anti-ferromagnetic links inhibit the propagation of errors.

Error suppression in continuous-time quantum computing

Jemma Bennett*, Tom O'Leary, Mia West, Adam Callison†, Nicholas Chancellor, and Viv Kendon‡

*jemma.e.bennett@durham.ac.uk

Durham University, UK, †University College London, UK, ‡Strathclyde University, UK

1. Summary

• Quantum Computers may be faster and/or more efficient at solving **optimisation problems** than classical computers. E.g. Max2Sat, MIS.

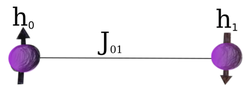
• Continuous-time quantum computing; **Quantum walks, Adiabatic Quantum Computing (AQC) and Quantum Annealing**, is an ideal setting for optimisation problems.

• However, there is limited **error correction** capabilities of continuous-time quantum computing.

• We connect **3 copies** of an **Ising spin glass** in a loop with **anti-ferromagnetic** couplings to improve **robustness** to error.

2. Transverse Ising model

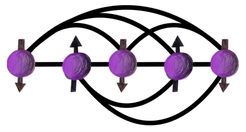
• Classical optimization problems can be encoded onto **Ising models.** Answer is the **ground state.**



• We use Sherrington-Kirkpatrick **Ising spin glasses:**

• Uniformly hard, NP-hard.

• Fields, h_i, couplings, J_{ij}. Drawn randomly from uniform range [-1, 1].



$$\hat{H}_P = \sum_{i=0}^{n-1} h_i \hat{Z}_i + \sum_{i \neq j=0}^{n-1} J_{ij} \hat{Z}_i \hat{Z}_j$$

3. Continuous-time Quantum computing

• Solve using **continuous-time** technique.

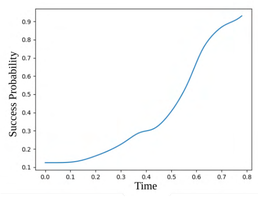
• Drive into **ground state** of $\hat{H}_P$:

$$\hat{H} = A(t)\hat{H}_I + B(t)\hat{H}_P$$

Using:

• Adiabatic Quantum Computing (AQC)

$$\hat{H}_I = n\hat{\mathbb{1}} - \sum_{j=0}^{n-1} \hat{X}_j$$

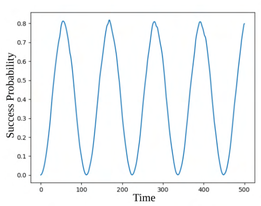


• Initial state:

$$|\psi(0)\rangle = \frac{1}{\sqrt{N}} \sum_{j=0}^{n-1} |j\rangle$$

• Quantum Walk

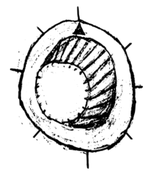
$$\hat{H}_I = \gamma(n\hat{\mathbb{1}} - \sum_{j=0}^{n-1} \hat{X}_j)$$

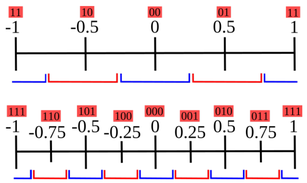


4. Error model

• We model limitations in the **resolution** of h's and J's.

• Define number of allowed values, in interval $[-1, 1]$:
$n_v = 2^p + 1$, where p is **precision**.

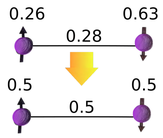




← allowed values for $p = 2$ and $p = 3$.

• Uncertainties in red and blue.

• When $p = 2$, an Ising model with true precision becomes →

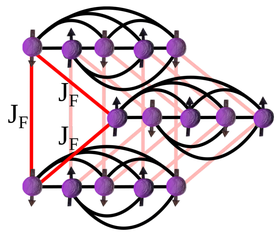


• If this change is **too large** a qubit could flip → **incorrect** ground state!

5. Error suppression method

• We aim to find the **correct** ground state with **lower** precision.

• We connect 3 copies **anti-ferromagnetically** in a loop, causing **frustration**.



The Hamiltonian for this is,

$$\hat{H} = \sum_{i=0}^{n-1} \sum_{k=0}^{m-1} (h_i + \epsilon^h_{i,k}) \hat{Z}_{i,k} + \sum_{i \neq j=0}^{n-1} \sum_{k=0}^{m-1} (J_{ij} + \epsilon^J_{i,j,k}) \hat{Z}_{i,k} \hat{Z}_{j,k}$$

$$- \sum_{i=0}^{n-1} \sum_{k \neq l=0}^{m-1} F Z_{i,k} Z_{i,l}$$

6. Frustration

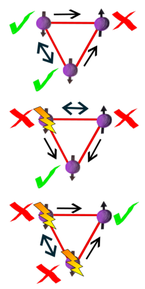
• Frustration inhibits: **error propagation**.

• **No errors:**
One copy is forced to be incorrect.

• **One error:**
One copy is forced to be correct.

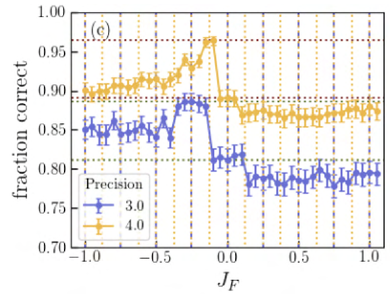
• **Two errors:**
One copy is forced to be correct.

• We only need **one or more** copies to be correct.



7. Optimal link strength

• Need to find **optimal strength** for couplings J_F linking copies.



• 1000 5 qubit Ising spin glasses

• Varied J_F between -1 and 1.

• **Fraction correct:** fraction of times we find the correct ground state.

• At $p = 3$ and $p = 4$.

• Optimal setting: J_F to **minimum** (negative) value allowed by precision.

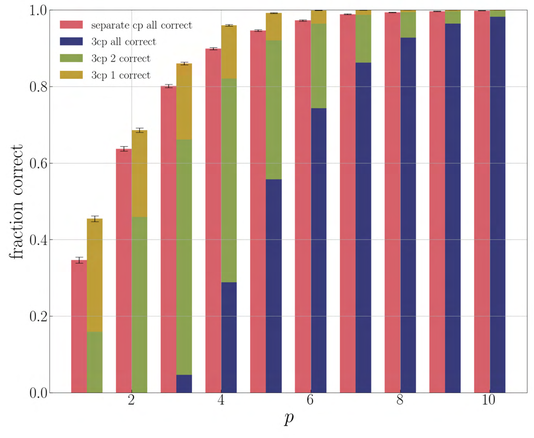
8. Results

• Subjected 10000 5 qubit spin glasses to precision p.

Measured **fraction correct** by:

• Finding **ground state** of each **single copy** (equiv. to 3 separate) using classical branch and bound technique.
→ LHS bars

• Did the same for **3 copies** connnected anti-ferromagnetically (loop). → RHS bars: 3 copies correct (blue), 2 copies correct (green), 1 copy correct (yellow).

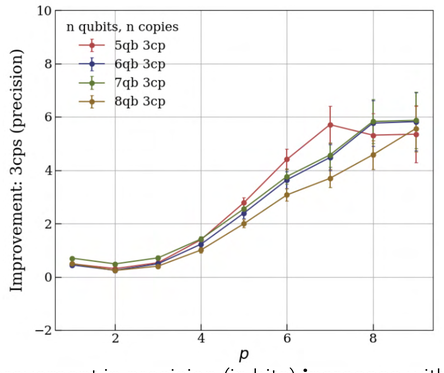


• 3 connected copies **outperforms** single copies.

• Trend continues for **larger** spin glasses. (6, 7, 8, 9 qubit spin glasses were tested).

9. Precision improvements

• Measured the difference (**improvement**) in precision between 1 and 3 copies of 5, 6, 7, 8 qubit spin glasses.



• Improvement in precision (in bits) **increases** with increasing precision of the spin glass.

• The improvement trend remains similar as the size of spin glass gets **larger**.

10. Conclusion

• Connecting **3 copies** of Ising **spin glasses** in a **loop** with **anti-ferromagnetic** links, set to **minimum** allowed precision, is **more robust** to lack of precision compared to three disconnected copies.

• Trend continues for **larger** (5, 6, 7, 8 qubit) spin glasses.

• Connecting copies technique, could be used to **increase effective precision** of computation.
→ This enables bypassing of limitations in hardware precision which arise as qubit number increases.

References

• Young, K. C., Blume-Kohout, R., Lidar, D. A. (2013). Adiabatic quantum optimization with the wrong Hamiltonian. Physical Review A - Atomic, Molecular, and Optical Physics, 88(6).

• Lidar, D. A. (2008). Towards fault tolerant adiabatic quantum computation. Physical Review Letters, 100(16), 1–4.

• Atalaya, J., Zhang, S., Niu, M. Y., Babakhani, A., Chan, H. C. H., Epstein, J., Whaley, K. B. (2020). Continuous quantum error correction for evolution under time-dependent Hamiltonians. 1–17. arXiv:2003.11248

Funding: EP/L022303/1 (VK); EP/S00114X/1 (NC); EPSRC DTG (JB)

Engineering and Physical Sciences Research Council

3 November 2021

Kieran Woodward – Nottingham Trent University

Tangible Fidgeting Interfaces for Personalised Affective Monitoring

Stress and anxiety are increasingly impacting people of all ages and socio-economic backgrounds. Modern lifestyles are heavily contributing to increased levels of daily stress and poor mental wellbeing, making the management of real-world mental wellbeing more vital than ever. Measuring and helping to improve mental wellbeing is extremely import as modern lifestyles are contributing to increased daily stress with 59% of UK adults experiencing work-related stress, costing the economy £2.4 billion each year.

Fidgeting is a common response to stress and is often used as a coping strategy as repetitive interactions help tap into an individual's psychological need to feel occupied, demonstrating potential to regulate stress. Fidget cubes embed a range of sensory tools to facilitate fidgeting catering for a wide range of needs in a small, unobtrusive design, helping to improve wellbeing by normalising stimming (self-stimulatory behaviour such as tapping or clicking). On the other hand, the ability to unobtrusively measure mental wellbeing states using non-invasive sensors has the potential to greatly improve affective monitoring. Recent developments in non-invasive sensors paired with deep learning classifiers introduce the possibility to quantify mental wellbeing in real-time. Non-invasive physiological sensors including Electrodermal Activity (EDA), Heart Rate (HR) and Heart Rate Variability (HRV) provide insight into wellbeing due to their correlation with the sympathetic nervous system, paving the way to continuously and non-invasively monitor wellbeing.

The ability to objectively infer affective states from multiple modalities such as physiological and motion sensor data is an exciting proposition, as it could enable better real-world management of wellbeing. However individual differences between people limit the generalisability of deep learning models. Wellbeing states are often personal with individuals experiencing large variations in physiological parameters, despite experiencing the same state of mental wellbeing. Advances in deep learning have helped increase affective modelling performance, however the majority of previous work has not considered the personalisation of models.

This research introduces the concept of Tangible Fidgeting Interfaces (TFIs) as custom-built physical fidgeting devices that enable repetitive physical interaction while also enabling objective sensor measurement. These fidgeting interfaces can be coupled with deep learning algorithms, paving the way for a new type of real-time interaction. TFIs can vary in form but by developing handheld interfaces that embed the necessary sensors, it is possible to develop devices that encourage engagement and improve wellbeing through fidgeting, acting as a preventative tool. In particular, this work presents the design of TFIs in the form of computerised fidget cubes that embed non-invasive sensors along with fidgeting mechanisms. Digitising traditional fidget cubes enables the real-time collection of physiological data and fidgeting interactions, potentially aiding relaxation and easing restlessness.

TFIs were provided to 15 participants to use in-situ for one week to collect labelled physiological and motion data and enable fidgeting. Participants were encouraged to use the device as frequently as possible and label their wellbeing as positive or negative each time they used the interface. It is essential to utilise the real-world labelled data collected from the TFIs to develop personalised, subject-independent models for wellbeing detection as when working with a heterogeneous population there are numerous factors that result in a variation of physiology. This work proposes a Transfer Learning (TL) approach that develops personalised real-world affective models classifying positive and negative mental wellbeing

from the time-series sensor data by adapting a controlled stressor source 1D CNN model. In order to achieve the transfer, the source network was frozen, the fully connected layer was removed and two fully connected layers were added, forming an adaption layer. The CNN was then re-trained with each user's labelled data to fine-tune the model and update the weights, developing a personalised affective model for each participant.

TL can help the development of personalised models but real-world data collection and the process of developing a custom model for each user remain labour and time intensive tasks, preventing mass adoption. This TL approach advances on previous state-of-the-art approaches by alleviating many of these challenges traditionally associated with affective modelling by only requiring few real-world labelled samples, which can then be used to automatically develop personalised models on-device. Collecting a small, labelled dataset directly from the TFI and then personalising models on-device significantly simplifies the process of developing the personalised, cross-domain models.

The results show adopting the TL approach significantly increased model performance with the highest accuracy achieved using the TL 1D CNN fusing physiological and motion data, demonstrating the importance of using HR, HRV, EDA and motion data when inferring positive and negative states of mental wellbeing. These TL models achieved an average accuracy of 93.47%, 21.8% higher than the comparative 1D CNN trained without the TL approach. Furthermore, when the personalised models were trained with other users' data there was a 36.09% reduction in accuracy, confirming the TL approach successfully personalised the models. Also, when the real-world datasets were tested on the source controlled experiment model there was a 43.55% reduction in accuracy confirming the TL approach also enabled the model to adapt cross-domain for real-world mental wellbeing recognition.

When using solely motion data to infer wellbeing, an average accuracy of 88.05% was achieved using the TL approach, again outperforming the equivalent non-TL model. While this is lower than models trained using physiological data, it is higher than expected, with one user achieving their highest respective accuracy of 97% using the univariate motion TL model. The obtained performance demonstrates that these new forms of tangible interfaces combined with deep learning classifiers have the potential to accurately infer wellbeing in addition to providing fidgeting tools that participants found “calming” and “relaxing”.

Overall TFIs present new methods of real-world physiological data collection whilst simultaneously enabling fidgeting interactions to help improve mental wellbeing. The proposed TL methodology helped overcome problems with affective model personalisation, thus improving on the performance of conventional deep learning methods. This approach creates the opportunity for future research applications to infer real-world wellbeing for the wider population without the need to first collect large, labelled datasets, greatly improving accessibility to personalised affective models.

Tangible Fidgeting Interfaces for Personalised Real-World Affective Modelling

Kieran Woodward
kieran.woodward@ntu.ac.uk

Eiman Kanjo
eiman.kanjo@ntu.ac.uk

Department of Computer Science, Nottingham Trent University

Introduction

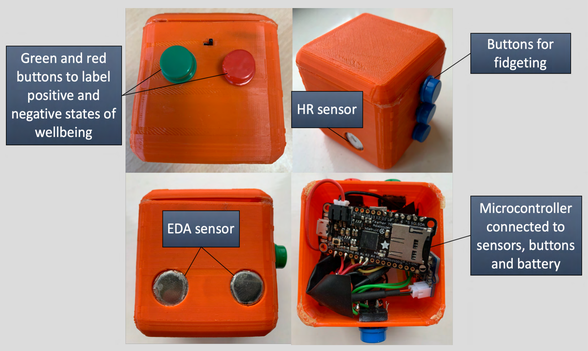
Modern lifestyles are heavily contributing to increased levels of daily stress and poor mental wellbeing making the management of real-world wellbeing more vital than ever.

Fidgeting is a common response to stress and is often used as a coping strategy as repetitive interactions help tap into an individual's psychological need to feel occupied.

On the other hand, recent developments in non-invasive physiological sensors paired with deep learning classifiers introduce the possibility to quantify mental wellbeing in real-time, paving the way to continuously and non-invasively monitor wellbeing.

Tangible Fidgeting Interfaces

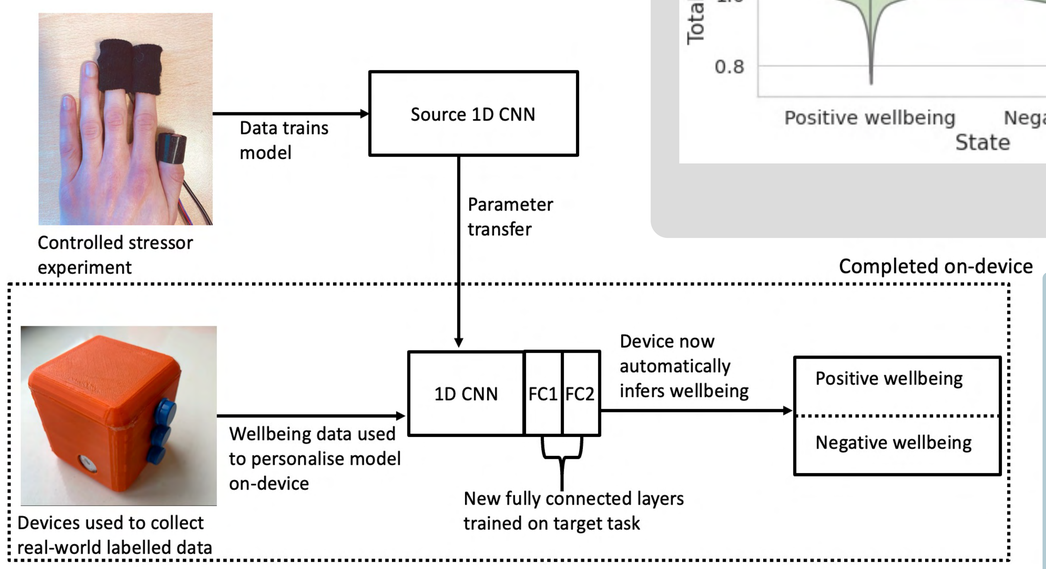
This research introduces the concept of Tangible Fidgeting Interfaces (TFIs) as custom-built physical fidgeting devices that enable repetitive physical interaction while also enabling real-world objective physiological and motion sensor measurement. These fidgeting interfaces can be coupled with deep learning algorithms, paving the way for a new type of real-time interaction.



On-device Personalisation TL Methodology

A Transfer Learning (TL) approach has been developed to personalise real-world affective models, classifying positive and negative mental wellbeing from the time-series sensor data by adapting a controlled stressor source 1D CNN model.

This TL approach advances on previous state-of-the-art approaches by alleviating many of these challenges traditionally associated with affective modelling by only requiring few real-world labelled samples, which can then be used to automatically develop personalised models on-device. Collecting a small, labelled dataset directly from the TFI and then personalising models on-device significantly simplifies the process of developing the personalised, real-world affective models.



Results

Adopting the TL approach significantly increased model performance with the highest accuracy achieved using the TL 1D CNN fusing physiological and motion data. These TL models achieved an average accuracy of 93.47%, 21.8% higher than the comparative 1D CNN trained without the TL approach.

When using solely motion data to infer wellbeing, an average accuracy of 88.05% was achieved using the TL approach, again outperforming the equivalent non-TL model. This shows changes in fidgeting behaviours can be used to infer wellbeing.

Table 1: Comparison of target user multivariate physiological models

	Non-TL Accuracy	TL Accuracy	TL AUC
User 1	74.1%	84%	86.2%
User 2	85.1%	90.4%	90.6%
User 3	72.5%	99.6%	99.6%
User 4	85.5%	95.8%	91.6%
User 5	86.2%	90.6%	90.7%
User 6	83%	93.5%	92%
User 7	85.3%	89%	89%
User 8	83.8%	91%	95%
User 9	87 %	97%	92%
Average	82.5%	92.3%	91.9%

Table 2: Comparison of target user multivariate physiological and motion models.

	Non-TL Accuracy	TL Accuracy	TL AUC
User 4	87%	96.3%	86.9%
User 5	79%	89.2%	89.2%
User 6	81.1%	90.3%	86.6%
User 7	54.8%	91%	91%
User 8	50%	96%	98%
User 9	78.3 %	98%	99%
Average	71.7%	93.5%	91.8%

Table 3: Comparison of target users' univariate motion models.

	Non-TL Accuracy	TL Accuracy	TL AUC
User 4	86%	89.1%	72.3%
User 5	80%	72.4%	72.6%
User 6	73%	82.8%	75.4%
User 7	27%	90%	90%
User 8	69%	97%	89%
User 9	74 %	97%	89%
Average	68.2%	88.1%	81.4%

Conclusion & Future Work

Overall TFIs present new methods of real-world physiological data collection whilst simultaneously enabling fidgeting interactions that participants found “calming” and “relaxing”, helping to improve mental wellbeing.

The proposed TL methodology helped overcome problems with real-world affective model personalisation, thus improving on the performance of conventional deep learning methods.

This approach creates the opportunity for future research applications to infer real-world wellbeing for the wider population without the need to first collect large, labelled datasets, greatly improving accessibility to personalised affective models. In the future, TFIs should be trialled with more users over a longer period of time to collect additional data and further explore the impact of fidgeting on wellbeing.

For further details, visit the articles or scan the QR Code

Woodward, K., & Kanjo, E. (2020). iFidgetCube: Tangible Fidgeting Interfaces (TFIs) to Monitor and Improve Mental Wellbeing. IEEE Sensors Journal.
Woodward, K., Kanjo, E., Brown, D. J., & McGinnity, T. M. (2021). Towards Personalised Mental Wellbeing Recognition On-Device using Transfer Learning “in the Wild.” IEEE International Smart Cities Conference 2021.

Thomas Johnson – Nottingham Trent University

DIGITALEXPOSOME: Unravelling The Impact of the Environment on Mental Wellbeing

The proliferation of miniaturized electronics has fuelled a shift towards environmental sensing technologies ranging from pollution to weather monitoring at a higher granularity. However, at present, little consideration has been focused on the relationship between environmental stressors (e.g., air pollution) and mental wellbeing. Air pollution levels are often exceeding normal allowed limits for cities with the UK spending between £8 to £20 billion pounds a year. The world health organisation (WHO) found that 91% of people are living in places where the air quality guidelines are not met and the use of non-clean fuels and household emissions in the atmosphere are causing over 4.2 million deaths each year.

Recent developments in urban sensing and Internet of Things (IoT) has created the possibility to utilise environmental and on-body sensing tools to monitor the environment and its impact on individuals. Sensor-based technologies are becoming increasingly popular due to their availability to collect data in real-time, affordability and small size. These advances continue to enable more opportunities for capturing environmental exposures in urban setting by providing the mechanisms to collect and analyse objective data, physiological changes, and behaviour markers of mental wellbeing, in real-time. In addition, the major advances and recent developments within data science have created greater opportunities to understand large multimodel datasets through machine learning, deep learning, and spatial visualisations. Personal sensors to measure individual exposure such as air pollution, noise, outdoor temperature, physical activity, and blood pressure have been a positive way forward in monitoring due to their ability to collect data continually and in real-time helping to reveal early health conditions. By combining these sensor data streams together and the possibility for an individual to continuously wear sensors, the data can show the exposures an individual encounters as well as predict early health conditions. To understand the relationship, we define the term 'DigitalExposome' as a conceptual framework that takes us closer towards understanding the relationship between environment, personal characteristics, behaviour, and wellbeing. The framework aims to quantify an individual's exposure to the environment by using a range of technological, mobile sensing and digital devices. Combining multiple data collection methods helps to support the DigitalExposome concept in gaining a better understanding into how exposures to environmental pollutants can impact mental wellbeing. Our concept extends the current work on the 'Exposome' concept by digitally providing a better understanding into the impact of exposure directly to an individual.

Through 'DigitalExposome', the aim to explore the many opportunities that we for-see with this concept in exploring the link between pollution and wellbeing. To support this work, we have developed a range of sensing devices and applications. DigitalExposome is primarily made up of two parts: data collection and data analysis. Both aspects make use of technological advances in order to calculate the exposome. In order to quantify the process, we propose the utilisation of data from sensors that show how an individual has been exposed to pollutants. We see this as being a key part of the exposome concept, where both terms are clearly connected through their vision of being able to capture the true exposure that an individual has been exposed to. Data that is generated through the use of technology, such as sensors is ideal to monitor various exposures and enable the possibility to link this to health. The experiment took the form of 25 participants walking around an urban environment carrying the sensing equipment, whilst simultaneously collecting (for the first time) multi-sensor data including urban environmental factors (e.g. air pollution including: PM1, PM2.5, PM10, Oxidised, Reduced, NH3 and Noise, People Count in the

vicinity), body reaction (physiological reactions including: EDA, HR, HRV, Body Temperature, BVP and movement) and individuals' perceived responses (e.g. self-reported valence). Using a range of mathematical and statistical analysis concepts (multivariate regression, principle component analysis and visualisations) on the collected fused sensor data, the results clearly highlighted the impact of certain environmental pollutants towards physiological variables and showing a negative impact to mental wellbeing. Particularly, it was possible to see first-hand the impact through the use of urban visualisations such as heat-maps and Voronoi. It could be seen that when each participant walked from an open, green space into a busy, polluted environment: physiological responses such as Heart-Rate, ElectroDermal Activity and Heart-Rate Variability were drastically impacted. In addition to this, participants would typically report (via the smartphone) when they were within a high PM2.5 environment that their wellbeing was negatively impacted.

Furthermore, machine learning was utilised in order to classify the five self-reported states of wellbeing using the pollution (PM1.0, PM2.5, PM10, Oxidised, Reduced, NH3 and Noise) and physiological (BVP, EDA, HR, HRV and body temperature) data. The ability for pollution data to increase overall accuracy demonstrates its impact on wellbeing and shows pollution should continue to be considered as a factor that influences changes in wellbeing. To conclude, although very appealing, the exposome approach is complex, both in terms of observing each factor, quantifying it and analysing its relationship to health. This novel research challenges existing approaches to understanding the ‘Exposome’ concept, providing a new perspective on how to quantify the approach between the environment and mental wellbeing. The use of statistical analysis including PCA, MultiVariant Linear Regression, Voronoi data spatial visualisations were implemented to explore the variation in data and the factor importance. Experiment results found that physiological (on-body) sensor data is directly correlated to pollution (Particulate Matter in particular) within the environment. In addition, Machine learning classification of the five states of mental wellbeing found that 80.8% of accuracy using the fused physiological and pollution data. The use of the ‘DigitalExposome’ concept shows that there is a clear relationship of environmental pollutants impacting towards a negative mental wellbeing.

DigitalExposome: Unravelling the relationship between the Environment and Mental Wellbeing

Thomas Johnson
thomas.johnson@ntu.ac.uk

Eiman Kanjo
eiman.kanjo@ntu.ac.uk

Department of Computer Science, Nottingham Trent University

Introduction

The long-term exposure to urban environmental stressors such as particulate matter, gases and noise have been found to significantly impact an individual's behaviour and physiological health. The World Health Organisation (WHO) find that 91% of people are living in places where the air quality guidelines are not met and the use of non-clean fuels and household emissions in the atmosphere are causing over 4.2 million deaths each year.
Developments in urban sensing and Internet of Things (IoT) has created the possibility to utilise environmental and on-body sensing tools to monitor the environment and impact to individuals. Sensor-based technologies are increasingly popular due to their availability to collect data in real-time, affordability and small size.
Mobile technology in previous research coupled with sensors have aimed to provide a deeper understanding into the impact of exposure to an individual in a particular location.
Developed in 2005, the Exposome Concept (Figure 1) encompasses each exposure that is subjected to a human from birth to death.

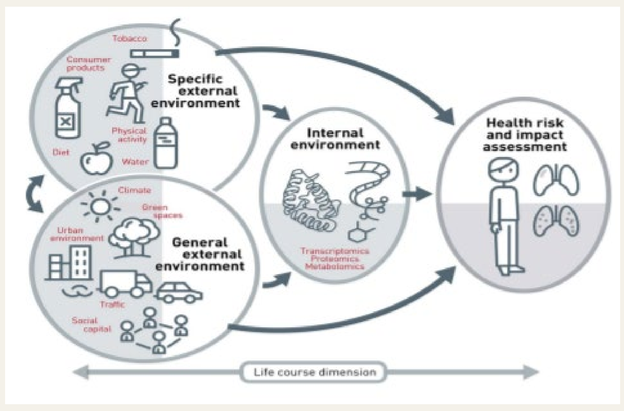


Figure 1. Demonstrating three stages of Exposome Concept

The concept in its current form can calculate some of the impact of environmental exposure, however, there remains some challenges. Most studies have found it challenging to address and understand Exposome fully because of its size, quantity of data required, and the overall quality of the data produced.

DigitalExposome

DigitalExposome is a framework (Figure 2) to quantify an individual's exposure to the environment by utilising a range of technological, mobile sensing and digital devices. The concept aims to measure multiple environmental factors using mobile technologies and then quantify them in real-life settings. The combination of multiple data collection methods helps to support DigitalExposome and gain a better understanding into how exposure to the environment can impact mental wellbeing.

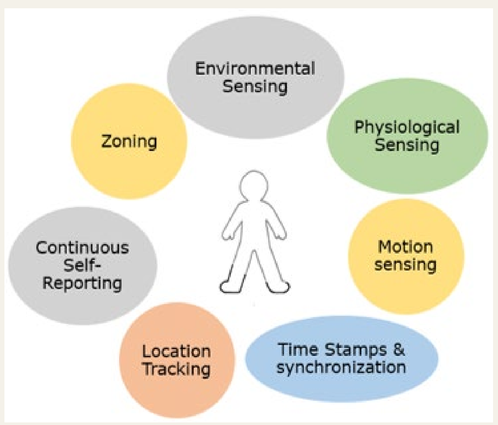


Figure 2. Data collection methods in unravelling DigitalExposome

DigitalExposome promotes the use of the Exposome concept by digitally providing a better understanding into the impact of exposure directly to an individual. There are many opportunities with this concept in exploring the link between pollution and wellbeing. The concept is primarily made up of two parts: data collection and data analysis. We see this as being a key part of the Exposome concept, where both terms are clearly connected through their vision of being able to capture the true exposure that an individual has been exposed to. Data that is generated through the use of technology, such as sensors is ideal to monitor various exposures and enable the possibility to link this to health.

Conceptual System Architecture

Conceptual layer explains the four main areas that can impact mental wellbeing include environmental, biological, social and cultural factors. (See Figure 3)

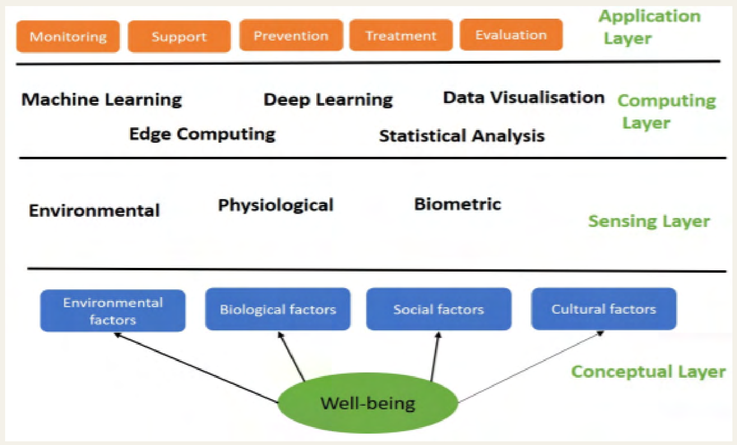


Figure 3. Conceptual and System Architecture of DigitalExposome

Methodology

Participants walked through a range of different urban environments (see Figure 4) from several green to busy and polluted spaces which would help to demonstrate the impact of different levels of exposure to air pollutants.

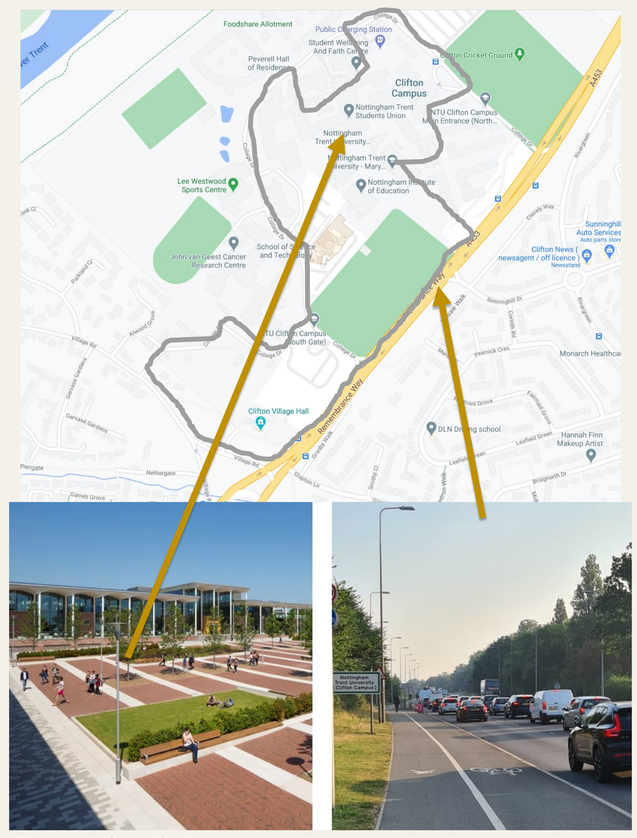


Figure 4. Specified route map demonstrating the two environment (left) Green spaces and (right) Busy, polluted

Data Collection

Participants were given a custom-built smartphone app to record momentary wellbeing, E4 Empatica to observe physiological changes and Enviro-IoT to capture environmental pollution (Figure 5).

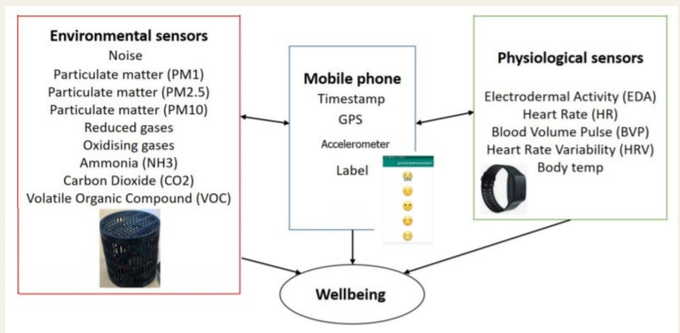


Figure 5. (left) Enviro-IoT, (middle) Wellbeing Application, (right) E4 Empatica along with collected variables

Results & Discussion

Figure 6, demonstrates the impact of wellbeing against levels of PM2.5. The bars on the chart are associated with how many times a particular user would label how they were feeling (reported wellbeing) whilst walking around the environment. The results of this indicate that high levels of PM2.5 are associated with a negative wellbeing, shown by participants choosing '1' on the device. Whereas where participants labelled '5' (very positive wellbeing), the levels of PM2.5 were much lower.

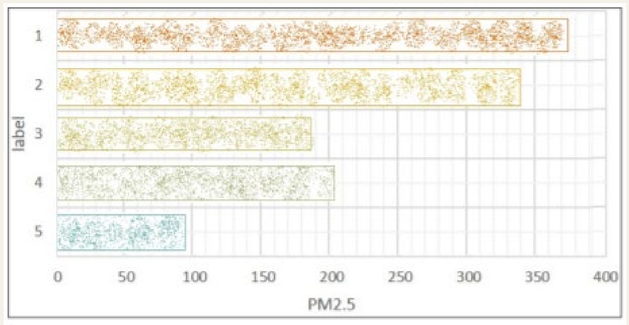


Figure 6. Depicts the relationship between the self-reported participant's wellbeing (label) and PM2.5

Spatial Visualisation – Heat Maps

Heat maps are a common method in investigating patterns with collected sensor data. Figure 7, presents several environmental and physiological sensor data, depicting the changes in both while the participant is travelling around the route. At each right corner if the map we can observe that moving from a green space into a busy polluted space found an increase of PM2.5 and Noise which also resulted in an increase of Heart-Rate Variability (HRV) and ElectroDermal Activity (EDA).

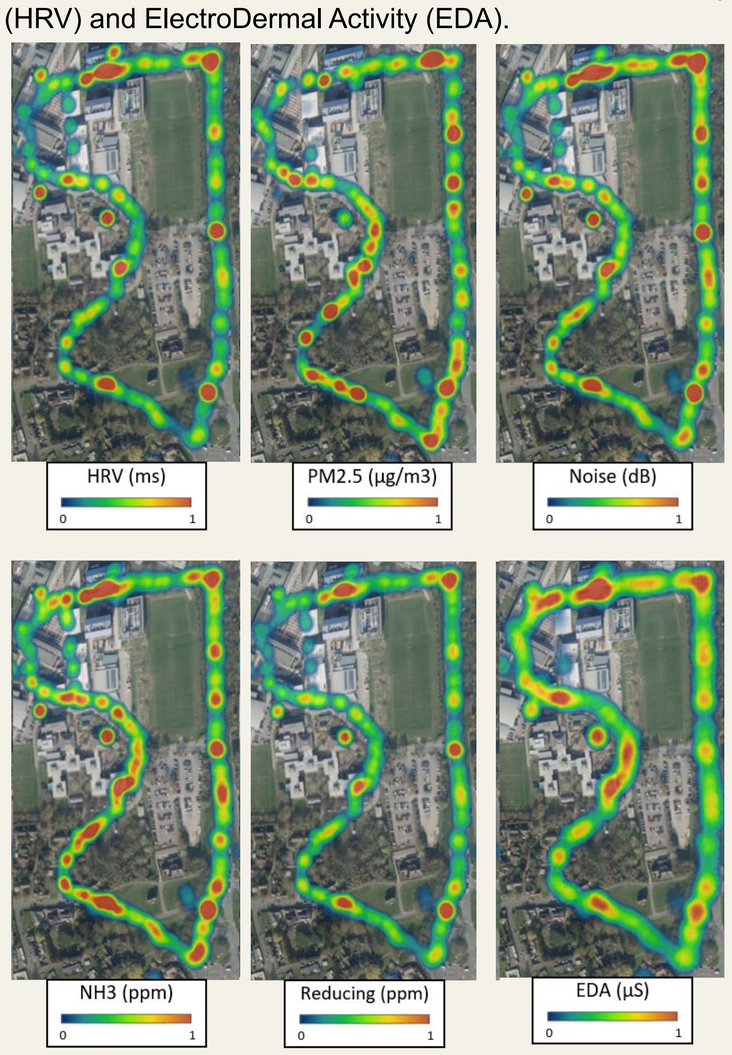


Figure 7. Several heatmaps demonstrating environmental and physiological sensors along the specified route

Voronoi Visualisation

Figure 8, presents the self-reported momentary wellbeing data using the app on the specified route for this experiment. The colour of the polygons represents the wellbeing data from low negative to high positive. The visualisation demonstrates that poor wellbeing (lighter colour) was most reported along the main road where high levels of pollution were also experienced whereas more positive states of wellbeing was recorded in less polluted areas such as fields and open spaces.

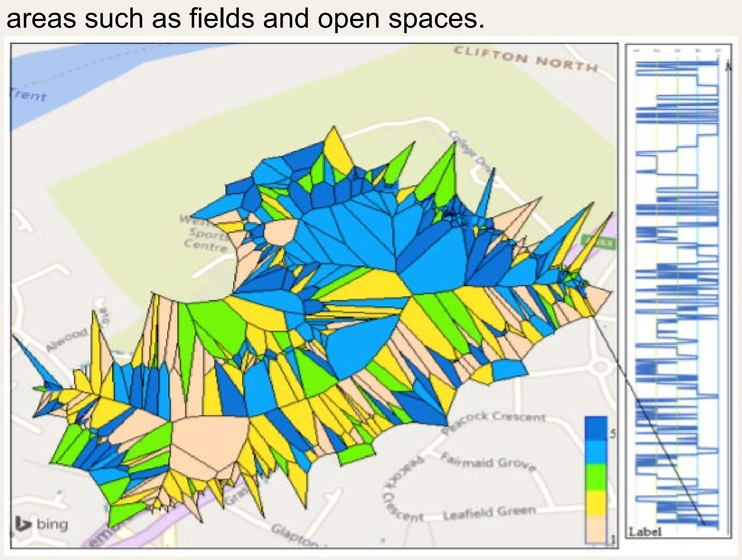


Figure 8. (left) Voronoi overlay from one participant data. Each polygon represents one location trace tagged with a wellbeing label while collecting the data in specified route (the map layer from Microsoft Bing), (right collected label data from start to end)

Conclusion & Future Work

DigitalExposome that demonstrated the potential of employing a multi-model mobile sensing approach to unravel the relationship between the environment and it's impact on mental wellbeing. We found that physiological (on-body) sensor data is directly impacted to high levels of pollution (PM in particular) within the environment. In the future environmental sensors to observe changes that may improve our sense of places and characterize the relationship between people and spatial settings, which in turns might influence the future design of urban spaces.

Find out more about this work:
Johnson, T., Kanjo, E., & Woodward, K. (2021). *DigitalExposome: Quantifying the Urban Environment Influence on Wellbeing based on Real-Time Multi-Sensor Fusion and Deep Belief Network.*

|epcc|

Comparison of I/O performance of ADIOS 2 using benchmark_c

Shrey Bhardwaj

October 23, 2021

Supervisors: Prof. Mark Parsons
Dr. Paul Bartholomew

ASiMoV Prosperity Partnership
Edinburgh Parallel Computing Centre (EPCC)
School of Informatics. University of Edinburgh
United Kingdom

Abstract

Current Petascale High Performance Computing (HPC) systems perform in the order of tens to hundreds of petaFLOPs [1] which represents 10^{15} computations per second. However future challenges require even faster computational speeds, which necessitates Exascale computing [1]. One of the major roadblocks to achieving these speeds is the I/O bottleneck. Current systems are capable of processing data quickly, but they are limited by how fast they can read and write data. For instance, the bandwidth achieved using a shared file for all processes on the NextGenIO cluster at EPCC is around 3 GB/s for writing and 7 GB/s for reading [2]. This limitation hampers startup times for scientific applications, and forces a trade-off between data resolution and allocated compute time. If this bottleneck is eliminated it would significantly increase the performance and efficiency of HPC systems.

The ADIOS 2 library has been developed to improve the I/O performance for large scientific datasets. This performance is achieved using an abstraction layer that handles data movement, and is independent of the back-end I/O layer [3]. ADIOS 2 allows individual write operations to be buffered, and data can be written to file using different write methods. To compare the I/O rates achieved using ADIOS 2 against other traditional I/O layers, a prototype C-based benchmarking application `benchmark_c` [4] was developed. This application extends `benchio`[5], a Fortran based benchmarking application by including an ADIOS 2 benchmarking test. Using this application the I/O performance from serial writes, MPI, HDF5 and ADIOS 2 systems were compared when writing to disk. Different abstract engines from ADIOS 2 were used such as the ADIOS 2-native BP4 engine and the HDF5 engine which uses the HDF5 I/O layer. Their performance was then compared with respect to average time taken and average bandwidth achieved for increasing size of datasets and varying number of nodes, tasks per processor and Lustre stripe counts.

The tests were carried out in two HPC systems at EPCC: the NextGenIO and Fulhame clusters to compare the results of software optimisations over more traditional I/O layers on different hardware. The NextGenIO HPC system uses Intel's new Optane DC Persistent Memory Module (DCPMM) [1]. This architecture combines high-performance processors with Storage Class Memory (SCM) in Non-Volatile Memory Random Access Memory (NVRAM) form, traditional Dynamic Random Access Memory (DRAM) memory and a high-performance network [1] resulting in unprecedented levels of capacity at near-DRAM speeds [6]. In its simplest mode, "Memory mode", it expands

the available memory capacity of a system which can deliver excellent performance for large workloads without having to change the application [6]. For the purpose of comparison with a representative standard memory system, the ARM Fulhame Cluster is used. It is comprised of 64 compute nodes, each with two 32-core processors and 128 GB of memory resulting in 256 GB memory per node and 4096 total number of cores [7].

Preliminary findings indicate that using ADIOS 2 with the HDF5 engine achieves a higher I/O rate than by using the HDF5 I/O layer directly. In the future this work will be used to set a baseline performance for ADIOS 2, investigate the performance benefits of ADIOS 2's runtime configurations, develop an integration with the FEniCSx finite element framework and utilise the NVRAM storage of NextGenIO.

Acknowledgements

The Fulhame HPE Apollo 70 system is supplied to EPCC, the supercomputing centre at the University of Edinburgh, as part of the Catalyst UK programme, a collaboration with Hewlett Packard Enterprise, Arm and SUSE to accelerate the adoption of Arm based supercomputer applications in the UK.

Bibliography

[1] A. Jackson, I. Panourgias, B. Homölle, A. Miranda, R. Nou, J. Conejero, M. Cintra, S. Smart, A. Bonanni, H. Brunst, C. Herold, S. Zafar, and T. Dresden, "NEXTGenIO Architecture White Paper," p. 16.

[2] A. Jackson, M. Weiland, M. Parsons, and B. Homölle, "An Architecture for High Performance Computing and Data Systems Using Byte-Addressable Persistent Memory," in *High Performance Computing*, M. Weiland, G. Juckeland, S. Alam, and H. Jagode, Eds. Cham: Springer International Publishing, 2019, pp. 258–274, series Title: Lecture Notes in Computer Science. [Online]. Available: http://link.springer.com/10.1007/978-3-030-34356-9_21

[3] S. Klasky, M. Wolf, M. Ainsworth, C. Atkins, J. Choi, G. Eisenhauer, B. Geveci, W. Godoy, M. Kim, J. Kress, T. Kurc, Q. Liu, J. Logan, A. B. Maccabe, K. Mehta, G. Ostrouchov, M. Parashar, N. Podhorszki, D. Pugmire, E. Suchyta, L. Wan, and R. Wang, "A View from ORNL: Scientific Data Research Opportunities in the Big Data Age," in *2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, Jul. 2018, pp. 1357–1368. [Online]. Available: https://ieeexplore.ieee.org/document/8416399/

[4] S. Bhardwaj, "benchmark_c." [Online]. Available: https://github.com/sb15895/benchmark_c.git

[5] EPCCed, "benchio." [Online]. Available: https://github.com/EPCCed/benchio.git

[6] M. Weiland, H. Brunst, T. Quintino, N. Johnson, O. Iffrig, S. Smart, C. Herold, A. Bonanni, A. Jackson, and M. Parsons, "An early evaluation of Intel's optane DC persistent memory module and its impact on high-performance scientific applications," Denver Colorado, Nov. 2019, pp. 1–19. [Online]. Available: https://dl.acm.org/doi/10.1145/3295500.3356159

[7] EPCC, "fulhame." [Online]. Available: https://www.epcc.ed.ac.uk/facilities/other-facilities/fulhame

benchmark_c: A tool to compare I/O performance from MPI-IO, HDF5 and ADIOS2

Shrey Bhardwaj [1]
Supervisors: Dr. Paul Bartholomew [1] Prof. Mark Parsons [1]
[1]EPCC, University of Edinburgh

Introduction

Current Petascale HPC systems perform 10^{15} computations per second. However future challenges require even faster computational speeds. One of the major roadblocks to achieving these speeds is the I/O bottleneck.

This I/O bottleneck can be seen in figure 1 obtained from the CFD code, Xcompact3D [5].

This poster introduces benchmark_c [2] which has been developed to benchmark I/O backends to find improvements in the I/O bandwidth.

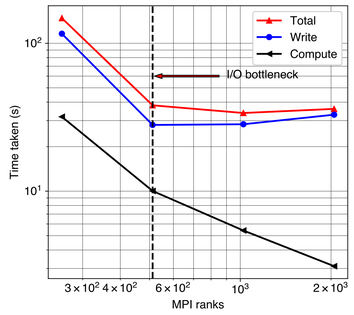
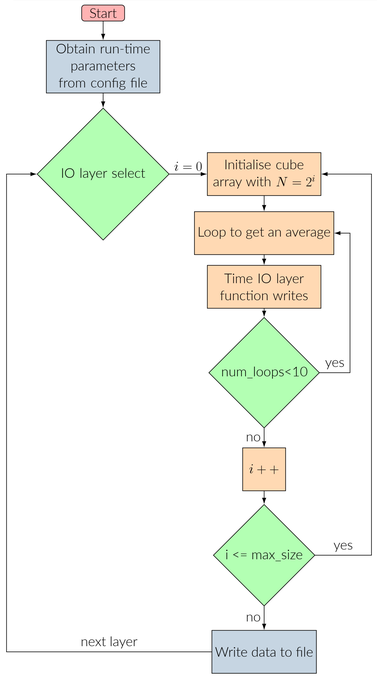


Figure 1:Parallel performance of Xcompact3D for increasing MPI ranks

benchmark_c



- benchmark_c [2] writes increasing array sizes using different I/O backends.
- It is derived from benchio [1], a fortran based application.
- A data array is passed to either MPI, HDF5, ADIOS2 HDF5 IO engine or ADIOS2 BP4 IO engine for writing to disk.
- The results were obtained by submitting this job 3 times and averaged to account for any system-wide noise.

Machines used	NextGenIO [4]	ARM Fulhame Cluster [3]
Total number of nodes	34	64
Cores per node	48	64
Memory per node (GB)	196	256
Compute environment	gnu/10.2.0 intel-mpi /2021.3.0 HDF5/1.12.0 ADIOS2/2.7.1	gnu/9.2.0 openmpi /4.0.2 HDF5/1.12.0 ADIOS2/2.7.1

Table 1:Details of HPC machines and modules used

Comparison of different backends

First different I/O backends were investigated for any improvement in I/O performance. This was conducted over multiple node configurations using NextGenIO with full Lustre striping across the run directory. In this experiment, the MPI ranks were used as follows; 1 node in serial (1 Processor), 1 node half full (24 Processors), 1 full node, 2 full nodes, 4 full nodes and 8 full nodes.

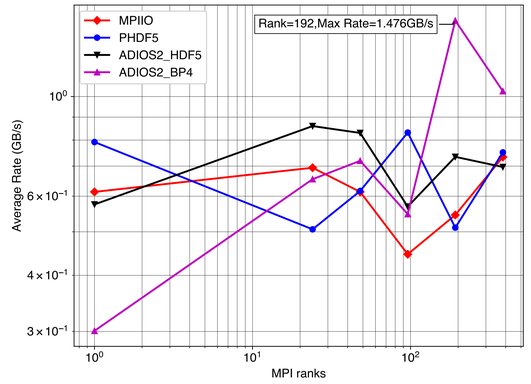


Figure 2:Comparison of achieved bandwidth using backends with local array size of 0.13GB

Comparison with different Machines

Next, this experiment was repeated on different machines, NextGenIO HPC system and ARM Fulhame Cluster. In addition to this, the jobs were run with maximum and minimum striping in both the machines with a local array size of 0.13GB with the same node configurations as before.

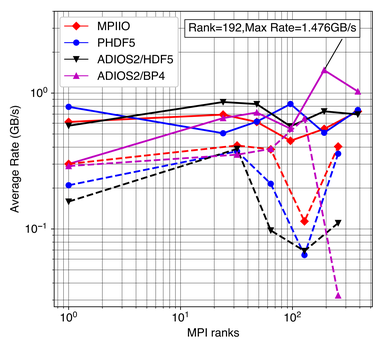


Figure 3:Benchmarking results on NextGenIO and Fulhame (marked by dashed line)

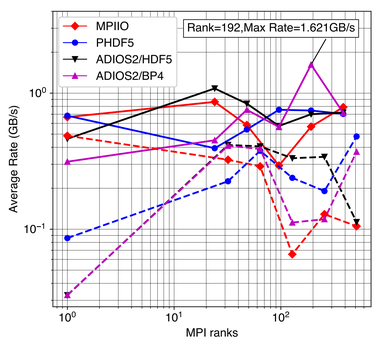


Figure 4:Benchmarking results on NextGenIO and Fulhame (marked by dashed line)

Speedup comparison of ADIOS2 I/O engines

Lastly, the ADIOS2 I/O engines were compared relative to HDF5 for performance advantages using their default settings. The experiments were run on NextGenIO with different MPI ranks and full Lustre striping across the run directory. In figure 5, two job sizes are compared, 1 MPI rank (serial) and 384 MPI ranks (8 Nodes) to compare the benefits of the two I/O engines with varying levels of parallelism upto a local array size of 0.13GB.

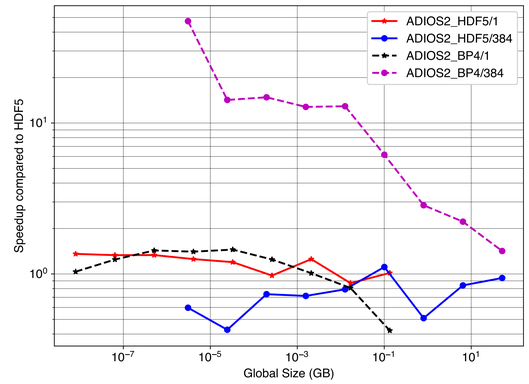


Figure 5:Speedup achieved w.r.t. I/O rates from HDF5 with local data size upto 0.13GB

Conclusions

It is observed that ADIOS2 BP4 I/O engine provides much better bandwidth than the other I/O layer backends. This is possibly due to factors such as the innovative BP4 native metadata system and write buffering system. From figure 4 it is observed that better rates are obtained by using NextGenIO compared to Fulhame.

In the future, larger sized arrays would be used for benchmarking. It would be useful to investigate the benefits of ADIOS2 BP4 I/O engine and its many configurable options. It is also planned to investigate advanced hardware such as the NVRAM storage of NextGenIO [6].

Acknowledgments

The Fulhame HPE Apollo 70 system is supplied to EPCC, the supercomputing centre at the University of Edinburgh, as part of the Catalyst UK programme, a collaboration with Hewlett Packard Enterprise, Arm and SUSE to accelerate the adoption of Arm based supercomputer applications in the UK.
The NEXTGenIO system was funded by the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement no. 671951.
This work was supported by an EPCC funded studentship as part of the ASiMoV project. Funding from EPCC is gratefully acknowledged.

References

[1] benchio. `https://github.com/EPCCed/benchio.git`.
[2] benchmark_c. `https://github.com/sb15895/benchmark_c.git`.
[3] Fulhame. `https://www.epcc.ed.ac.uk/facilities/other-facilities/fulhame`.
[4] Nextgenio. `http://www.nextgenio.eu`.
[5] Xcompact3d. `github.com/xcompact3d/Incompact3d`.
[6] Adrian Jackson, Michèle Weiland, Mark Parsons, and Bernhard Homölle. An Architecture for High Performance Computing and Data Systems Using Byte-Addressable Persistent Memory. 2019. doi:10.1007/978-3-030-34356-9_21.


COMPUTING
INSIGHT UK 2022
UKRI
Science and Technology Facilities Council
SAVE THE DATE
1-2 December 2022
Manchester Central, UK
SEE YOU IN 2022
www.stfc.ac.uk/ciuk