WHAT IS RISC-V AND WHY SHOULD WE CARE?

Nick Brown (EPCC)

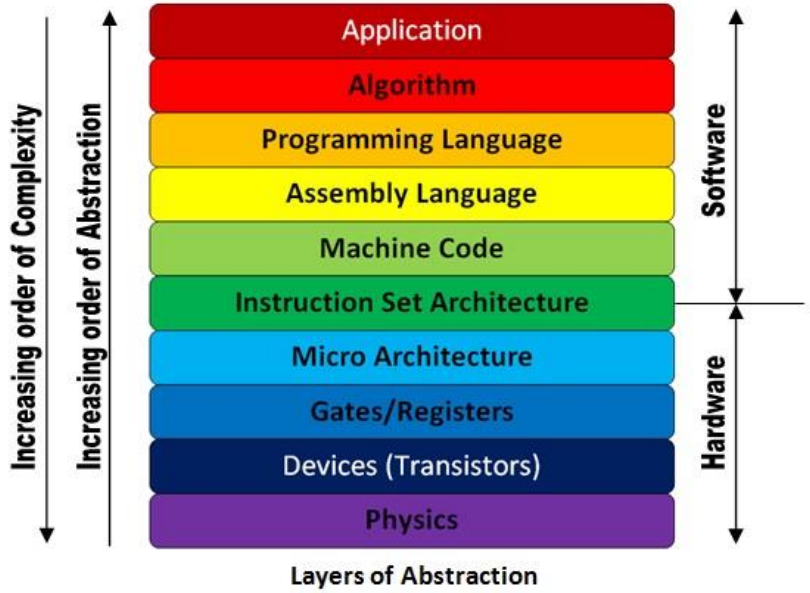
n.brown@epcc.ed.ac.uk

epcc

THE UNIVERSITY OF EDINBURGH

What is RISC-V

- Started out by Berkley in 2012
- An open Instruction Set Architecture (ISA) which is overseen by RISC-V International
  - Standardisation activities driven by expert members
  - Numerous areas of focus ranging from HPC & ML to the data centre to embedded computing



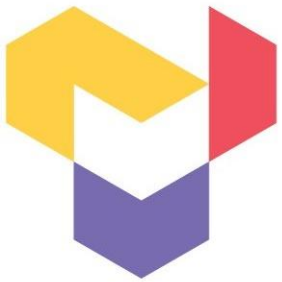
- RISC-V is an Open Standard Instruction Set Architecture (ISA)
  - Software uses the ISA to tell the hardware what to do.
  - At the base level, the RISC-V ISA and extensions ratified by RISC-V International are royalty free and open base building blocks for anyone to build their own solutions and services on

RISC-V International is the global standards body

- **4k+ individuals in 60+ RISC-V work groups** and committees
- **330+ RISC-V solutions** online including cores, SoCs, software, tools, and developer boards
- **35 local RISC-V community** groups, with more than **6,800+ engineers**
- **Over 4200** members in **70+** countries

RISC-V®

Lots of serious players here



tenstorrent

Qualcomm

SiFive

RiVOS

InspireSemi™

epcc

THE UNIVERSITY OF EDINBURGH

RISC-V HPC Special Interest Group (SIG)

- The group in RISC-V International that is concerned with driving the update of RISC-V in HPC
- Lots of members across a wide range of organisations and vendors
- Do a whole load of activities:
  - Organise workshops, panels etc at SC, ISC, HPC Asia
  - Marketing of RISC-V in HPC
  - Gap analysis to understand what is missing
  - Explore how other parts of the standard/activities can be leveraged for HPC

Why for HPC?

- Modularity and freedom to design bespoke hardware is the key advantage
- Especially as we see an increased focus on energy efficiency

ISA	Chips?	Architecture License?	Commercial Core IP?	Add Own Instructions?	Open-Source Core IP?
x86	Yes, ***three*** vendors	***No***	***No***	***No***	***No***
ARM	Yes, ***many*** vendors	Yes, ***expensive***	Yes, ***one*** vendor	***No*** *(Mostly)*	***No***
RISC-V	Yes, ***many*** vendors	Yes, ***free***	Yes, ***many*** vendors	***Yes***	***Yes, many available***

- We have set up a RISC-V testbed for HPC, enabling free access for scientific developers to experiment with the technology for their workloads
- Looks/feels like any other HPC machine, with login node, shared filesystem, module environment and Slurm queue providing access to compute nodes

epcc | THE UNIVERSITY OF EDINBURGH

Initially a scrabble to get hardware…..

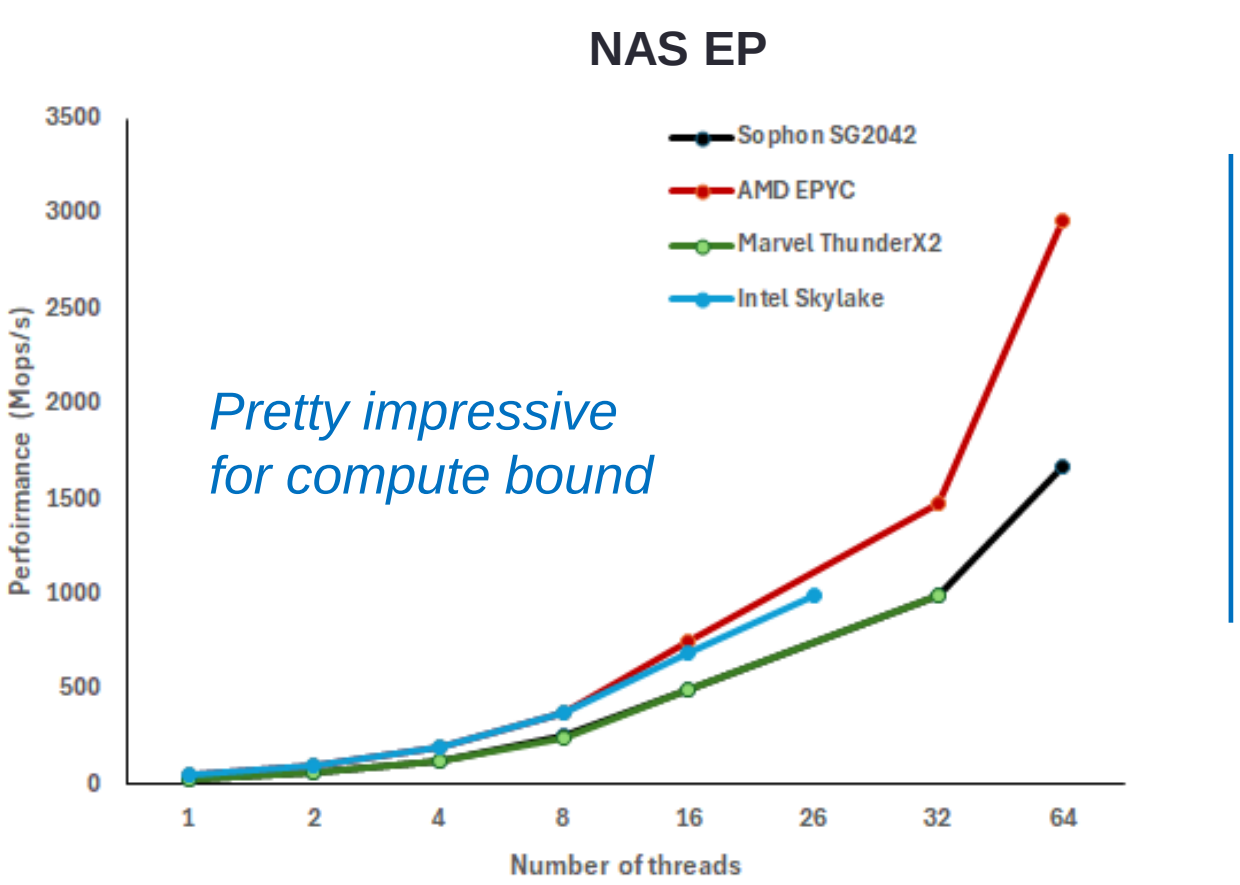
- When we initially set this up in early 2022, hardware availability was a big challenge
  - Had to rely on SBC designed more for embedded workloads

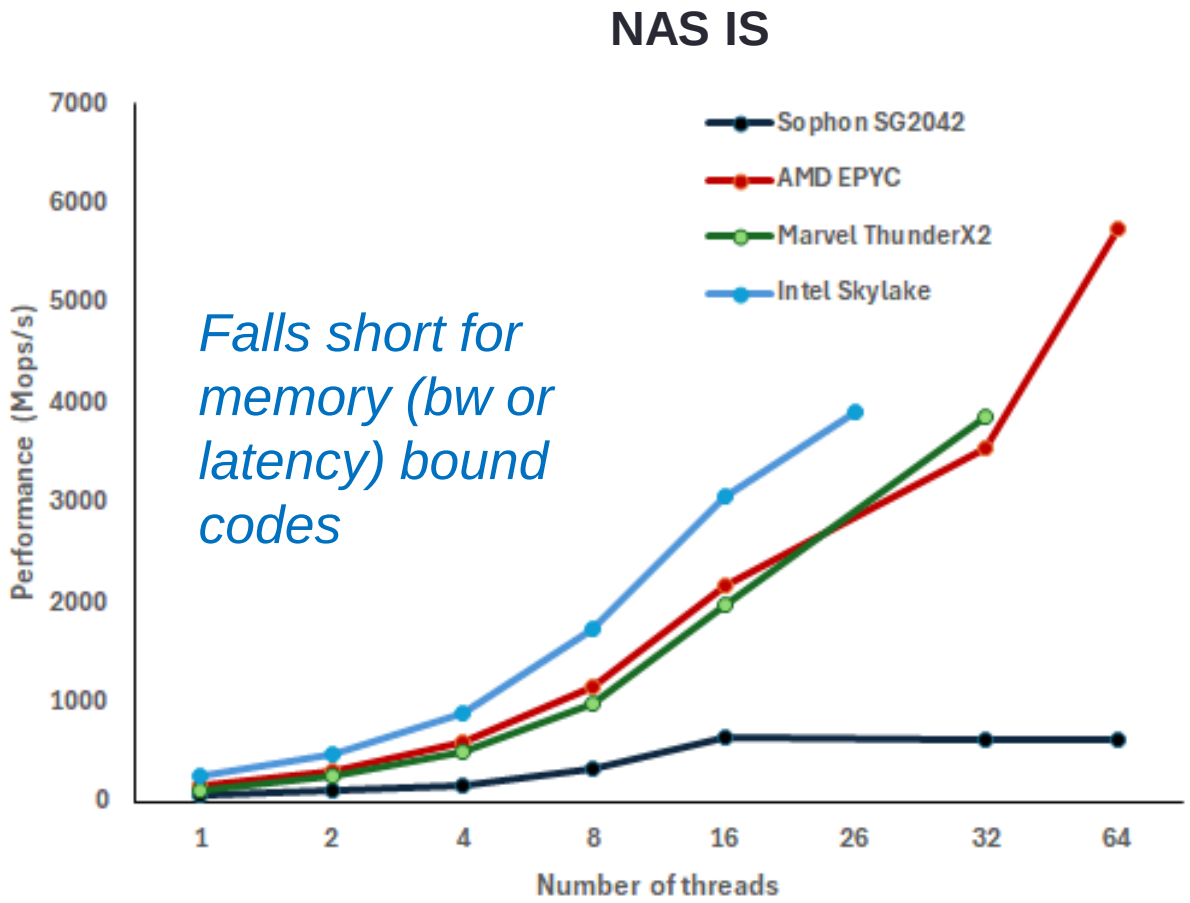
- In 2023 Sophon released their 64-core SG2042 RISC-V CPU designed for high performance workloads
  - A much more realistic proposition for HPC
  - Core for core, each for of the SG2042 is at-least three times as fast (and sometimes many times) what is in these SoCs

epcc

THE UNIVERSITY OF EDINBURGH

How do RISC-V CPUs currently compare?





Lots more details at https://arxiv.org/pdf/2406.12394 and https://browse.arxiv.org/pdf/2309.00381.pdf

epcc

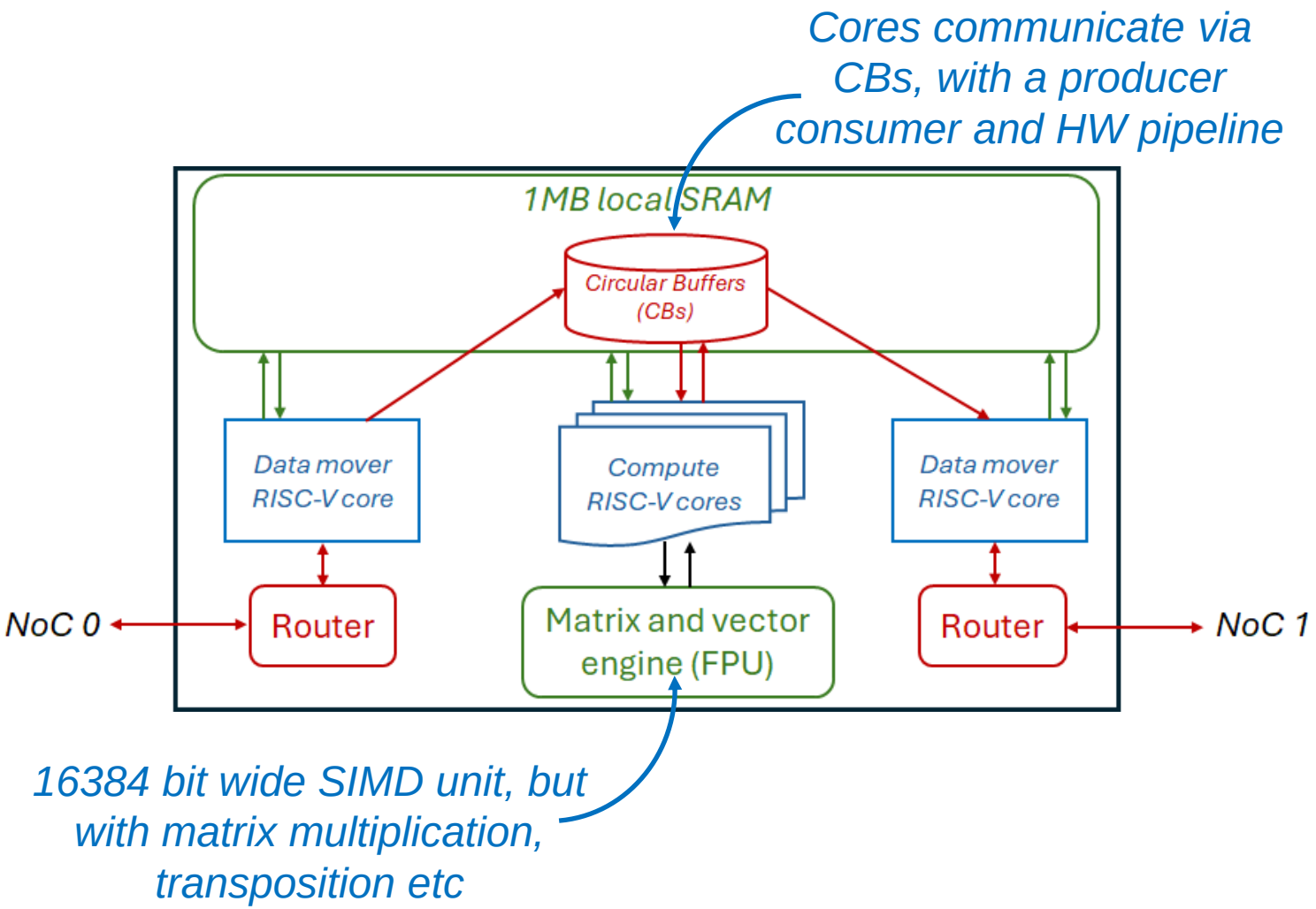
THE UNIVERSITY OF EDINBURGH

RISC-V PCIe accelerator cards

epcc

THE UNIVERSITY OF EDINBURGH

Deep dive: Tenstorrent accelerator



- 120 of these Tensix cores in the Grayskull with 8GB of DDR4
- 128 of these Tensix cores in the Wormhole with 24GB of DDR6
- Goes back to the ability to specialise here
- RISC-V means a common compiler can be used, although need to use their specific API for programming

epcc | THE UNIVERSITY OF EDINBURGH

Experimenting with Tenstorrent Grayskull

- The architecture has been designed for AI inference, but TT develop an open source Metalium framework that enables direct programming

- We ported a stencil code to the accelerator
  - Stencils are ubiquitous in scientific computing, whilst this is a simple Jacobi solver for LaPlace's equation for diffusion it enabled us to explore the appropriate code techniques in detail
  - Initially, was way slower than a core of the CPU (Xeon Platiunum Cascade Lake)

Version	Performance (GPt/s)
CPU single core	1.41
Initial	0.0065

Initially a Tensix core was considerably slower than a CPU core!

epcc

THE UNIVERSITY OF EDINBURGH

Experimenting with Tenstorrent Grayskull

- We undertook a range of experimentation to understand the most appropriate approaches for data movement and where bottlenecks may lie
  - Considering how best to map the domain to the FPUs in the Tensix cores, and most effectively leverage the data movement capabilities

- Ultimately, slightly better performance but at five times less energy.

Based upon this optimisation, we were able to improve performance on a single Tensix core by around 160 times

Across the entire chip, we can slightly outperform the 24-core Xeon Platinum (Cascade Lake) but at five times less energy usage

Type	Total cores	Cores in Y	Cores in X	Performance (GPt/s)	Energy (Joules)
CPU	1	-	-	1.41	1657
CPU	24	-	-	21.61	588
e150	1	1	1	1.06	2094
e150	2	1	2	2.48	893
e150	4	1	4	2.92	744
e150	8	4	4	7.99	276
e150	32	8	4	9.20	240
e150	64	8	8	12.96	170
e150	72	8	9	17.26	128
e150	108	12	9	22.06	110

Much more detail at https://arxiv.org/pdf/2409.18835

epcc

THE UNIVERSITY OF EDINBURGH

Conclusions & get involved!

- RISC-V is growing extremely rapidly
  - There is still plenty to do here, but I think lots of potential for scientific computing and HPC more generally. This is our opportunity to shape this technology to suit our needs.

- A wide range of activities are going on to push RISC-V more into HPC and mature the ecosystem

- Get involved!
  - https://riscv.epcc.ed.ac.uk
  - RISC-V International is free to join for individuals and academic institutions
    - HPC SIG meets virtually every month (and lots of other SIGs too!)

epcc

THE UNIVERSITY OF EDINBURGH