



**Barcelona
Supercomputing
Center**

Centro Nacional de Supercomputación



**EXCELENCIA
SEVERO
OCHOA**



**UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH**



Challenges and Opportunities for RISC-V Architectures towards Genomics Workloads

Gonzalo Gómez, **Aaron Call**, Xavier Teruel, Lorena Alonso, Ignasi Moran, Miguel Ángel Pérez, David Torrents, Josep Ll. Berral

Universitat Politècnica de Catalunya
Barcelona Supercomputing Center
Institut Català de Recerca i Estudis Avançats

May/2023



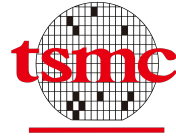
MEEP

MareNostrum Experimental
Exascale Platform



Motivation: EU-Technological sovereignty

- All electronic components nowadays have a chip (semiconductor) inside.
- Current semiconductors are manufactured mainly in Taiwan and a small part in the USA.



- Chips are designed either by US or Chinese companies
- Europe does neither design nor manufactures its own chips: thus it has a technological dependency on outside countries.



- Consequently we do have a lack of technological sovereignty



Opportunities & Challenges

- EU advocates for:
 - Open-hardware
 - Open-repositories



- Challenges
 - Lack of support for a full software stack.
 - Not actual hardware on RISC-V high-performing as x86 architectures.



Our Contributions

- Contribution 1: A Benchmark for Scientific HPC-based Analytics Application for RISC-V, adapted to the capabilities of current RISC-V implementations.
- Contribution 2: The identification of the challenges explaining the performance differences between RISC-V implementations and x86 on real HPC applications.
- Contribution 3: A discussion and recommendations on the progress and improvement in RISC-V towards next step designs.
- Contribution 4: The creation of a publicly available open-data repository of benchmarks to run on RISC-V platforms.

GENOMIC WORKLOAD VARIANT INTERACTION ANALYSIS



MEEP

MareNostrum Experimental
Exascale Platform



**Barcelona
Supercomputing
Center**

Centro Nacional de Supercomputación



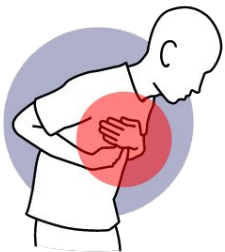
NEAR DATA



VITAMIN V

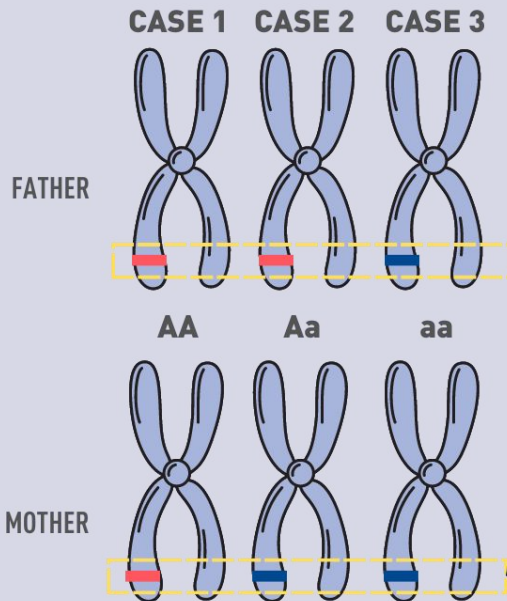
THE DATA

COMPLEX DISEASES



ASTHMA
TYPE 2 DIABETES
ALZHEIMER'S
...

GENOMIC VARIANTS



THREE POSSIBLE CASES

AA: reference-reference
Aa: reference-alternative
aa: alternative-alternative

INTERACTION EFFECT

	VARIANT 3	VARIANT 4
VARIANT 1		
VARIANT 2		

THE DATA

INPUT DATA



1,883,192 PAIRS

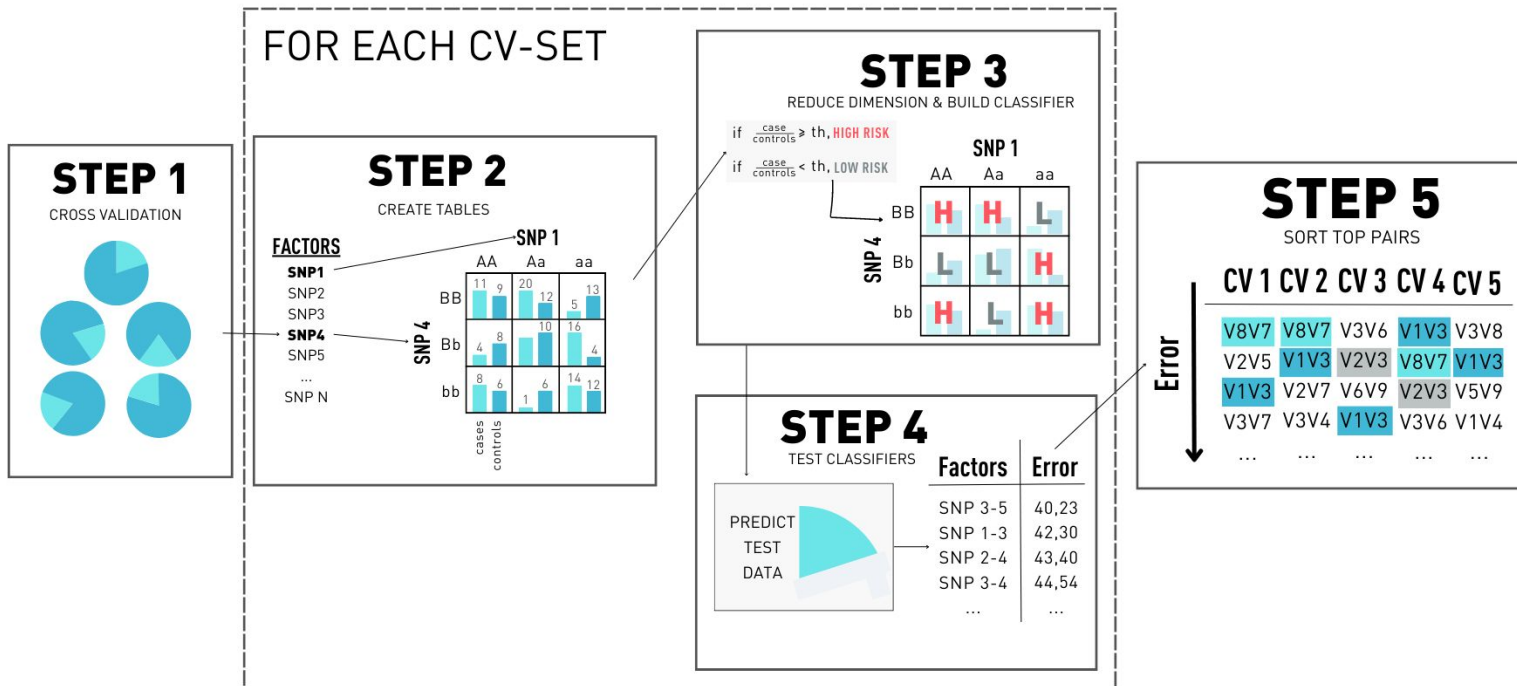
1,128 PATIENTS

LABELS

PATIENT 1	CASE
PATIENT 2	CONTROL
PATINET 3	CASE
...	..
PATIENT 1,128	CONTROL

	p1	p2	p3	p4	p5	...
VARIANT 1	AA	AA	Aa	AA	aa	...
VARIANT 2	aa	AA	Aa	Aa	AA	...
VARIANT 3	Aa	aa	AA	AA	Aa	...
...						

THE WORKLOAD



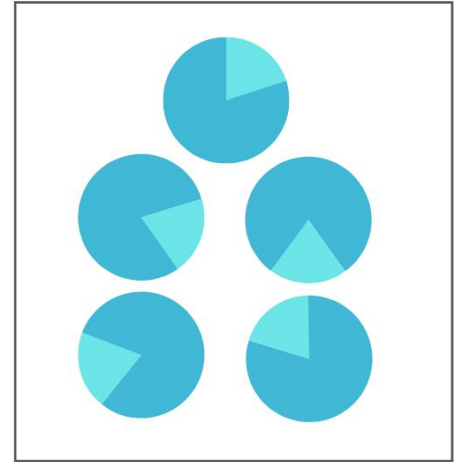
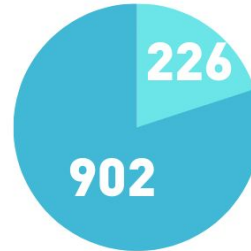
STEP 1 - CROSS VALIDATION

INPUT DATA

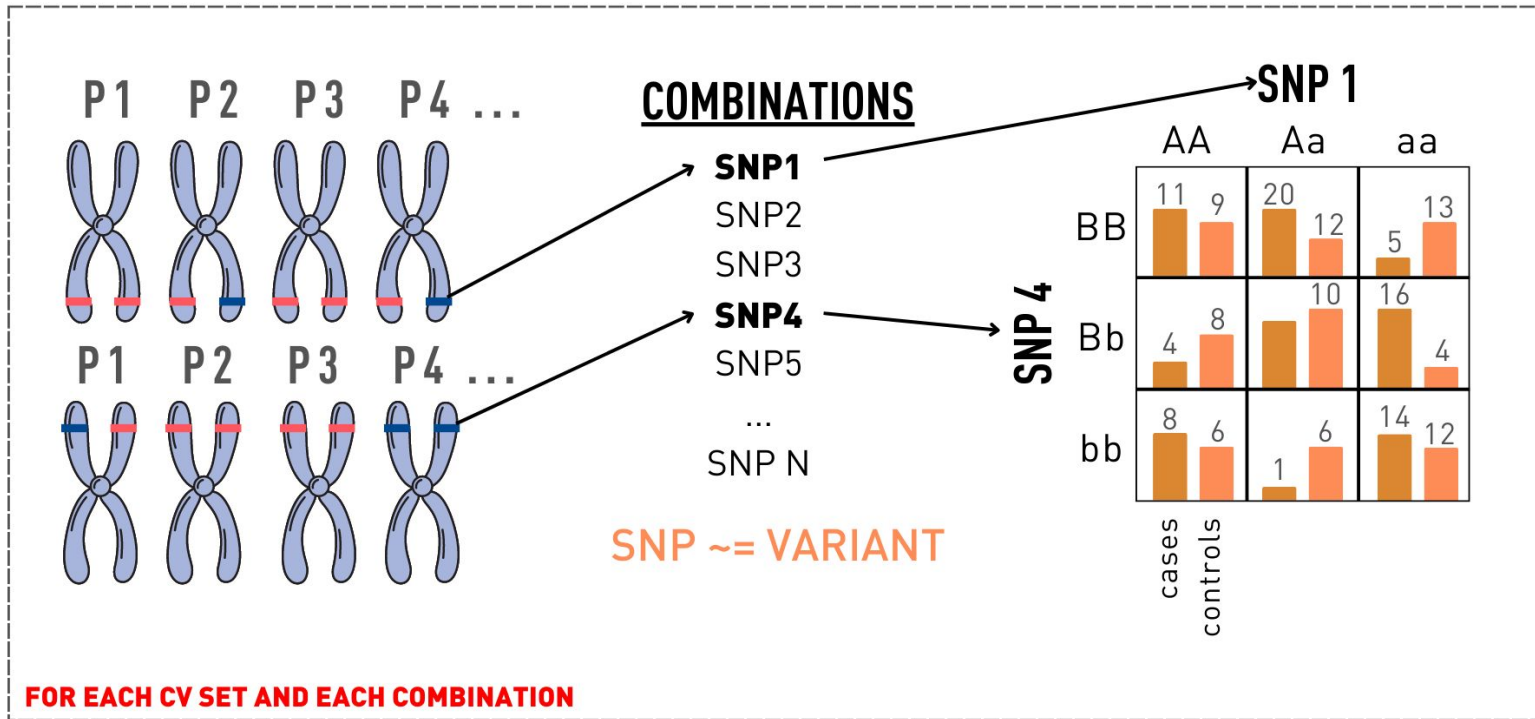


1,128 PATIENTS

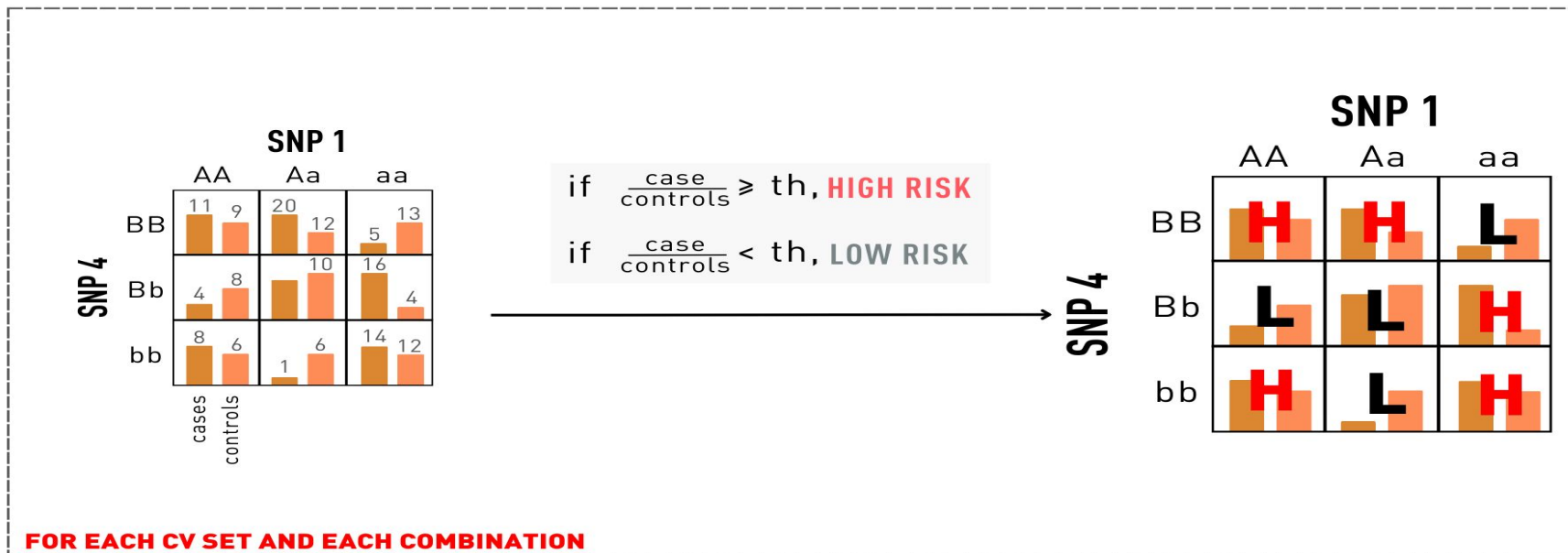
MAKE 5 CV SETS



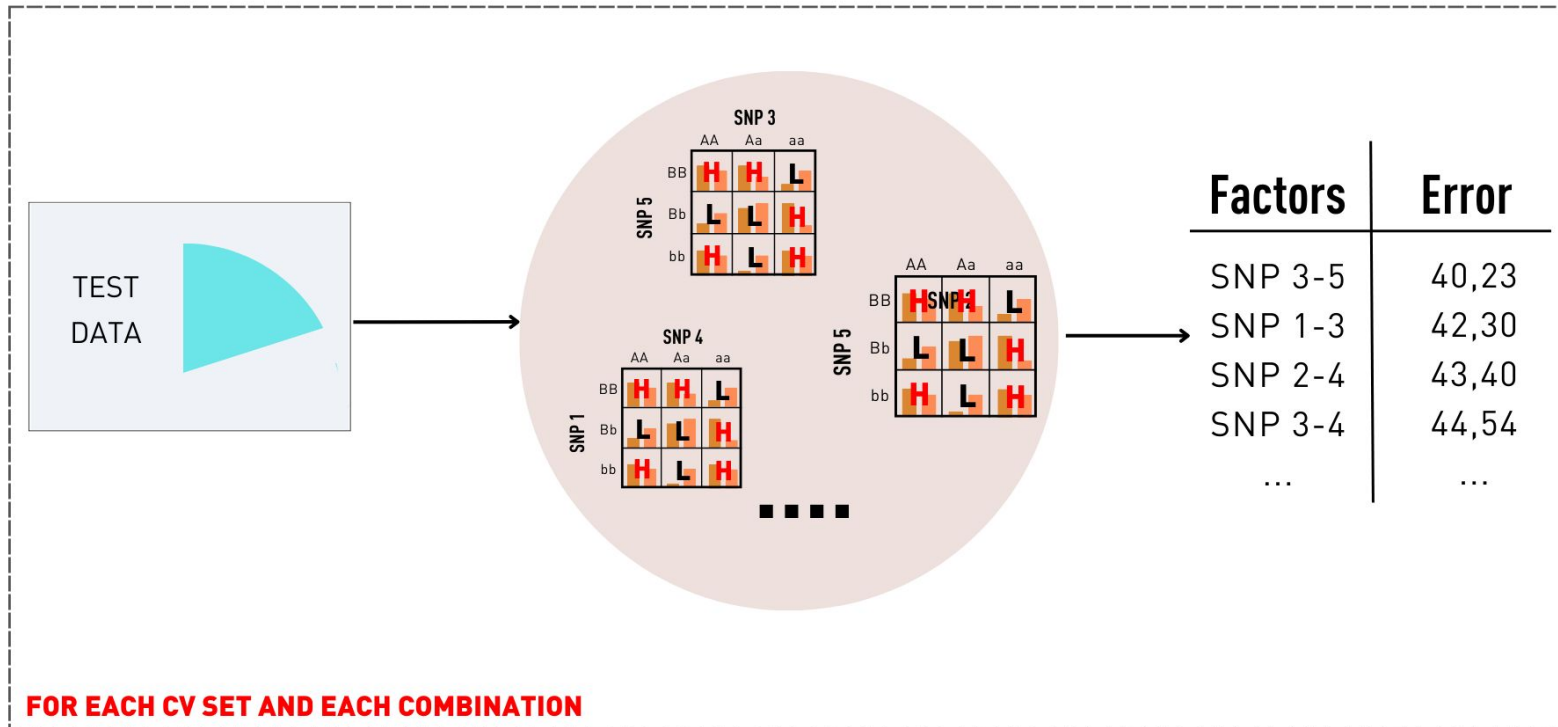
STEP 2 - CREATE CONTINGENCY TABLES



STEP 3 - BUILD CLASSIFIERS



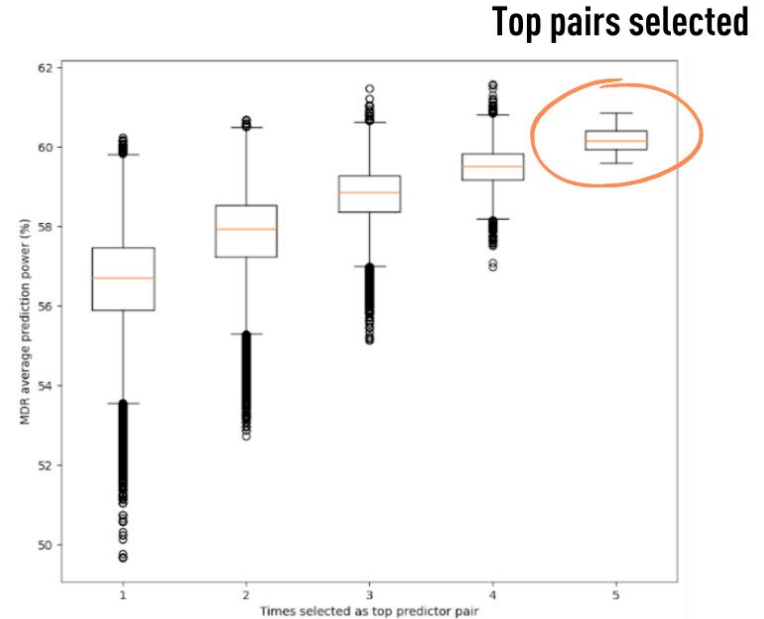
STEP 4 - TEST CLASSIFIERS



STEP 5 - SELECT TOP PAIRS

Error ↓

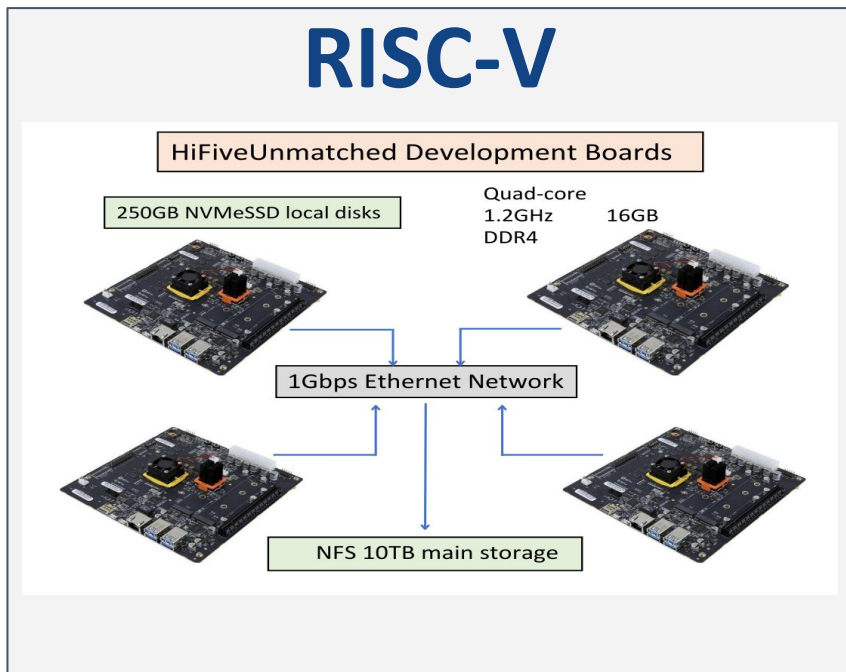
	CV 1	CV 2	CV 3	CV 4	CV 5
	SNP 8-7	SNP 8-7	SNP 3-6	SNP 1-3	SNP 3-8
	SNP 2-5	SNP 1-3	SNP 2-3	SNP 8-7	SNP 1-3
	SNP 1-3	SNP 2-7	SNP 6-9	SNP 2-3	SNP 5-9
	SNP 3-7	SNP 3-4	SNP 1-3	SNP 3-6	SNP 1-4



BENCHMARKING EXPERIMENTS

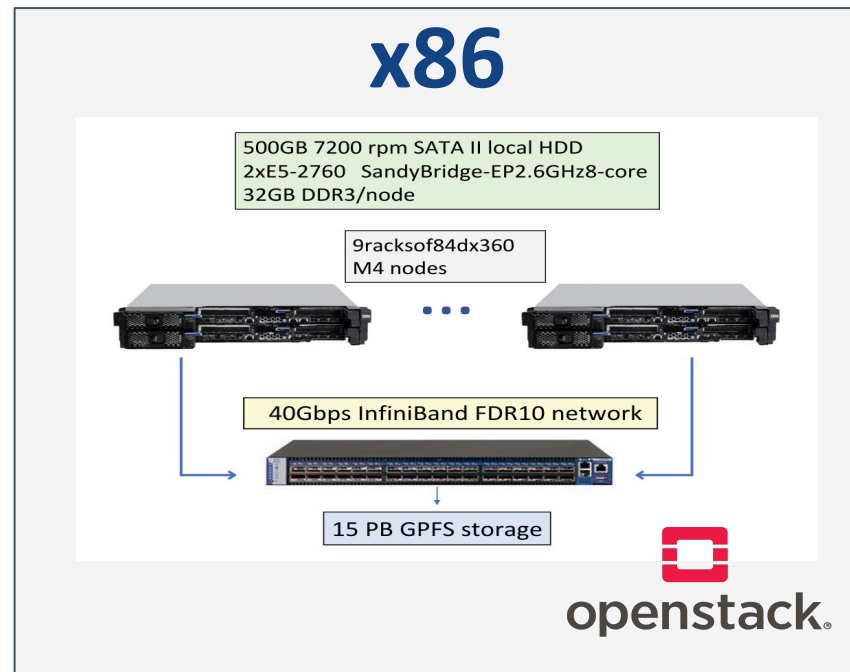
Environment Setup

RISC-V



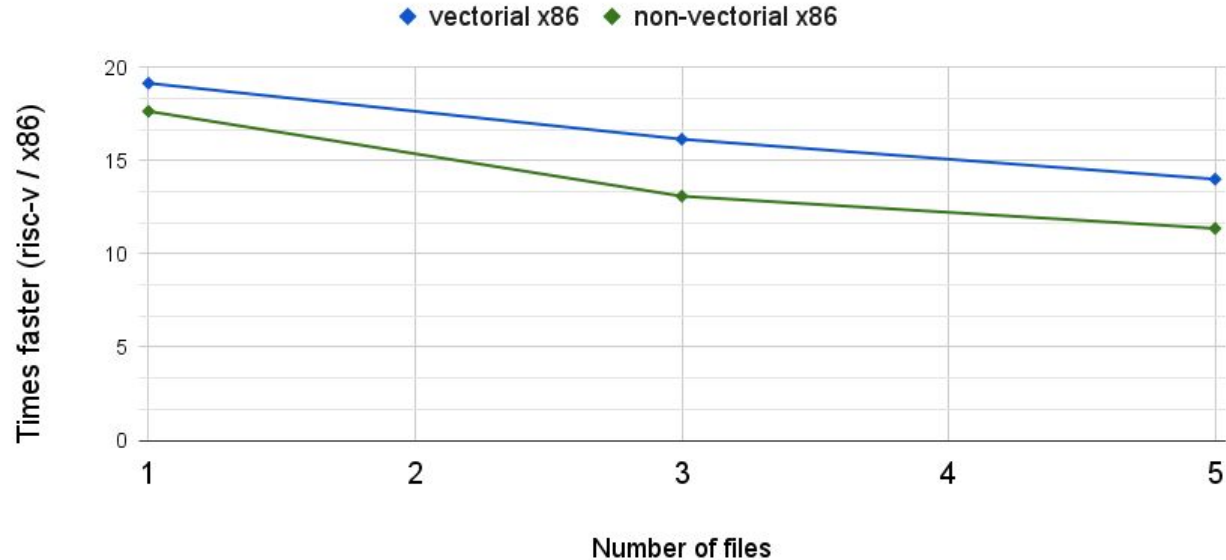
VS

x86



Experiment 1: Vectorial vs non-vectorial

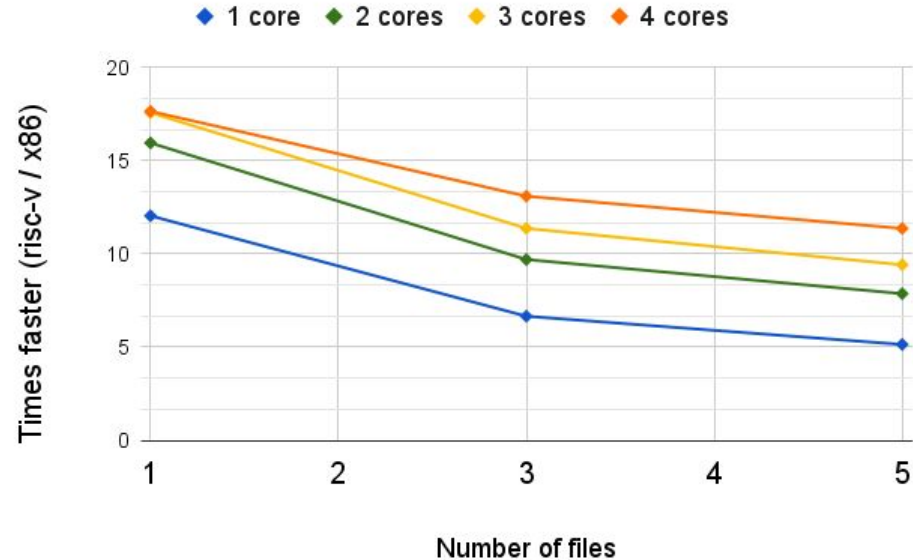
2 NODES AND 4 CORES



CONCLUSION: vectorial ops are an important element to decrease the gap with x86

Experiment 2: Cores scalability

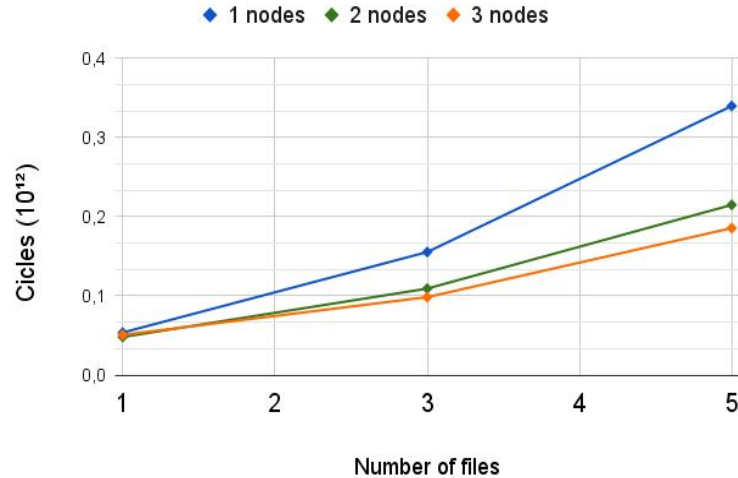
2 NODES



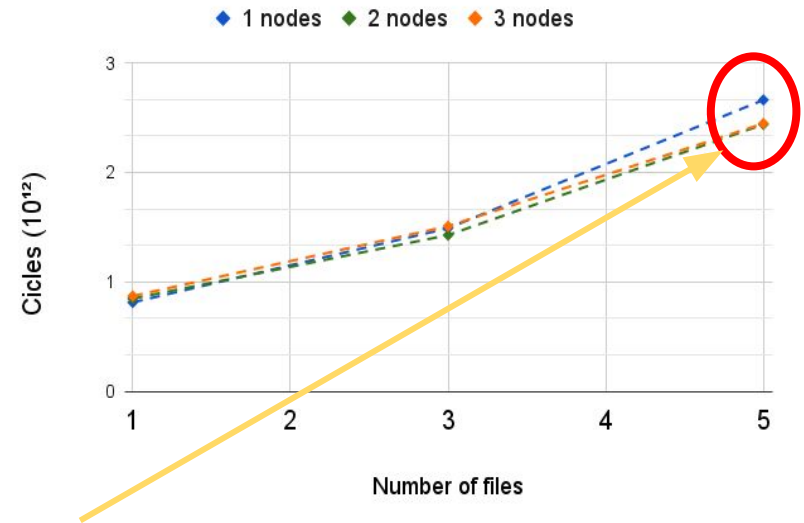
CONCLUSION: both scale on cores, but x86 is faster with more cores

Experiment 3: Nodes scalability

X86 USING 4 CORES

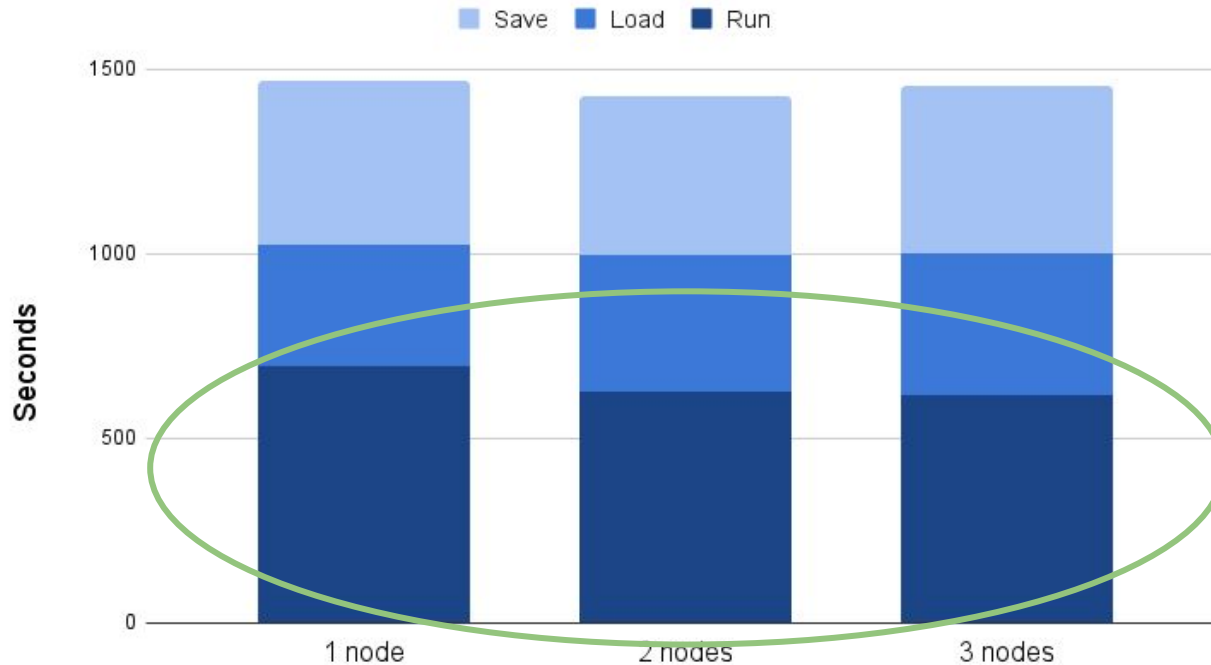


RISC-V USING 4 CORES



Why is not scaling by nodes??

Experiment 4: workload times in Risc-v



CONCLUSION: save and load time are hiding the improvements on running time

Open repository

- A list of results can be found in:
https://github.com/MortI2C/genomics_riscv_openrepo
- And the workload is available at:
<https://gitlab.bsc.es/datacentric-computing/via>
- WiP: available from public website

Conclusions & Future Work

- Vectorial instructions are a significant element to cover the performance gap with x86
- Data loading process is expensive on RISC-V systems and avoids to scale properly
 - It could be improved via using HDFS - which performs data distribution prior to workloads' execution -.
 - Fine-grained monitoring tools in our system made the runs slower, preventing to acquire valid and detailed data
- There is a need to find a proper mapping between x86 and RISC-V architectures so they can be run equivalently
- If we want RISC-V to become the new standard we need to fulfill end-users requirements in performance as well

Thanks for your attention

gonzalo.gomez@bsc.es
aaron.call@bsc.es