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
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
THE DATA

COMPLEX DISEASES

ASTHMA
TYPE 2 DIABETES
ALZHEIMER’S
...

GENOMIC VARIANTS

CASE 1  CASE 2  CASE 3

FATHER

AA  Aa  aa

MOTHER

THREE POSSIBLE CASES
AA: reference-reference
Aa: reference-alternative
aa: alternative-alternative

INTERACTION EFFECT

<table>
<thead>
<tr>
<th>VARIANT 3</th>
<th>VARIANT 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

HELP!
THE WORKLOAD

FOR EACH CV-SET

STEP 1
CROSS VALIDATION

STEP 2
CREATE TABLES

FACTORS
SNP1
SNP2
SNP3
SNP4
SNP5
SNP N

SNP 1
BB
AA
Aa
aa

STEP 3
REDUCE DIMENSION & BUILD CLASSIFIER

SNP 1
BB
AA
Aa
aa

STEP 4
TEST CLASSIFIERS

Factors | Error
---|---
SNP 3-5 | 40.23
SNP 1-3 | 42.30
SNP 2-4 | 43.40
SNP 3-4 | 44.54

STEP 5
SORT TOP PAIRS

CV 1 CV 2 CV 3 CV 4 CV 5
V8V7 V8V7 V3V6 V1V3 V3V8
V2V5 V1V3 V2V3 V8V7 V1V3
V1V3 V2V7 V6V9 V2V3 V5V9
V3V7 V3V4 V1V3 V3V6 V1V4

Error
STEP 1 - CROSS VALIDATION

INPUT DATA

1,128 PATIENTS

MAKE 5 CV SETS

226

902
STEP 2 - CREATE CONTINGENCY TABLES

FOR EACH CV SET AND EACH COMBINATION
STEP 3 - BUILD CLASSIFIERS

FOR EACH CV SET AND EACH COMBINATION

if \( \frac{\text{case}}{\text{controls}} \geq \text{th} \), HIGH RISK

if \( \frac{\text{case}}{\text{controls}} < \text{th} \), LOW RISK

SNP 1

<table>
<thead>
<tr>
<th>SNP 4</th>
<th>AA</th>
<th>Aa</th>
<th>aa</th>
</tr>
</thead>
<tbody>
<tr>
<td>BB</td>
<td>11</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>Bb</td>
<td>4</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>bb</td>
<td>8</td>
<td>6</td>
<td>14</td>
</tr>
</tbody>
</table>

SNP 1

<table>
<thead>
<tr>
<th>SNP 4</th>
<th>AA</th>
<th>Aa</th>
<th>aa</th>
</tr>
</thead>
<tbody>
<tr>
<td>BB</td>
<td>H</td>
<td>H</td>
<td>L</td>
</tr>
<tr>
<td>Bb</td>
<td>L</td>
<td>L</td>
<td>H</td>
</tr>
<tr>
<td>bb</td>
<td>H</td>
<td>L</td>
<td>H</td>
</tr>
</tbody>
</table>
STEP 4 - TEST CLASSIFIERS

FOR EACH CV SET AND EACH COMBINATION

<table>
<thead>
<tr>
<th>Factors</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNP 3-5</td>
<td>40.23</td>
</tr>
<tr>
<td>SNP 1-3</td>
<td>42.30</td>
</tr>
<tr>
<td>SNP 2-4</td>
<td>43.40</td>
</tr>
<tr>
<td>SNP 3-4</td>
<td>44.54</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
STEP 5 - SELECT TOP PAIRS

<table>
<thead>
<tr>
<th>CV 1</th>
<th>CV 2</th>
<th>CV 3</th>
<th>CV 4</th>
<th>CV 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNP 8-7</td>
<td>SNP 8-7</td>
<td>SNP 3-6</td>
<td>SNP 1-3</td>
<td>SNP 3-8</td>
</tr>
<tr>
<td>SNP 2-5</td>
<td>SNP 1-3</td>
<td>SNP 2-3</td>
<td>SNP 8-7</td>
<td>SNP 1-3</td>
</tr>
<tr>
<td>SNP 1-3</td>
<td>SNP 2-7</td>
<td>SNP 6-9</td>
<td>SNP 2-3</td>
<td>SNP 5-9</td>
</tr>
<tr>
<td>SNP 3-7</td>
<td>SNP 3-4</td>
<td>SNP 1-3</td>
<td>SNP 3-6</td>
<td>SNP 1-4</td>
</tr>
</tbody>
</table>

Top pairs selected

Error
BENCHMARKING EXPERIMENTS
Environment Setup

**RISC-V**

- HiFiveUnmatched Development Boards
- 250GB NVMeSSD local disks
- Quad-core 1.2GHz DDR4
- 16GB
- 1Gbps Ethernet Network
- NFS 10TB main storage

**x86**

- 500GB 7200 rpm SATA II local HDD
- 2xE5-2760 SandyBridge-EP2.6GHz-8-core
- 32GB DDR3/node
- 9racksof84dx360 M4 nodes
- 40Gbps InfiniBand FDR10 network
- 15 PB GPFS storage

VS

**openstack**
Experiment 1: Vectorial vs non-vectorial

CONCLUSION: vectorial ops are an important element to decrease the gap with x86
Experiment 2: Cores scalability

CONCLUSION: both scale on cores, but x86 is faster with more cores
Experiment 3: Nodes scalability

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