Performance of the Vipera framework for DSLs on micro-core architectures

Citation for published version:
Jamieson, M & Brown, N 2023, Performance of the Vipera framework for DSLs on micro-core architectures. in LNCS Euro-Par DSL-HPC workshop proceedings. vol. 13835, Springer Nature Switzerland AG, International workshop on DSLs for HPC, Glasgow, United Kingdom, 22/08/22.

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Early version, also known as pre-print

Published In:
LNCS Euro-Par DSL-HPC workshop proceedings

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Performance of the Vipera framework for DSLs on micro-core architectures

Maurice Jamieson and Nick Brown
EPCC, The University of Edinburgh
Edinburgh, UK
maurice.jamieson@ed.ac.uk
https://www.epcc.ed.ac.uk

Abstract. Vipera provides a compiler and runtime framework for implementing dynamic Domain-Specific Languages on micro-core architectures. The performance and code size of the generated code is critical on these architectures. In this paper we present the results of our investigations into the efficiency of Vipera in terms of code performance and size.

Keywords: Domain-specific languages · Python · native code generation · RISC-V · micro-core architectures

1 Introduction

In order to reduce the power consumption of new High-Performance Computing (HPC) machines, the use of hybrid HPC architectures with graphics processing units (GPUs) as accelerators has increased, such as the 4:1 ratio of GPUs to central processing units (CPUs) per node of the new OLCF Frontier exascale supercomputer[1]. Other novel architectures for HPC have been introduced, including innovative micro-core\(^1\) processor architectures that consist of many, low energy cores combined with small amounts of memory on a single chip, such as the 256 core Kalray MPPA, the 256 core Sunway SW26010 , the 1024 Adapteva Epiphany-V and the 2048 core PEZY-SC2 . These micro-core architectures have the promise of overcoming the power wall due to the high energy efficiency of their designs, for example, the class-leading 70 GFLOPS per Watt of the the 64-core Adapteva Epiphany-IV [3]. Whilst these architectures provide the high energy efficiency and low overall power consumption levels, micro-cores are notoriously difficult to program and take advantage of; each technology is different with its own idiosyncrasies, such as the topology of the Network-on-Chip (NOC), and they each present a different low-level interface to the programmer. Although manufacturers have made great progress in developing the hardware, parallel programming and compilation techniques have not evolved

\(^1\)Although the term manycore is commonly used, we define micro-cores as many-cores with extremely small amounts of on-chip, scratchpad RAM (circa 32 - 64KB) without hardware cache support.
quickly enough to exploit this effectively[23]. Fundamentally, writing parallel, scalable code is difficult and requires the programmer to consider multiple levels of parallelism to get good performance[25]. However, to date, these technologies have tended to result in significant performance overheads, required the programmer to ensure their code fits within the limited on-chip memory, provided limited choices around data location and size, and provided little, if any, portability across architectures. As evidenced by ePython[12], a Python interpreter for the Epiphany-III, dynamic programming languages can significantly reduce the programming effort required to overcome these complexities in comparison to the provided, low-level C software development kits (SDKs)[22].

In this paper we present the investigations into the efficiency of our Vipera framework for dynamic programming languages, in terms of code performance and size, relative to handwritten (native) C, on a variety of micro-core and traditional CPU architectures.

2 Background and related work

Whilst Python is currently the most popular programming language[2], its use of an interpreter results in performance significantly slower than statically compiled languages, such as C and Fortran. This has driven the need to overcome the performance overhead of the interpreter and the restrictions imposed by the global interpreter lock (GIL). This has resulted in technologies to increase the performance of existing Python codes through the compilation to native code, including Cython[11], MicroPython[9], Numba[24], Copperhead[14], Parakeet[26], ALPyNA[19] and PyCUDA[4]. The high-level approach of Numba, Copperhead and Parakeet is similar, whereby they define an embedded domain specific language (eDSL) and utilise Python function decorators (directives) to annotate the code to be compiled to native code or offloaded to GPUs. ALPyNA adopts a different technique to generating GPU code than the eDSL and function decorator approach. Rather than requiring the programmer to select and annotate the Python functions that will be generated as GPU kernels, ALPyNA analyses loop data dependencies and performs automatic loop parallelisation to generate CUDA kernels for GPUs. However, unlike Numba, Copperhead, Parakeet and ALPyNA, PyCUDA does not abstract the generation of GPU code but instead embeds CUDA C code directly within the Python source code. MicroPython performs the compilation of bytecode to native code on the device[8] similar to JIT except that the bytecode is not profiled as is common for JIT compilers, rather the bytecode is just lowered to native code. An alternative approach was taken for Vipera, similar to that employed by the Pallene / Titan compiler[17] for Lua[18]. Here, the source language compiler, running on the host, emits C source code that is then compiled to generate native binary executables.

2.1 Vipera dynamic language framework

The Vipera[13] framework was created to support the development of dynamic languages on micro-core architectures. The framework consists of a layered ar-
chitecture with components running on the host and micro-core devices. Vipera manages the compilation of code, the transfer and launch of kernels on the micro-core devices, and the transfer of data. vPython is a development of ePython, a subset of the Python programming language specifically designed for micro-core architectures. Vipera provides two implementations of this; the first compiles down to bytecode that executes on a tiny virtual machine (c. 24KB on the Adapteva Epiphany-III[12]) running on the device and the second generates Olympus abstract machine code that is compiled to provide device native code. In this paper we will focus on the Olympus abstract machine version of vPython.

vPython can either be run standalone on the device or as a Domain-Specific Language (DSL) within Python running on the host, offloading kernels for execution to the device. More information on the parallel programming, offloading and dynamic code loading capabilities of the language can be found in [22] and [21].

3 Benchmarking

3.1 CPU selection

In order to support the assessment of the Vipera vPython compiler and Olympus abstract machine, a number of different platforms and processors were selected, including the Adapteva Epiphany-III, Xilinx MicroBlaze and PicoRV32 RISC-V micro-cores and the AMD64 (x64), ARM Cortex-A9 (ARM32), MIPS32, SPARCv9 and U740 RISC-V (RISCV64) traditional CPUs. As processor ISAs can have a significant impact on both the compiled kernel performance and binary size, the CPUs were selected to test the impact of the Olympus abstract machine design and to test the portability of Olympus between 32 bit and 64 bit processors with varying alignment constraints and byte ordering.

For the selected benchmarks, LINPACK[15] and the Sieve of Eratosthenes[16], the source vPython codes were compiled to Olympus abstract machine C source code and wrapped by Eithne[20] API calls for execution on a single core of the CPUs.

3.2 LINPACK performance

Figure 1 shows the single-core performance results for LINPACK\(^2\) on the target processor architectures, compiled using the -O0 and -O3 compiler optimisation levels. Whilst the results vary widely across the architectures, the performance difference between the Olympus and native C kernels is very small. However, the Olympus LINPACK kernel compiled at -O0 is 1.7 times faster than native C on the Epiphany-III and is marginally faster (1.5\%) on the ARM32. Although the performance advantage of Olympus kernels on the ARM32 is reversed at -O3, where native C is 4.6\% faster, the advantage is actually slightly increased

\(^2\)Due to the extremely small memory available on the micro-core devices, the problem size \(n\) was 50 and for traditional CPUs \(n = 1000\).
analysing the performance advantage of Olympus kernels over native C on the Epiphany-III and at -O3 on the ARM32 requires knowledge of the peculiarities of the Epiphany-III and looking at the assembly language generated by the C compiler. In the case of the Epiphany-III, there are four modes for the floating point unit (FPU) that can be specified at compile time[6]. The default FPU mode is caller, which results in native C kernels being 1.7 times faster than Olympus. The truncate FPU mode does not provide a significant improvement (2.1%) of native C kernels over Olympus. The round-nearest mode provides a 2.1 times performance improvement of native C over the Olympus abstract machine. The int FPU mode, executing integer operations as well as floating point operations in the FPU, delivers a 1.66 and 1.83 times performance advantage of Olympus kernels over native C at for -Os and for -O3, respectively. This result is surprising but considering that the Epiphany-III is a superscalar design that can execute two floating point operations and one integer instruction per clock cycle[10], it is possible to surmise that the Olympus mnemonics can take advantage of the additional two integer operations per clock cycle afforded by the int FPU mode and prevent the pipeline from stalling.

The minor performance advantage of Olympus over native C on the ARM at the -Os compiler optimisation level can be explained by the additional 21 APSR_nzcv opcodes in the native C kernel. This opcode transfers the floating-
point status flags are transferred the ARM application program status register (APSR) and, as [7] state:

*These instructions stall the ARM until all current NEON or VFP operations complete.*

It is also interesting to determine from the disassembly listing of the ARM Olympus kernel that the ARM NEON vector / SIMD instructions (e.g. VLDR, VLMUL and VSTR) are being issued by the C compiler for the Olympus mnemonics, thereby taking advantage of this parallel processing capability of the ARM processor for the LINPACK benchmark.

### 3.3 LINPACK code size

Figure 2 illustrates that the C kernels are significantly smaller than the Olympus kernels on all platforms, at GCC optimisation levels -Os and -O3, for the LINPACK benchmark. The difference in kernel size ranges from around 1.5 times bigger than native C on the Epiphany-III to 2.6 times bigger on the MIPS32, using -O3. Interestingly, the difference ranges from around 2 times bigger than native C on the Epiphany-III to around 3 times bigger on the MIPS32 and AMD64. This suggests that the Olympus mnemonics generate *wordy* C code, whereby a significantly larger number of underlying operations (machine op-codes) are generated by the C compiler in comparison to the equivalent native C operation. However, it should be noted that the Olympus kernels include a full compacting heap manager and other runtime functions required to support the dynamic features of ePython that are absent from the static native C LINPACK kernel.

![Fig. 2. LINPACK benchmark native C and Olympus kernel size](image-url)
The figures for the MicroBlaze reflect the use of the floating-point emulation option for the LINPACK benchmark. Unsurprisingly, the code size difference is greater on the MicroBlaze in comparison to the Epiphany-III, at between 2 and 2.6 times larger (for both compiler optimisation levels), due to the increased number of operations generated by the Olympus mnemonics over native C, which is amplified by the floating-point emulation code required for the MicroBlaze LINPACK benchmark. There is up to a 20% advantage, on the Epiphany-III, in terms of code size in selecting -Os over -O3. However, for the SPARCv9 the advantage is minimal (0.048%) and is actually detrimental on the ARM32 (-1.67%). Overall, there is an average increase in code size of 7.5% selecting -O3 over -Os, which needs to be considered relative to any performance advantage gained by selecting the higher compiler optimisation level. For a micro-core architecture, such as the Epiphany-III, the code size saving of 20% (approximately 1.8KB) could be significant. Therefore, it is important to understand any performance differences between the two compiler optimisation levels.

3.4 Sieve of Eratosthenes performance

The LINPACK benchmark tests the floating point performance of the Olympus abstract machine. Therefore, the Sieve of Eratosthenes⁴ (Sieve) benchmark was selected to determine the size efficiency and integer performance of Olympus relative to handwritten (native) C. Figure 3 shows that, across compiler optimisation levels -Os and -O3, the Sieve benchmark displays a wider performance gap between the Olympus and native C kernels than was observed for the LINPACK benchmark, discussed in Section 3.2. The Olympus Sieve kernel performance ranges from approximately 1.4 times slower than native C on the Epiphany-III to over 5.5 times slower⁵ on the RISCV64. For all CPUs apart from the AMD64, the difference between Olympus and native C kernel performance is smaller at compiler optimisation level -O3 than at -Os. On the RISCV64, the native C Sieve kernel is 5.5 times faster than the Olympus kernel at -Os but is only 4 times faster at -O3.

Whilst the kernel performance difference between the -Os and -O3 GCC optimisation levels is greatest for the RISCV64, all of the RISC CPU Olympus kernels close the performance gap with the native C kernels at -O3. In comparison, the CISC AMD64 native C kernels are 1.7 times faster than Olympus at -Os and 2 times faster at -O3. This suggests that GCC is able to leverage the additional registers available on the RISCV64 over those available on the AMD64 to optimise the Olympus abstract machine code at -O3 optimisation level. However, the results for the Epiphany-III, MIPS32 and SPARCv9 suggest that the additional registers available on the Epiphany-III do not provide an advantage over the 32 available on the the MIPS32 and SPARCv9.

⁴Due to the limited memory on the micro-core devices, the flag array size was reduced (SIZE = 4095) on the Epiphany-III, MicroBlaze and PicoRV32 micro-cores, on the other CPUs, per the original benchmark, SIZE = 8190.

⁵Both bounds of the range at compiler optimisation level -O3.
On the PicoRV32, the Olympus -O3 kernel froze and did not return a value to the host, even though the kernel successfully executed when compiled at optimisation level -Os. As the LINPACK PicoRV32 kernels also failed to execute correctly at -O3, it is likely that the version of the RISC-V compiler used (riscv32-unknown-elf-gcc 8.2.0) is generating code that is invalid for the PicoRV32 at this level of optimisation.

### 3.5 Sieve of Eratosthenes code size

Figure 4 shows the size of the Sieve kernels compiled with -Os and -O3 compiler optimisation levels for all CPUs. Whilst the Olympus kernel sizes are between near parity\(^6\) and 1.9 times\(^7\) that of the native C kernels for the other CPUs, the difference for the PicoRV32 is striking, with the Olympus kernel size around 5 times larger for both -Os and -O3. The Olympus kernel binary size for the PicoRV32 is explained by the fact that the GCC compiler allocates space in the kernel ELF file for the statically allocated C array used for the heap in the Olympus abstract machine. This is best illustrated by the size of the .bss segment reported by the GNU size utility for the Olympus Sieve kernel on the Epiphany-III, as shown in Listing 1.1, where the Olympus abstract machine heap is 24KB, and the 8MB default heap size of the RISCV64 desktop (threaded) kernel, as shown in Listing 1.2.

\[^6\]MicroBlaze GCC optimisation level -Os.

\[^7\]Epiphany-III GCC optimisation level -O3.
Fig. 4. Sieve benchmark native C and Olympus kernel size

<table>
<thead>
<tr>
<th>text</th>
<th>data</th>
<th>bss</th>
<th>dec</th>
<th>hex</th>
<th>filename</th>
</tr>
</thead>
<tbody>
<tr>
<td>4666</td>
<td>1208</td>
<td>25336</td>
<td>31210</td>
<td>79ea</td>
<td>e_task.elf</td>
</tr>
</tbody>
</table>

Listing 1.1. Output of GNU size for Epiphany-III Olympus Sieve kernel

<table>
<thead>
<tr>
<th>text</th>
<th>data</th>
<th>bss</th>
<th>dec</th>
<th>hex</th>
<th>filename</th>
</tr>
</thead>
<tbody>
<tr>
<td>9185</td>
<td>928</td>
<td>8001440</td>
<td>8011553</td>
<td>7a3f21</td>
<td>threaded_sieve.elf</td>
</tr>
</tbody>
</table>

Listing 1.2. Output of GNU size for RISCV64 Olympus Sieve kernel

<table>
<thead>
<tr>
<th>text</th>
<th>data</th>
<th>bss</th>
<th>dec</th>
<th>hex</th>
<th>filename</th>
</tr>
</thead>
<tbody>
<tr>
<td>54668</td>
<td>0</td>
<td>0</td>
<td>54668</td>
<td>d58c</td>
<td>rv_task.elf</td>
</tr>
</tbody>
</table>

Listing 1.3. Output of GNU size for PicoRV32 Olympus Sieve kernel

In contrast, for the PicoRV32, as shown in Listing 1.3, there is is only a single `.text` segment, containing the executable code, static values, strings and the Olympus heap array. This is due to the custom GNU linker file that is required to set up the memory map on the bare-metal PicoRV32 micro-core. The Epiphany-III and MicroBlaze micro-cores require similar custom linker files. However, the PicoRV32 file is unique in that the `KEEP` command is used to prevent the linker from performing dead code removal on the `.text` segment, which is vital to ensure that the PicoRV32 register initialisation is performed. As the register initialisation subroutine is not referenced in the C source code, it would be removed by the GCC linker when the kernel binary is created, if the `KEEP` command was not used. As all functions are placed in the `.text` segment
and no dead code removal is performed, all unused library functions will also be kept in the final binary, unlike the binaries for other CPUs. Although this is an issue for PicorV32 binaries, it impacts both Olympus and native C kernels. Therefore, a more detailed discussion of possible mitigations for this issue will not be provided, except to highlight the benefits of the Olympus dynamic code loading mechanism discussed in [21].

3.6 Optimising loops

Whilst the performance of the Olympus abstract machine closes the gap with native C, the question remained as to whether the Olympus code generator could leverage the constrained vPython for loop to increase performance. Although it is considered unpythonic to use range to provide an index variable to iterate through the elements of a list[5], as shown in lines 2 and 3 of Listing 1.4, rather than accessing an iterator directly as shown in lines 5 and 6, the iterator is immutable and the list element is cannot be updated, whereas the unpythonic approach allows the list element to be updated.

```python
1 arr = ["a", "b", "c"]
2 for i in range(0, len(arr)):
3    arr[i] = "x"
4
5 for i in arr:
6    i = "y"
```

Listing 1.4. Unpythonic and Pythonic list access

Although a while loop with a manual index variable is often used in this case, the unpythonic for loop approach provides a performance benefit in vPython. As the iterator is managed by the Olympus abstract machine and not the programmer, the vPython for loop can leverage a native C local loop index variable, for example $iter_i$ in Listing 1.5. This C variable not only controls the loop iteration but also is used to update the vPython list element, as shown in line 2 of Listing 1.5. In contrast, the while loop requires a lookup of the index variable in the Olympus environment for both loop control and list element updates, as shown in lines 5, 6 and 7 of Listing 1.5.

```c
1 FOR($iter_i$,0,LDI(ADDRL(2)),1)
2 STAI(ADDRL(4),$iter_i$,TRUE);
3 END

5 WHILE((LDI(ADDRL(10))<LDI(ADDRL(2))))
6 STAI(ADDRL(4),LDI(ADDRL(10)),TRUE);
7 STI(ADDRL(10),(LDI(ADDRL(10))+1));
8 END
```

Listing 1.5. Example Olympus abstract machine code for vPython loop constructs
Two vPython variants of the Sieve benchmark were used to determine the performance benefits of the for loop over the while loop alternative. These were compiled at GCC optimisation levels -Os and -O3, and run on the RISCV64. A native C version of the Byte Sieve benchmark was also compiled at both optimisation levels and run for comparison with the vPython variants. As detailed in Table 1, the for loop variant of the vPython Byte Sieve benchmark is approximately 3 times faster at both -Os and -O3 GCC optimisation levels than the while loop variant. Furthermore, the for loop variant closes the performance gap with native C to around 1.5 times slower from approximately 5 times slower for the while loop variant (both at GCC optimisation level -Os).

<table>
<thead>
<tr>
<th>Code variant</th>
<th>GCC -Os</th>
<th>GCC -O3</th>
</tr>
</thead>
<tbody>
<tr>
<td>vPython while</td>
<td>7.22</td>
<td>5.55</td>
</tr>
<tr>
<td>vPython for</td>
<td>2.33</td>
<td>1.92</td>
</tr>
<tr>
<td>Native C</td>
<td>1.33</td>
<td>1.20</td>
</tr>
</tbody>
</table>

Table 1. Byte Sieve benchmark runtime performance (seconds)

The new version of the Olympus abstract machine for Vipera that separates the object addressing from operation within the mnemonics, not only enables direct access to native C variables, as shown in Listing 1.5, to increase performance but also simplifies the implementation of object references within the abstract machine, enabling the integration of Olympus applications with C frameworks, such as the Eithne benchmarking framework[20] and MPI (Message Passing Interface).

4 Conclusion

Whilst the vPython virtual machine provided a productive environment to deploy parallel codes written in a dynamic language to micro-core architectures, the performance overhead of the interpreter limited its use for real-world codes. However, the Olympus abstract machine approach resulted in kernel performance that was comparable to or, in some cases could exceed, native C kernels, as confirmed for the LINPACK benchmark in Section 3.2, and, at a worst-case, was around five times slower than native C for the Sieve of Eratosthenes benchmark (Section 3.4). Crucially, as shown in Section 3.6, this gap can be lowered to just over 1.5 times slower by leveraging the for loop’s native C iterator. Furthermore, a single Python code is portable across these architectures, which is not the case for the standard C codes.

Vipera has also addressed the portability of user codes and underlying runtime support. All of the benchmarks run unmodified across all the supported platforms and the Olympus abstract machine builds from a single codebase,

---

8Standalone versions, not run within the Eithne framework per Section 3.4.
which results in significant programmer productivity gains. All device-specific code is managed within the mnemonics and runtime support functions, with the generated Olympus abstract machine code remaining the same across all platforms. Furthermore, the vPython virtual machine was also shown to be portable to a number of micro-core architectures with the minimum of effort.

Further work includes exploring automatic memory management for data and code, optimisation of the Olympus abstract machine, automatic dynamic function selection for the dynamic loading support discussed in [21], additional data types (byte arrays) to minimise the memory footprint of data and additional device support (GPUs and FPGAs) using OpenCL C and Xilinx HLS C.

Whilst this paper has focused on the assessment of the Olympus code generation model using vPython, we also believe that Vipera has a wider applicability to other dynamic programming languages targeting micro-core architectures.

References

2. index | TIOBE - the software quality company, https://www.tiobe.com/tiobe-index/