Abstract

Six implementations of different lazy functional languages are compared using a common benchmark of a dozen medium-sized programs. The experiments that were carried out two years ago have been repeated to chart progress in the development of these compilers.

The results have been extended to include all three major Haskell compilers. Over the last two years, the Glasgow Haskell compiler has been improved considerably. The other compilers have also been improved, but to a lesser extent. The Yale Haskell compiler is slower than the Glasgow and Chalmers Haskell compilers. The compilation speed of the Clean compiler is still unrivalled.

Another extension is a comparison of results on different architectures so as to look at architectural influences on the benchmarking procedure. A high-end architecture should be avoided for benchmarking activities, as its behaviour is uneven. It is better to use a midrange machine if possible.

1 Introduction

In the previous benchmark paper [10], which will be referred to as the benchmark-I paper, five compilers for lazy functional languages were benchmarked. It is interesting to see what progress has been made in the development of these compilers during the past two years. The benchmark procedure has been extended to take into account the third major Haskell compiler, from Yale.

The benchmark has been executed on different architectures to look at the influence of architectural parameters on the reliability of the results. Previously, certain anomalies could only be explained by referring to the complicated architecture on which the experiments were performed. The inclusion of results for different architectures make it possible to firm up some conclusions about the results.

To chart purely the progress made in the various compiler developments the benchmarking procedure has not been altered, only extended. The benchmark-I results are therefore directly comparable to the benchmark-II results.

The next section summarises the main points from the benchmark-I paper. Section 3 describes an extension to the benchmarking procedure. Section 4 presents the new results obtained. Section 5 compares the results obtained using different architectures. The last section gives the conclusions.

2 A summary of the method

In the benchmark-I paper, four languages and five implementations were used. These were Clean (version 0.8) [9] and its compiler from Nijmegen [15]; a subset of Miranda\(^1\) [8] and our own FAST/FCG compiler [9, 13]; Lazy ML [2] and its compiler from Chalmers [3] and finally Haskell [1] with the compilers from Chalmers [1] and Glasgow [17]. In this paper the third major Haskell compiler from Yale [7] is also included. A summary of the offerings of the four languages is presented in the benchmark-I paper.

A new version of Clean (version 1.0) is currently under development. The results presented here apply to version 0.8 of Clean. It is hoped that the new Clean results may be included in a future version of the benchmark.

Benchmarking different languages requires a common benchmark program to be written in different syntaxes and using different primitive functions. This is possible, as we have demonstrated with the single program used in the Pseudoknot benchmark [11]. The effort involved in translating by hand a larger set of programs would be considerable. More importantly, it is difficult to do so fairly. It is easy to make mistakes and to introduce bias towards one of the languages. In particular it is difficult to avoid bias towards the language in which the benchmarks were originally written. The solution adopted here is to write all the benchmarks in one language, whilst taking into account two important considerations to reconcile the languages.

Firstly, we restrict the use of special syntax to those forms that can be translated mechanically into a functionally equivalent, and equally efficient syntax, in all the languages with which we are concerned. The main problem here is that the Haskell compilers need to resolve operator overloading, whereas the other compilers do not. Compile and run time resolution of operator overloading has consistently been avoided by mechanically annotating each operator with the exact type of its operands. A minor problem is that some compilers (Clean and FAST) support comparison operators only on numeric data. In the benchmarks, such operators are therefore only applied to numeric data, with explicit annotation of the exact types of the operands.

Secondly, we use a set of essential primitive functions that must be present in any language for it to be of interest.

---

\(^1\)Miranda is a trademark of Research Software Ltd.
for general purpose programming purposes. The first problem here is that some systems offer single precision floating point numbers (FAST and Yale Haskell) and others offer double precision floats. This difference could not be ironed out, so the results for essentially floating point programs will have to be interpreted with care. The second problem is that all compilers except Clean offer arrays. An array implementation based on binary trees has been provided for Clean. The third problem is that complex numbers use lazy data constructors in Haskell, but strict constructors in the other systems. This affects one of the 14 benchmark programs only (the fft program).

The benchmark programs are written in Miranda and translated to LML, Haskell or Clean by the FAST compiler, which is also capable of autonomous compilation. The generated LML, Haskell and Clean code is further processed by the appropriate compiler. This procedure is shown in Figure 1.

The benchmark set contains small and medium-sized programs, each of which is run on a realistic input data set. Refer to the benchmark-I paper for a summary description of the individual programs.

The benchmark programs were all compiled while optimising for execution speed. The compile time and run time options used are shown in Table 1 together with the characteristics of the two platforms used.

Most compilers generate stand-alone executables, the exception being the Yale Haskell compiler. This system is embedded in a Lisp system (CMU Common Lisp). As part of the compilation process a compiled Lisp image is generated. This image can be executed under control of the CMU common Lisp system and the entire process can be timed. From the total time taken, the time necessary just to start the Common Lisp system (9 seconds on the SUN 4/690) are subtracted. This brings the Yale Haskell timing as close as possible to the timing of the stand-alone binaries generated by the remaining compilers.

Execution time measurements of the stand-alone executables (or the Lisp image in the case of Yale Haskell) were taken using the UNIX /bin/time command, taking the sum of user and system time as the total execution time. Each executable has been run a large number of times, on a quiet

---

**Figure 1:** The organisation of the benchmarking procedure, showing how the benchmarks are translated into LML, Haskell and Clean.

**Table 1:** The computer system used and the compiler and run time options. The versions used previously and the current versions are shown as “benchmark-I” → “benchmark-II”.

---

**Systems used:**

- **SUN 4/690**: a dual processor machine with 64M memory and 64K cache (1 way associative) under SunOS 4.1.2.
- **SPARC 10/41**: a single processor machine with 128M of memory, 20K primary instruction cache (4 way associative), 16K primary data cache (5 way associative) and 1M secondary cache under Solaris 2.3

In the (Bourne) shell commands below, the variable $P represents the name of the program, $I is the input parameter (Only used at command level by the FAST/FCG system). The heap sizes shown all correspond to a maximum of 48M space.

**Clean version 0.8.2 → 0.8.4:**

```bash
fast -clean $P.i
clm $P -o $P.clean.out
$P.clean.out -b -h 48m -nt -s 8m
```

**FAST/FCG version 31 → 34:**

```bash
fast -fcg $P.i
$P.fcg.out -v 1 -h 12000000 -s 2000000 $I
```

**LML version 0.999.4 → 0.999.7:**

```bash
fast -lml $P.i
lmlc -H20000000 -0 $P.lml.m fast2lml.o -o $P.lml.out
$P.lml.out -H48000000 -A1000000 -V1000000
```

**Chalmers Haskell version 0.999.4 → 0.999.7:**

```bash
fast -hbc $P.i
hbc -H40000000 -0 -ihbclib: $P.hbc.hs -o $P.hbc.out
$P.hbc.out -H48000000 -A1000000 -V1000000
```

**Glasgow Haskell version 0.11 → 0.22:**

```bash
fast -ghc $P.i
ghc -Rmax-heapsize 60M -Rmax-stksize 2M -0 \ -Lghclib -Lghclib -lfast $P.ghc.hs \ -syslib hbc -o $P.ghc.out
$P.ghc.out +RTS -H48M -KEM
```

**Yale Haskell version → 2.3b** (with CMUCL 17e):

```bash
fast -yale $P.i
cmu-haskell
    :compile $P
cmu-haskell
    :((setf ext:*bytes-consed-between-gcs* 24000000)
    :load $P
    :((ext:gc)
    :run $P
```

---

2
system, taking the best execution time as the ultimate performance measure, because it minimises the error in the time measurement. The real times were found to be at most 1 second higher than the execution time in over 83% of all measurements. This implies that the reported times are a reliable indication of program performance, in the sense that the machines used were not otherwise loaded.

The measured execution times include garbage collection time. Each stand-alone executable has been run 10 times with each of six different heap sizes: 8, 16, 24, 32, 40 and 48M. The Yale Haskell system has been run likewise, while setting the bytes-consed-between-gcs parameter to 4, 8, 12, 16, 20 and 24M. The setting of this parameter causes the garbage collector to be run after the specified amount of space has been allocated. As the CMU Common Lisp system uses a two-space garbage collector, this setting should correspond closely to the maximum heap sizes of 8...48M for the other systems.

In all cases, the selected parameters allow a two-space copying garbage collector two semi-spaces of 4, 8...24M each. One of the heap sizes will give the best execution time. A small heap means slow execution because of a large number of garbage collections. In this case the user time will be high. Having a large heap slows execution down because of excessive paging activity. This will result in a high system time. As the sum of the system and user time is taken as the total execution time, there will be an optimum in the execution time. The benchmarking procedure is designed to find this optimum.

All implementations were benchmarked using the default garbage collector. For the Glasgow Haskell compiler, this is a generational garbage collector; the other systems use a two-space copying garbage collector.

3 Optimising prelude functions

A bone of contention with the benchmark-I method has been the use of the prelude functions such as map, filter, foldr etc. These functions come with the benchmark programs, but are also available in the standard prelude provided by most of the compilers. These often-used prelude functions can be optimised. In particular, the Yale Haskell compiler uses special versions of the prelude functions that are amenable to the foldr/build deforestation optimisation. None of the other compilers provide this optimisation (yet). To gauge the effectiveness of this (and possibly other optimisations) the Haskell programs have been compiled and run in two versions: using the built-in Haskell prelude functions and using the prelude functions supplied with the original Miranda versions of the code. This has no significant influence on the execution time measured for the Glasgow and Chalmers Haskell compilers.

As one might expect, the Yale Haskell compiler produces faster code when using the Yale Haskell prelude functions. However, the improvements are slight with three exceptions: wang runs 20% faster and listcopy and listcompr run approximately 40% faster using the Yale Haskell prelude. These results tally with the findings reported by Gill and Peyton Jones using the Glasgow Haskell system. They report [6] that none of the nifi benchmark [16] programs show a significant change, when run with or without a form of deforestation similar to that used in the Yale Haskell system [14]. It is not sure whether the performance improvements due to deforestation are sustained when the overall performance of the Yale Haskell system is brought in line with its competitors. Compiler optimisations can not be studied properly in isolation, as the many points of interaction may cause the performance of the system as a whole to deteriorate as well as improve [1].

In the results that follow, the Haskell programs use the built-in Haskell prelude functions. The other compilers use the translated Miranda versions. This introduces a small incompatibility between the benchmark-I and the benchmark-II results. The reason for making this particular choice is that future developments must lead to more successful optimisations if the Haskell preludes are used. This would make it necessary to switch over to Haskell preludes at a future date.

4 New results

Table 2 shows the new compile time and run time performance measurements. The compilation speed is reported in lines per minute of real time, where the number of lines in the original Miranda program determines the size of the program. For each compiler the minimum and maximum compilation speed is reported, as found over the whole range of benchmarks. The compilation speed of the various implementations has barely changed over the past two years. The compilation speed of the Clean system is still unrivalled.

The execution time for each compiled program is measured in seconds. For each stand-alone executable the best time out of 10 x 6 = 60 runs is reported. The heap size required to achieve this optimum is shown next to the execution time. Fixing the heap size to the same value for all experiments shows somewhat larger execution times, but the relative ranking of the compilers does not change.

Each row in Table 2 bears at least one star. The stars mark the best result for that particular row. This shows that it depends to some extent on the application which compiler generates the fastest code. The stars awarded for benchmark-I are shown here for ease of reference as dots. Clean, FAST/FCG and LML produced the fastest code for benchmark-I. For benchmark-II the fastest code is generated by the Glasgow Haskell compiler. This compiler has been improved considerably. The major improvements are the provision of a strictness analyser which was previously lacking and a completely rewritten simplifier. Combined with a careful tuning of the system, a considerable overall improvement is the result.

The Clean compiler and the two Chalmers compilers have been improved over the whole range of programs, but not significantly so. No improvements have been made to the FAST/FCG compiler, so a stand-still of two years is sufficient to be denoted from the top of the league.

We have it on good authority (but alas undocumented) that on larger programs than the ones used here, the Chalmers and Glasgow Haskell compilers are closer than would appear from the measurements presented here.

At first there appears to be various remarkable differences in the amount of space used to achieve fast execution. Compare the SPARC 10/41 performance for listcopy as delivered by the Chalmers and Glasgow Haskell systems. The former uses only 8M; the latter requires 48M. Another program, fft, shows quite the opposite behaviour. Figure 2 shows the best execution times of listcopy, with each of the six heap sizes used, for both Chalmers and Glasgow Haskell. There is little variation in the execution times and it just so happens that the slopes of the two curves are in oppo-
<table>
<thead>
<tr>
<th>language compiler</th>
<th>input</th>
<th>Clean</th>
<th>FAST/FCG</th>
<th>LML</th>
<th>Haskell</th>
</tr>
</thead>
<tbody>
<tr>
<td>minimum</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>maximum</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Compilation speed in lines per minute real time, SUN 4/690

Heap space in Mb byte and execution time in seconds, SUN 4/690

(a) The new compilation and execution results using the SUN 4/690

<table>
<thead>
<tr>
<th>language compiler</th>
<th>input</th>
<th>Clean</th>
<th>FAST/FCG</th>
<th>LML</th>
<th>Haskell</th>
</tr>
</thead>
<tbody>
<tr>
<td>minimum</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>maximum</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Heap space in Mb byte and execution time in seconds, SPARC 10/41

(b) The new execution results using the SPARC 10/41.

Table 2: The compile time and run time results of the benchmarking (* = Best execution time for Benchmark-II; = Best execution time for Benchmark-I; ∞ = Stack overflow in CMU Common Lisp; - = results not available yet)
The amounts of heap space reported thus do not imply that one compiler requires significantly more heap space than another to achieve fast execution. The exception to this rule is the Yale Haskell compiler which runs between 1.3 and 2.5 times slower with the smallest heap size of 8M.

Some of the more specific remarks that were made in the benchmark-I paper need to be reconsidered.

It is no longer the case that the Haskell compilers from Chalmers and Glasgow are less mature than the other, older compilers. Work is in progress to improve the Yale Haskell performance. Better results will hopefully appear in a future version of this paper.

The implementation of arrays should generally be improved. This is demonstrated by the relatively slow execution of the code generated by the all but the FAST/FCG compiler on the (array intensive) wave program.

All systems provide floating point support at comparable levels (programs that use floating point arithmetic are wang, fft, wave and solid). The results do not indicate that single precision (FAST/FCG and Yale Haskell) or double precision (all other compilers) floating point support has an effect on these programs.

Both the FAST compiler and the Glasgow Haskell compiler generate C programs; the Yale Haskell compiler generates Lisp code and the remaining compilers generate assembler programs. Going via C is not a disadvantage. Going via Lisp apparently does not offer an easy route to high performance. It should be pointed out that embedding the system in Lisp enables the Yale Haskell programming environment to be more elaborate than that offered by the other compilers.

To give a rough indication of the progress that has been made in constructing lazy functional compilers over the last two years, consider the task of running all 14 the programs through the five benchmark-I compilers. The execution times for benchmark-I add up to 2969 seconds, those for benchmark-II add up to 2586 seconds. This implies a progress of 17%. Most of this is due to the Glaswegian efforts.

### 5 Comparing different architectures

Some performance differences that were observed in the benchmark-I project could not be explained properly. To investigate this further, all stand-alone binaries have been run on a SUN 4/690 (the same machine as for benchmark-I, although over the last two years the operating system has undergone various updates) as well as on a second, rather different machine: a SPARC 10/41 (c.f. Table 1). The SUN 4/690 that has been used is a high-end model. The SPARC 10/41 is a more modern machine, but is a midrange model. Nevertheless, the SPARC 10/41 is between 1.5 and 2 times as fast as the SUN 4/690 on the benchmark set. A highly optimised, high-end machine responds in a less even fashion to varying loads than an average, midrange machine. Comparing the timings obtained on these different machines, using the same binaries presents an opportunity to look at the architectural and system differences.

Consider the execution times for the programs listcopy and listcomp as measured for the Clean and Glasgow Haskell compilers. These programs are virtually the same, the difference being that listcopy makes an extra copy of the (rather long) output list (see the benchmark-I paper for details). Such a small amount of extra work should only make it slightly more expensive to compute the full answer. The difference between 23 and 30 seconds (Clean results, benchmark-I) can therefore only be blamed on the uneven response of the architecture to slightly different loads. The new Clean measurements on the SUN 4/690 are close, as expected (20 and 21 seconds respectively). The same problem is now manifest with the Glasgow Haskell compiler, which in the benchmark-I results showed an acceptably small difference (33 versus 36 seconds), but now 34 versus 41 seconds. Looking at the results obtained with the SPARC 10/41 shows that both Clean and Glasgow Haskell yield similar performance on the two programs: 10 versus 11 seconds for Clean and 20 versus 22 seconds for Glasgow Haskell. The two programs listcopy and listcomp were also run on seven other systems, with various different characteristics as shown in Table 3. Only the SUN 4/690 showed anomalous behaviour.

The two Chalmers compilers and the FAST/FCG compiler use an optimisation, which definitely should make the execution times of listcomp and listcopy come out rather differently (see benchmark-I for details). This observed uneven behaviour of the SUN 4/690 does not invalidate the results that have been reported. The unevenness of the architecture may affect any of the 14 x 5 experiments (14 x 6 for benchmark-II). This observation is confirmed by the awarding of the stars to the various programs and compilers. The SUN 4/690 awards (Table 2-a) are slightly more in favour of Glasgow Haskell than the SPARC 10/41 awards, (Table 2-b) but the difference is small.

<table>
<thead>
<tr>
<th>SUN</th>
<th>SunOS</th>
<th>Memory</th>
<th>Cache (I/D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4/40</td>
<td>4.11</td>
<td>48M</td>
<td>64K</td>
</tr>
<tr>
<td>4/65</td>
<td>4.11</td>
<td>64M</td>
<td>64K</td>
</tr>
<tr>
<td>4/75</td>
<td>4.11</td>
<td>64M</td>
<td>64K</td>
</tr>
<tr>
<td>4/50</td>
<td>4.11</td>
<td>48M</td>
<td>64K</td>
</tr>
<tr>
<td>4/370</td>
<td>4.11</td>
<td>64M</td>
<td>64K</td>
</tr>
<tr>
<td>4/690</td>
<td>4.12</td>
<td>64M</td>
<td>64K</td>
</tr>
<tr>
<td>SS10/30</td>
<td>4.13</td>
<td>32M</td>
<td>1M+20/16K</td>
</tr>
<tr>
<td>SS10/41</td>
<td>5.3</td>
<td>128M</td>
<td>1M+20/16K</td>
</tr>
<tr>
<td>SS20/51</td>
<td>5.3</td>
<td>32M</td>
<td>1M+20/16K</td>
</tr>
</tbody>
</table>

Table 3: Nine different SUN systems used to benchmark listcopy and listcomp.
Six compilers for lazy functional languages have been benchmarked using a set of 14 medium-sized programs.

Two years of development can make a significant difference to the quality of a compiler. Implementors who want to keep up with the developments are not wise to lapse into such a long period of inactivity. The Glasgow Haskell compiler now generates the fastest code of the six compilers. There is a small gap between this compiler and four of the others. The performance of the Yale Haskell compiler is significantly worse than that of the other two Haskell compilers, but improvements will be made shortly.

Deforestation seems to be an interesting compiler optimisation, but does not fulfill its promise at this moment. Some evidence has been found that deforestation works for the other two Haskell compilers, but improvements will be made shortly.

The benchmarks are available from ftp.fwi.uva.nl, file pub/functional/packages/benchmark.tar.Z.

Acknowledgements

The help of Lennart Augustsson, John van Groningen, Sandra Loosemore, Jim Mattson, Will Partain and John Peterson with the benchmarking of the Haskell, LML and Clean compilers was greatly appreciated. Marcel Beemster, Hugh McEvoy, Jon Mountjoy, Henk Muller and Wim Vree have helped to improve the paper.

References


