



**HAL**  
open science

## A framework for automatic and parameterizable memoization

Loïc Besnard, Pedro Pinto, Imane Lasri, João Bispo, Erven Rohou, João  
Manuel Paiva Cardoso

► **To cite this version:**

Loïc Besnard, Pedro Pinto, Imane Lasri, João Bispo, Erven Rohou, et al.. A framework for automatic and parameterizable memoization. *SoftwareX*, 2019, 10, pp.100322. 10.1016/j.softx.2019.100322 . hal-02305415

**HAL Id: hal-02305415**

**<https://inria.hal.science/hal-02305415>**

Submitted on 4 Oct 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# A Framework for Automatic and Parameterizable Memoization

Loïc Besnard<sup>a</sup>, Pedro Pinto<sup>b</sup>, Imane Lasri<sup>a</sup>, João Bispo<sup>b</sup>, Erven Rohou<sup>a</sup>,  
João M. P. Cardoso<sup>b</sup>

<sup>a</sup>*Univ Rennes, Inria, CNRS, IRISA Campus de Beaulieu, 35042 Rennes France. Email: loic.besnard@irisa.fr, {imane.lasri, erven.rohou}@inria.fr*

<sup>b</sup>*Faculty of Engineering, University of Porto, Porto, Portugal. Email: {p.pinto, jbispo, jmpc}@fe.up.pt*

---

## Abstract

Improving execution time and energy efficiency is needed for many applications and usually requires sophisticated code transformations and compiler optimizations. One of the optimization techniques is memoization, which saves the results of computations so that future computations with the same inputs can be avoided. In this article we present a framework that automatically applies memoization techniques to C/C++ applications. The framework is based on automatic code transformations using a source-to-source compiler and on a memoization library. With the framework users can select functions to memoize as long as they obey to certain restrictions imposed by our current memoization library. We show the use of the framework and associated memoization technique and the impact on reducing the execution time and energy consumption of four representative benchmarks.

*Keywords:* memoization, compiler optimizations, source-to-source

---

## 1. Motivation and significance

2 Improving execution time and energy efficiency is of paramount impor-  
3 tance in many applications. A possible optimization that can be used for such  
4 improvements is memoization [1], which is the technique of saving the results  
5 of computations so that future computations can be skipped. Performance  
6 can be improved by caching execution results of memoizable functions, and  
7 retrieving them instead of recomputing a result when a new call is performed  
8 with repeated inputs. In this work we introduce a memoization framework,  
9 integrated in the ANTAREX tool flow [2], which can automatically apply  
10 this technique to memoizable functions.

11 We define memoizable functions [3, 4] as pure functions, i.e., determinis-  
12 tic and without side effects, with three additional constraints imposed by our  
13 current implementation of memoization. First, functions must have between  
14 1 and 3 parameters of the same type  $T$ . This limitation arises from our cur-  
15 rent choice of hash function, which combines all the arguments and, as such,  
16 requires them to have the same bit length. After combining all arguments  
17 with *XOR* operations, we take the high-order and the low-order bits of the  
18 result and combine them again using a *XOR*, leaving us with half the original  
19 bits. We repeat this process until we arrive to the number of bits required to  
20 index the hash table, masking any bits if necessary. The second constraint  
21 is that functions must return data of the same type  $T$ . Finally, the third  
22 constraint is that  $T$  is one of `int`, `float` or `double`.

23 In applications with memoizable functions in critical code sections, mem-  
24 oization may provide important execution time reductions and energy con-  
25 sumption savings. Our memoization framework relies on internal tables that  
26 store the results of previous computations and replace future calls by ta-  
27 ble lookups. The elements of the internal tables are indexed with a hash  
28 calculated from the call arguments of the memoized function.

29 A memoization approach can start by profiling the application and by  
30 identifying the contribution of memoizable functions to the overall execu-  
31 tion of the application. A profiling step may also provide values to setup  
32 an initial version of the internal table of the memoization technique. Our  
33 framework allows loading internal tables before application runs (e.g., with  
34 profiling data) and/or update them at runtime (allowing adaptation to ex-  
35 ecution contexts not considered during the profiling phase). This solution  
36 may enable savings in multiple kinds of applications and allow users to write  
37 runtime adaptivity strategies to make applications more resilient to context  
38 changes and able to achieve predetermined execution thresholds.

39 The memoization technique can be integrated into any C or C++ ap-  
40 plication that has memoizable functions or methods. Based on the selected  
41 memoizable functions, our framework generates a new version of the applica-  
42 tion enhanced with memoization support by relying on the Clava<sup>1</sup> source-to-  
43 source compiler. The framework is also responsible to generate the C library  
44 that contains the core of the memoization implementation, which is linked  
45 with the newly generated application. The compilation process of Clava  
46 is controlled by the LARA [5] Domain-Specific Language (DSL). With this  
47 setup, it is also possible to use LARA to perform analyses and transforma-  
48 tions on the code, e.g., where and how to apply the memoization technique in

---

<sup>1</sup><https://github.com/specs-feup/clava>

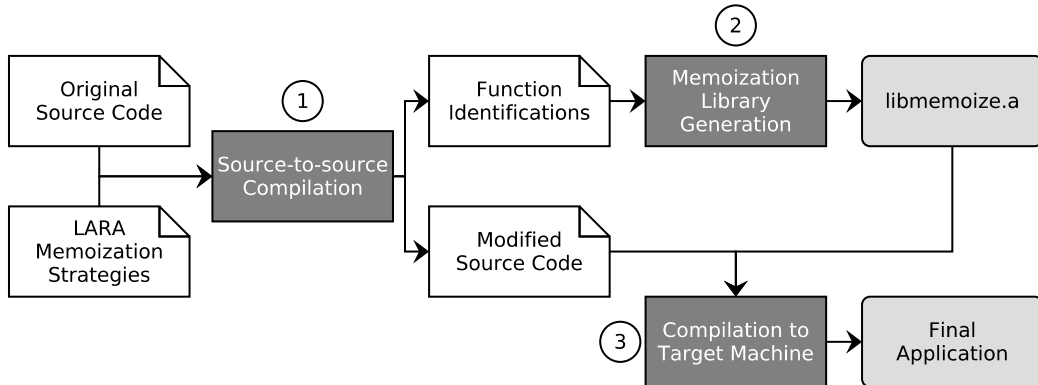


Figure 1: Tool flow of the proposed memoization framework.

49 this specific case. The main modular unit in LARA code is the *aspect*, which  
 50 can be considered as a function, with input and outputs, but that has the  
 51 capability of interacting with the internal representation of the application  
 52 source code. When several aspects are developed to achieve a specific goal,  
 53 we have a LARA program, which we often refer to as a *strategy*.

54 Our memoization framework is based on the memoization approach pre-  
 55 sented in [4], that uses a dynamic library and applies memoization at binary  
 56 level. Our framework has been generalized for C++, adds extensions that  
 57 allow application flexibility, adds options regarding table loading and table  
 58 runtime updating, and is applied at the level of the application source code.

## 59 2. Software description

60 The software solution presented in this paper relies on two different com-  
 61 ponents which are combined to easily enhance an application with memo-  
 62 ization support. The first component is Clava, a source-to-source compiler  
 63 able to automatically generate a new version of the application code with  
 64 the necessary changes to support memoization. This is performed through a  
 65 series of LARA aspects the user can call and parameterize, all distributed as  
 66 part of the Clava memoization library.

67 The second component is the memoization C library itself, which is gener-  
 68 ated from a configuration file that identifies the functions on which to apply  
 69 memoization. The library is compiled and linked against the version of the  
 70 input application that has been modified by Clava. As shown in Figure 1, the  
 71 use of this memoization solution consists of three steps: 1) source-to-source  
 72 compilation; 2) generation and compilation of the memoization library; and  
 73 3) compilation of the application linked with library.

---

```

1 float foo (float p)
2 {
3     /* code of foo without side effects */
4 }
5
6 float foo_wrapper(float p)
7 {
8     float r;
9
10    /* already in the table ? */
11    if (lookup_table(p, &r)) return r;
12
13    /* calling the original function */
14    r = foo(p);
15
16    /* updating the table or not */
17    update_table(p, r);
18
19    return r;
20 }

```

---

Figure 2: A memoizable C function and its wrapper.

74 Consider the memoizable C function `foo` shown in Figure 2. This is a pure  
75 function which takes a single `float` parameter and returns data of the same  
76 type. The source-to-source step consists of two phases, with the first phase  
77 being the addition of another function to the program. This new function  
78 *wraps* the target function and includes the memoization logic to interface to  
79 the memoization library. This wrapper is also illustrated in Figure 2. The  
80 second phase replaces all calls (it is possible to specify a LARA aspect that  
81 applies memoization to only certain call sites) to the original function, `foo`,  
82 with calls to its corresponding wrapper, `foo_wrapper`. This technique works  
83 for C and C++ functions as well as for C++ methods, and takes into account  
84 name mangling, function overloading, and references to objects.

85 The user can, through LARA aspects or manually, further change the  
86 application to interact with the memoization library. For instance, the mem-  
87 oization library exposes a set of variables to control its internal behavior. To  
88 dynamically stop and restart the memoization for a specific function, it is pos-  
89 sible to change the variable `_Memoize__<mname>`, where `<mname>` corresponds  
90 to the mangled name of the target function. Similarly, to control the table  
91 update policies the user can change the variables `_alwaysReplace__<mname>`  
92 and `_FullyOffline__<mname>`. This can be used to control the runtime be-  
93 havior of the memoization library with respect to the update of the internal  
94 table.

95 The generation and compilation of the memoization library start from  
96 the function identifications file (`funs-static.def`). The LARA library for

97 memoization, besides changing the source of the application, generates this  
98 file. In this article, we propose the use of memoization through Clava, by  
99 relying on aspects programmed with the LARA DSL. However, it is also  
100 possible to use the standalone library<sup>2</sup> without relying on LARA and the  
101 Clava compiler. To do so, the user must manually write the function iden-  
102 tifications file (`funcs-static.def`), change the source code to include the  
103 wrappers and replace calls from the original calls to the wrappers. The rest  
104 of the flow remains the same: generate the library and link it when compiling  
105 the application.

106 The advantage of using LARA strategies is that the memoization library  
107 is integrated into the application without performing manual modifications  
108 of the source code. The code generated by Clava is then compiled and linked  
109 with the associated generated memoization library.

110 In order to enhance an application with memoization with LARA strate-  
111 gies, users can call the library aspects and define a number of parameters.  
112 First, it is possible to define the size of the internal table that stores the  
113 computation data. This decision becomes a tradeoff that can be fine-tuned  
114 according to the needs of the current application or scenario. A larger ta-  
115 ble holds more data but requires more memory. On top of that, depending  
116 on the underlying architecture, the size of the table may impact cache per-  
117 formance. Second, one can specify a file from which to load a previously  
118 initialized memoization table. If none is specified, the program starts with  
119 an empty table. This allows building memoization tables from one execution  
120 to the following and keep the most frequently requested outputs. For this  
121 feature, we can also specify a file on which the results of the execution are  
122 saved. Finally, if one intends to use the approximation capabilities, it is also  
123 possible to configure the library at this step to disregard a number of bits  
124 from the input representation of the arguments to the memoized function.

### 125 *2.1. Technique Details*

126 The first operation of the wrapper is to perform a lookup to check whether  
127 the output for the provided input is already stored. The lookup operation will  
128 hash the input to get the correct slot in the internal table. If this position is  
129 empty, the lookup fails and returns `false`, signaling a miss. If this happens,  
130 the original target function is called and the correct value is returned. On the  
131 other hand, if the table slot is not empty, we compare the stored input with  
132 the input of the current call. If these values are the same, which we consider  
133 a hit, the lookup function sets the output value and returns successfully. The

---

<sup>2</sup><https://gforge.inria.fr/projects/memoization>

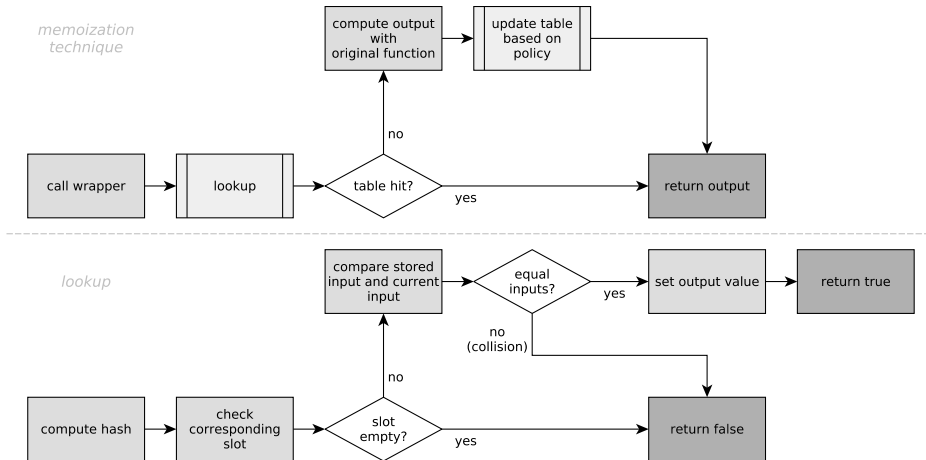


Figure 3: Flowcharts describing the overall memoization technique and the lookup process.

134 wrapper will then return this value. If the inputs are not equal we have  
 135 a collision, meaning that two different inputs were hashed into the same slot.  
 136 In this case, the lookup function returns `false`. Once again, the return value  
 137 is computed by calling the original function, which is eventually returned.  
 138 The correct value is always returned whether there was a hit, a miss or  
 139 a collision. The table update logic happens after the correct output value  
 140 has been computed (or fetched from the table), so it is inconsequential to  
 141 the safeness of the technique. Figure 3 shows two flowcharts describing the  
 142 overall memoization technique (on top) and the lookup process (on bottom).

## 143 2.2. Software Architecture

144 Figure 4 shows an overview of the architecture our memoization frame-  
 145 work. The four main components can be divided into two groups, *generation*  
 146 and *execution*. The generation group, with Clava and the Library Gener-  
 147 ator, is responsible for generating the execution group, composed by the  
 148 augmented application code and the memoization library. The dashed ar-  
 149 rows indicate that one component generates another and the solid arrows  
 150 indicate runtime interactions between components.

151 Clava is a source-to-source compiler, written in Java, that uses Clang<sup>3</sup> as  
 152 a front-end to parse C and C++ source code. The compiler maintains an  
 153 internal representation of the application which is analyzed and transformed  
 154 according to the strategies defined in the input LARA code. Clava includes

<sup>3</sup><https://clang.llvm.org/>

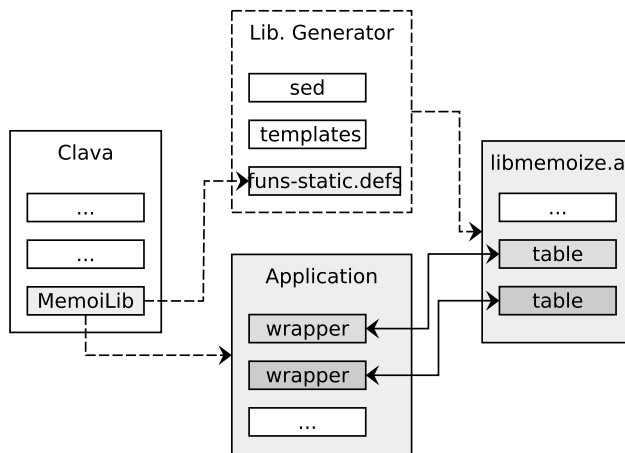


Figure 4: Main components of the proposed memoization solution.

155 multiple LARA libraries that can be used for different purposes, including  
 156 the memoization framework presented in this article.

157 The Library Generator is a `sed` script that takes the function definitions  
 158 file (`*.def`) and several code templates to generate the final customized code  
 159 of the memoization library. This code can then be compiled into a library.  
 160 The library itself maintains a table for each memoized function as well as the  
 161 associated variables for runtime control and the interface functions. These  
 162 are the functions called from each of the wrappers inserted with the Clava  
 163 memoization strategies.

164 The final application retains the original functionality. In the source  
 165 code, the difference is that it has a wrapper for each target function. This  
 166 wrapper is called instead of the original and performs calls to the generated  
 167 memoization library.

### 168 *2.3. Software Functionalities*

169 The memoization framework presented in this paper allows users to semi-  
 170 automatically enhance an application with memoization support. The typical  
 171 user intervention is related to the configuration options and to the selection  
 172 of the functions to memoize. However, function selection can be also per-  
 173 formed without user intervention as the LARA library also provides code  
 174 to automatically detect memoizable functions and methods. This option re-  
 175 turns the list of detected memoizable functions and users can decide whether  
 176 to apply memoization on these returned functions.



---

```

1 import antarex.memoi.Memoization;
2
3 aspectdef Launcher
4     call Memoize_Initialize( );
5     call Memoize_MathFunctions(['cos' , 'acos' , 'sqrt']);
6     call Memoize_Function('myfunc');
7     call Memoize_Finalize( );
8 end

```

---

Figure 5: An example of LARA Launcher aspect defined for memoization.

177 The LARA library integrated into the Clava compiler has several aspects  
178 that are used to apply memoization to different types of functions: mathe-  
179 matical functions (from `math.h`), C functions and C++ functions. For each  
180 of these, there are LARA aspects that allow the user to memoize a function  
181 with default parameters or by defining all the needed parameters. Addition-  
182 ally, the library includes an aspect that automatically performs memoization  
183 on all `math.h` functions, and aspects that apply memoization to C++ func-  
184 tions and methods taking into account function overloading.

185 Since our framework includes source-to-source compilation, it is possible  
186 to write sophisticated analysis and transformation strategies using LARA  
187 aspects. This can be used, for instance, to select which functions to memoize  
188 using any user-devised heuristics on top of the list of possible memoizable  
189 function returned by the library.

190 Another feature of our memoization library is the possibility of using  
191 approximation when dealing with floating point numbers. This essentially  
192 removes a number of (user-specified) bits from the inputs to the function,  
193 grouping together close numbers and returning the same output for those.  
194 More precisely, we remove a number of the least significant bits of the man-  
195 tissa. In scenarios where some loss of accuracy is tolerable (e.g., in some  
196 cases of image processing) this can provide further reductions in execution  
197 time and energy consumption.

198 All of these features enhance an application with the capability of taking  
199 advantage of repeating inputs on some specific and critical functions, while  
200 still being able to accommodate changes, since the table can be dynamically  
201 updated. Furthermore, users can explore runtime strategies to decide when  
202 to change the update policies of the table or completely turn memoization  
203 off, thanks to the variables that expose this type of control for each target  
204 function.

---

```

1 DEF(0, acos, acos_wrapper, 1, double, 0, none, no, none, no, 65536)
2 DEF(0, cos, cos_wrapper, 1, double, 0, none, no, none, no, 65536)
3 DEF(0, sqrt, sqrt_wrapper, 1, double, 0, none, no, none, no, 65536)
4 DEF(2, myfunc, myfunc_wrapper, 3, double, 0, none, no, none, no, 65536)

```

---

Figure 6: An example of `funs-static.def` file.

### 205 3. Illustrative Examples

206 This section presents an example of a LARA aspect that illustrates the  
 207 interface between the user and the memoization library described in the  
 208 previous section.

209 Consider an example of a C application that uses the mathematical func-  
 210 tions `cos`, `acos` and `sqrt`. Moreover, the profiling of the application shows  
 211 that there is a large number of calls to a user function called `myfunc`. Figure 5  
 212 shows a LARA aspect, named *Launcher*, that can be used to apply memoiza-  
 213 tion. First, we import the class that contains the memoization aspects (line  
 214 1). Then, the functions to be memoized are identified (lines 5-6), preceded  
 215 by an initialization (line 4) and followed by a finalization stage (line 7). This  
 216 is all the LARA code user needs to write to enhance a C/C++ application  
 217 with memoization.

218 The execution of the aspect will produce (1) a new version of the applica-  
 219 tion in which all the references to FOO (FOO is one of `cos`, `acos`, `sqrt`,  
 220 `myFunc`) are replaced by FOO\_wrapper, and (2) the `funs-static.def` file  
 221 that contains a line per memoized function as shown in Figure 6. Each line  
 222 encodes all the required parameters for the generation of the memoization C  
 223 library.

224 The `Memoize_Function` aspect is a helper aspect that automatically sets  
 225 some default parameters when calling other, more general aspects. For in-  
 226 stance, line 4 of Figure 6 exposes some of the parameters that can be specified  
 227 on those more general aspects. It specifies that it is a C user function (value  
 228 2), followed by the name of the function (`myfunc`) and its associated wrap-  
 229 per (`myfunc_wrapper`), and the identification that this function has 3 inputs  
 230 arguments of `double` type. Other parameters are provided to control the  
 231 memoization framework: in particular the 2 last ones (`no`, `65536`) are used  
 232 to specify the policy to manage possible conflicts when updating the table  
 233 (replace or not) and to define the size of the internal table. LARA aspects are  
 234 provided to the user to set these parameters, but in the presented example  
 235 we call simpler versions that use default values.

236 For example, Figure 7 shows an example of the implementation of one of  
 237 the LARA aspects that are part of the memoization library. This is library

---

```

1 aspectdef Memoize_Method_overloading_ARGS
2 input
3   aClass,      // Name of a class
4   aMethod,     // Name of a method of the class aClass
5   pType,      // Name of the selected type
6   nbArgs,     // Number of parameters of the method
7   fileToLoad, // filename for init of the table, or 'none'
8   FullOffLine,// yes/no. yes for a fully offline strategy
9   FileToSave, // filename to save the table, or 'none'
10  Replace,     // yes/no. yes: always replace in case of collisions
11  approx,     // Number of bits to delete for approximation.
12  tsize      // Size of the internal table.
13 end
14 // Control on the parameters of the aspect: nbArgs in [1,3]
15 ...
16 // Searching the method.
17 var MethodToMemoize, found=false;
18 select class{aClass}.method{aMethod} end
19 apply
20 if (! found) {
21   found = isTheSelectedMethod($method, nbArgs, pType);
22   if (found) MethodToMemoize=$method;
23 }
24 end
25 if (!found)
26 { /* message to the user */}
27 else {
28   GenCode_CPP_Memoization(aClass, aMethod, pType, nbArgs,
29   fileToLoad, FullOffLine, FileToSave, Replace, approx, tsize);
30   call CPP_UpdateCallMemoization(aClass, aMethod, pType, nbArgs);
31 }
32 end

```

---

Figure 7: An example of LARA aspect defined for C++ memoization.

238 code and not the code a user needs to write. This is used to target C++  
239 methods and allows the user to specify all memoization parameters. This  
240 example defines the memoization (lines 1-13) of a C++ method (`aMethod`)  
241 of a class (`aClass`) with `nbArg` parameters of the same type as the returned  
242 type (`pType`).

243 After some verifications, which are not presented in this example, the  
244 target method is searched in lines 17–24. In case of success, the code of the  
245 wrapper is added (line 28) to produce the memoization library, and the code  
246 of the application is modified for calling the generated wrapper (line 30),  
247 which is also declared as a new method of the class.

## 248 4. Impact

249 In order to show the impact of our memoization framework we have ap-  
250 plied it to four C/C++ programs. For each program we have tested the

251 original version as well as four different memoized versions representing a  
252 combination of two parameters. The first parameter is the internal table  
253 size and we tested with tables holding 256 and 65,536 elements. The second  
254 parameter is whether to perform table updates in case of hash collisions. For  
255 this experiments we do not consider the setup of tables prior to the exe-  
256 cution of the application. Thus, when the memoization-enhanced program  
257 starts executing, its tables are empty and start being filled as requests are  
258 performed. It is possible that two different inputs for the same function pro-  
259 duce the same hash value, which leads to a collision. Our strategy in this  
260 example is to either ignore the collision and not change the table, or replace  
261 the previous entry with the newest one.

262 Our benchmark applications are:

- 263 • **atmi** [6] is a library for modeling steady-state and time-varying tem-  
264 perature in microprocessors, with examples using `j0`, `j1`, `exp` and `sqrt`.
- 265 • **equake** is an application extracted from SPEC OMP. It has calls to  
266 the mathematical functions `sin`, `cos` and `sqrt` in its critical region.
- 267 • **fft** is a Fast Fourier transform implementation extracted from the  
268 BenchFFT<sup>4</sup> benchmark suite. It calls the functions `sin` and `cos`.
- 269 • **rgb2hsi** is a benchmarking kernel that converts images from RGB  
270 model to HSI model. It calls `cos`, `acos`, `sqrt`, and a pure user function.

271 The tests have been performed on an Intel(R) Core(TM) i7-5600U CPU  
272 @ 2.60GHz and the C/C++ codes were compiled using GCC with -O3.

273 To address possible concerns that this technique may be too dependent  
274 on which inputs it is tested on, we have used several different inputs for  
275 the **fft** and **rgb2hsi** benchmarks. For **fft**, we tried three numbers of sam-  
276 ples (100000000, 200000000, 300000000) and three frequencies (123.6, 256.89,  
277 88.7). For **rgb2hsi**, we used eight different images. The other benchmarks  
278 were tested with a single input due to the lack of available workloads.

279 Table 1 presents the speedups obtained with the four tested memoization  
280 configurations, over the original version (i.e., without memoization). Table 2  
281 presents, for the same configurations, the energy consumption improvements  
282 over the original. This was measured using a utility tool that relies on Intel  
283 RAPL counters. Values below 1 represent a reduction in energy consumption.  
284 For instance, for the **atmi** benchmark with a *65536-nu* configuration, we only  
285 consume 71% of the original (0.71 in the table).

---

<sup>4</sup><http://www.fftw.org/benchfft>

286 Overall, the use of our memoization framework allows us to achieve con-  
287 siderable reductions in both execution time and energy consumption. This  
288 is confirmed with the *Mean* line, which presents the geometric mean across  
289 benchmarks of the results obtained with each tested configuration. Consider-  
290 ing execution time, the worst average result is a  $0.97\times$  slowdown when using  
291 a small table and no update (in the case of collisions). On the other hand, a  
292 larger table with updates enabled, achieves an average speedup of  $1.16\times$ .

293 With respect to the **atmi** and **rgb2hsi** benchmarks, a larger table has  
294 a positive impact, while this does not happen in the other two benchmarks.  
295 **atmi** and **rgb2hsi** show only a small number of hits (i.e., results found on  
296 the table) on the 256 table, which grows considerably when the table size  
297 is increased, while on the **equake** and **fft** benchmarks this growth is not as  
298 large. This can happen because of the way inputs to the memoized functions  
299 change over time and how many different values these inputs take.

300 We note that enabling update in the case of hash collision improves the  
301 obtained results. This is the case with every benchmark tested, except with  
302 **atmi** in the case of using a large table. When updating is disabled, it is pos-  
303 sible to initially fill the table with inputs that are among the least frequently  
304 requested during the rest of the execution. This is somewhat corrected by  
305 turning on the ability to update in case of collision, since most frequent in-  
306 puts should be able to push out other, less-frequent, inputs and remain in  
307 the table for most of the execution. However, in the specific case of **atmi**,  
308 the Bessel functions show an increasing number of collisions when updating  
309 is enabled, which seems to indicate that two frequent inputs are colliding and  
310 continually pushing each other off the table, causing misses and forcing the  
311 recalculation of their corresponding outputs.

312 The **fft** benchmark shows a performance degradation when using memo-  
313 ization without updates, regardless of the table size. When looking at more  
314 detailed execution data (not presented here), it is possible to see that there  
315 is a single hit for **cos** (out of 200,000,000 calls) and 0 hits for **sin** (out of  
316 100,000,000 calls). This means the tables are initially filled with inputs that  
317 are not requested again in the future, causing an overhead which slightly  
318 degrades the original performance. On the other hand, when the updates  
319 are enabled, these initial values are eventually overwritten by more frequent  
320 inputs. The number of hits increases to 137,227,265 and 74,893,090 for **cos**  
321 and **sin**, respectively.

322 The **rgb2hsi** benchmark shows a severe drop in performance when using  
323 a small table without update on collisions. This benchmark benefits greatly  
324 from both a larger table and updates to remove less frequent values from the  
325 table. The number of hits increases by over  $6\times$  by simply using the larger  
326 table, in the case of no update. This may happen when inputs are distributed

Table 1: Speedups of the memoization versions over the original version. Each of the four columns corresponds to memoized benchmarks with two different parameters, table size and collision update policy (*nu* for no update, and *u* for update). The last line of the table presents the geometric mean of the speedups per configuration.

| Benchmark   | 256-nu          | 256-u           | 65536-nu        | 65536-u         |
|-------------|-----------------|-----------------|-----------------|-----------------|
| atmi        | 1.01            | 1.02            | 1.46            | 1.26            |
| equake      | 1.02            | 1.06            | 1.01            | 1.04            |
| fft         | $0.98 \pm 0.01$ | $1.13 \pm 0.02$ | $0.98 \pm 0.02$ | $1.12 \pm 0.02$ |
| rgb2hsi     | $0.89 \pm 0.02$ | $1.13 \pm 0.10$ | $1.18 \pm 0.06$ | $1.22 \pm 0.08$ |
| <i>mean</i> | $0.97 \pm 0.06$ | $1.08 \pm 0.05$ | $1.14 \pm 0.22$ | $1.16 \pm 0.10$ |

Table 2: Energy consumption improvements of the memoization versions over the original version. Each of the four columns corresponds to memoized benchmarks with two different parameters, table size and collision update policy (*nu* for no update, and *u* for update). Values below 1 represent a reduction in energy consumption. The last line of the table presents the geometric mean of the improvements per configuration.

| Benchmark   | 256-nu          | 256-u           | 65536-nu        | 65536-u         |
|-------------|-----------------|-----------------|-----------------|-----------------|
| atmi        | 1.02            | 1.01            | 0.71            | 0.83            |
| equake      | 0.96            | 0.93            | 0.95            | 0.94            |
| fft         | $1.03 \pm 0.03$ | $0.89 \pm 0.03$ | $1.04 \pm 0.03$ | $0.90 \pm 0.03$ |
| rgb2hsi     | $1.20 \pm 0.03$ | $0.95 \pm 0.08$ | $0.88 \pm 0.04$ | $0.86 \pm 0.05$ |
| <i>mean</i> | $1.05 \pm 0.11$ | $0.94 \pm 0.05$ | $0.89 \pm 0.14$ | $0.88 \pm 0.04$ |

327 in such a way that most calls use the same set of inputs, but this set is large  
 328 enough so that it cannot fit into a small table.

329 In practice we do not frequently use tables as small as 256 elements.  
 330 However, this illustrates the tradeoffs a user can perform to fine tune this  
 331 framework to the current needs. In some scenarios, such as when targeting  
 332 embedded systems with limited resources, a smaller table could be needed  
 333 due to memory restrictions.

334 Table 3 presents the results of a two-tailed, paired-sample *t-test* performed  
 335 on the execution time measurements of the **fft** and **rgb2hsi** benchmarks (the  
 336 two benchmarks tested with multiple inputs). Tests on the energy results  
 337 follow the same pattern. This statistical test was applied to compare the  
 338 results of each memoization configuration against the results of the original  
 339 version (i.e., without memoization), when running with different inputs (9  
 340 **fft** configurations and 8 **rgb2hsi** images). There is a single configuration  
 341 (highlighted in the table) on which we cannot claim statistical significance

Table 3: Results of a two-tailed, paired-sample *t*-test.

| Benchmark                   | 256-nu                    | 256-u                                       | 65536-nu                  | 65536-u                   |
|-----------------------------|---------------------------|---------------------------------------------|---------------------------|---------------------------|
| fft<br>( <i>df</i> = 8)     | $t = 3.50$<br>$p = 0.008$ | $t = 6.45$<br>$p = 0.000$                   | $t = 2.91$<br>$p = 0.020$ | $t = 6.01$<br>$p = 0.000$ |
| rgb2hsi<br>( <i>df</i> = 7) | $t = 3.63$<br>$p = 0.008$ | $t = 2.02$<br><b><math>p = 0.083</math></b> | $t = 2.50$<br>$p = 0.041$ | $t = 2.53$<br>$p = 0.039$ |

342 with  $\alpha = 0.05$ , the testing of the **rgb2hsi** benchmark with a table of 256  
343 elements and no update in case of collisions. For the other seven out of  
344 eight configurations, the *t*-test considers the data to be statistically different.  
345 Considering there are no other variables on the experiments other than how  
346 memoization is applied, we can confidently say the shown differences are  
347 caused by the technique.

## 348 5. Conclusions

349 This article presented a framework to automatically apply memoization  
350 to C/C++ applications. The resulting applications store outputs of pure  
351 functions mapped by their inputs. If these pure functions are called with  
352 the same inputs, the framework returns the stored value instead of recom-  
353 puting it. Hence, this technique may lead to execution time and energy  
354 consumption improvements by simply avoiding unnecessary computations  
355 in scenarios where critical functions are called with repeating inputs. Our  
356 framework consists of two main components. First, a source-to-source com-  
357 piler that modifies input C/C++ applications by targeting memoizable func-  
358 tions and replacing their calls with calls to wrappers that interface with a C  
359 memoization library. This library, the second component of our framework,  
360 maintains all the data structures and contains memoization logic needed to  
361 enhance applications with this optimization. The source-to-source compiler  
362 can be controlled by the user to perform analyses and decide which func-  
363 tions to target. We showed the impact of the memoization technique in four  
364 representative benchmarks and the usefulness of our software by measuring  
365 reductions in execution time and energy consumption.

## 366 Acknowledgments

367 This work was partially funded by the ANTAREX project through the EU  
368 H2020 FET-HPC program under grant no. 671623. Bispo acknowledges the

369 support provided by *Fundação para a Ciência e Tecnologia* (FCT), through  
370 the Postdoctoral scholarship SFRH/BPD/118211/2016.

## 371 **References**

- 372 [1] D. Michie, “memo” functions and machine learning, *Nature* 218 (5136)  
373 19 (1968).
- 374 [2] C. Silvano, G. Agosta, J. Barbosa, A. Bartolini, A. R. Beccari, L. Benini,  
375 J. Bispo, J. M. P. Cardoso, C. Cavazzoni, S. Cherubin, R. Cmar, D. Gadi-  
376 oli, C. Manelfi, J. Martinovi, R. Nobre, G. Palermo, M. Palkovi, P. Pinto,  
377 E. Rohou, N. Sanna, K. Slaninov, The ANTAREX tool flow for monitor-  
378 ing and autotuning energy efficient HPC systems, in: 2017 International  
379 Conference on Embedded Computer Systems: Architectures, Modeling,  
380 and Simulation (SAMOS), 2017, pp. 308–316.
- 381 [3] A. Suresh, B. Narasimha Swamy, E. Rohou, A. Sez nec, Intercepting Func-  
382 tions for Memoization: A Case Study Using Transcendental Functions,  
383 *ACM Trans. Archit. Code Optim.* 12 (2) (2015).
- 384 [4] A. Suresh, E. Rohou, A. Sez nec, Compile-Time Function Memoization,  
385 in: 26th International Conference on Compiler Construction, Austin, TX,  
386 United States, 2017, pp. 45–54.
- 387 [5] J. M. Cardoso, J. G. Coutinho, T. Carvalho, P. C. Diniz, Z. Petrov,  
388 W. Luk, F. Gonçalves, Performance-driven instrumentation and mapping  
389 strategies using the lara aspect-oriented programming approach, *Soft-  
390 ware: Practice and Experience* 46 (2) 251–287 (2016).
- 391 [6] P. Michaud, Y. Sazeides, ATMI: analytical model of temperature in mi-  
392 croprocessors, Third Annual Workshop on Modeling, Benchmarking and  
393 Simulation (MoBS) (2007).

394 **Current code version**



| <b>Nr.</b> | <b>Code metadata description</b>                                | <b>Please fill in this column</b>                                                                                                                                                                              |
|------------|-----------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| C1         | Current code version                                            | Clava: v3.0.13<br>Memoi: v1.0                                                                                                                                                                                  |
| C2         | Permanent link to code/repository used for this code version    | Clava: <a href="https://github.com/specs-feup/clava">https://github.com/specs-feup/clava</a><br>Memoi: <a href="https://gforge.inria.fr/projects/memoization">https://gforge.inria.fr/projects/memoization</a> |
| C3         | Legal Code License                                              | Clava: Apache License, 2.0<br>Memoi: LGPL V3                                                                                                                                                                   |
| C4         | Code versioning system used                                     | Clava: git<br>Memoi: svn                                                                                                                                                                                       |
| C5         | Software code languages, tools, and services used               | Clava: C++, Java, LARA<br>Memoi: C, sed                                                                                                                                                                        |
| C6         | Compilation requirements, operating environments & dependencies | Linux, Windows, macOS                                                                                                                                                                                          |
| C7         | If available Link to developer documentation/manual             |                                                                                                                                                                                                                |
| C8         | Support email for questions                                     | Clava: joaobispo@gmail.com<br>Memoi: loic.besnard@irisa.fr                                                                                                                                                     |

Table 4: Code metadata