

OCamlJIT 2.0 - Faster Objective Caml

Benedikt Meurer
Compilerbau und Softwareanalyse
Universität Siegen
D-57072 Siegen, Germany
`meurer@informatik.uni-siegen.de`

Abstract

This paper presents the current state of an ongoing research project to improve the performance of the OCAML byte-code interpreter using Just-In-Time native code generation. Our JIT engine OCAMLJIT2 currently runs on x86-64 processors, mimicking precisely the behavior of the OCAML virtual machine. Its design and implementation is described, and performance measures are given.

1 Introduction

The OCAML [25, 30] system is the main implementation of the Caml language¹, featuring a powerful module system combined with a full-fledged object-oriented layer. It comes with an optimizing native code compiler `ocamlopt`, for high performance; a byte-code compiler `ocamlrun`, for increased portability; and an interactive loop `ocaml`, for experimentation and rapid development.

`ocamlrun` and `ocaml` translate the source code into a sequence of byte-code instructions for the OCAML virtual machine, which is based on the ZINC machine [24] originally developed for Caml Light [26]. The optimizing native code compiler `ocamlopt` produces fast machine code for the supported targets, but is only applicable to *static program compilation*. It cannot be used with multi-stage programming in METAOCAML [32], or the interactive top-level `ocaml`.

This paper introduces OCAMLJIT2, a new open-source² implementation of the OCAML virtual machine, based on Just-In-Time native code generation. OCAMLJIT2 is developed on 64-bit Mac OS X 10.6, but should also run on Linux/amd64 and FreeBSD/amd64 systems without modifications.

¹<http://caml.inria.fr/>

²The full source code is available from <http://gitorious.org/ocamljit2/> under the terms of the QPL and LGPL licenses.

The paper is organized as follows: Section 2 presents the existing systems, including the relevant parts of the OCAML implementation, and the previous OCAMLJIT implementation [31] which inspired the present work. Section 3 details the design and implementation of OCAMLJIT2. Performance measures are given in section 4. Section 5 and 6 list related and future work, followed by the conclusion in section 7.

2 Existing systems

We present the existing OCAML system and the previous OCAMLJIT [31] implementation, which is based on the GNU LIGHTNING library [6]. For the sake of completeness we repeat the overview of the OCAML compiler and runtime given in [31] below. Readers already familiar with the internals of the OCAML implementation can skip to 2.3.

2.1 OCaml compiler usage

The OCAML system can be used either interactively via the `ocaml` command, which provides a “*read, compile, eval*” loop, or in batch mode via the `ocamlc` command, which takes a set of `*.ml` or `*.mli` source or `*.cmi` or `*.cmo` byte-code object files³ and produces a byte-code object file (to be reused by a subsequent invocation of `ocamlc`) or a byte-code executable file. `ocamlc` also handles `*.c` sources and `*.so` shared libraries (for external C functions invoked from OCAML) and also deals with byte-code library files `*.cma`. The byte-code executable files produced by `ocamlc` are interpreted by the `ocamlrun` program⁴ upon execution.

The goal of the present work is to substitute the interpreter part of the OCAML runtime `ocamlrun` with a Just-In-Time compiler, which incrementally compiles the byte-code to native code and runs the generated native code, instead of interpreting the byte-code. Unlike OCAMLJIT [31], we do not aim to provide a separate `ocamljitrun` runtime, since that way one would have to explicitly use the runtime replacement to achieve the benefits of Just-In-Time compilation.

2.2 Overview of the OCaml system

This section gives a brief overview of the OCAML byte-code compiler and runtime. Feel free to skip to 2.3 if you are already familiar with the details.

2.2.1 Compiler phases and representations

Compilation starts by parsing an OCAML source file (or a source region in interactive mode) into an abstract syntax tree (AST, see file `parsing/parsedtree.mli` of the OCAML source

³Source files are either module interfaces `*.mli` or module implementations `*.ml`. Byte-code object files are either compiled module interfaces `*.cmi` or compiled module implementations `*.cmo`.

⁴Technically speaking, an OCAML byte-code executable file is a `#!/usr/bin/ocamlrun` Unix script, so `execve` of an OCAML byte-code executables invokes `ocamlrun`.

code). Figure 1 illustrates the abstract syntax tree for the expression `let x=1 in x+3`. Compilation then proceeds by computing the type annotations to produce a typed syntax tree (see file `typing/typedtree.mli`) as shown in figure 2.

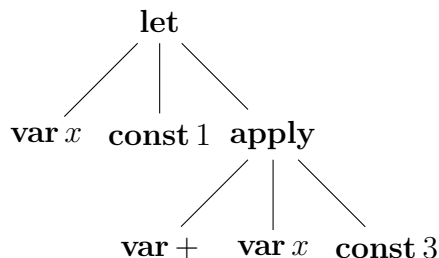


Figure 1: Abstract syntax tree

From this typed syntax tree, the byte-code compiler generates a so called *lambda representation* (see file `bytecomp/lambda.mli`)⁵ as shown in Figure 3, inspired by the untyped call-by-value λ -calculus. This intermediate representation is not directly related to the source code, because source file positions and some names are lost at this stage, and does not contain any kind of explicit type information.

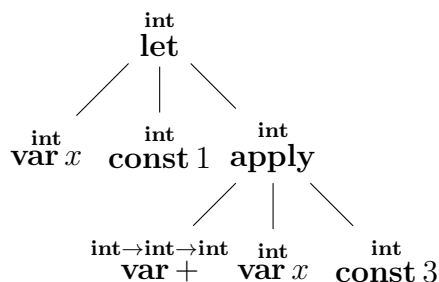


Figure 2: Typed syntax tree

This lambda tree representation includes leaves for variables and constants as well as nodes for function definitions and applications, `let` and `let rec`, primitive operations (like addition, multiplication, array access, etc.), switches (used in compiled forms of pattern matchings), exception handling (`raise` and `try...with` nodes), imperative traits (sequences, `for` and `while` loops), etc.

(let (x/1038 1) (+ x/1038 3))

Figure 3: Lambda representation

After several optimizations are applied (mostly peephole optimizations, transforming lambda trees into *better* or smaller lambda trees), the lambda tree representation is trans-

⁵The native code compiler `ocamlc` has a similar representation (see file `asmcomp/clambda.mli`), which adds explicit direct or indirect calls and closures.

formed into a list of byte-code instructions as shown in figure 4. This instruction list is optimized and afterwards written into the generated byte-code file (or kept in a memory buffer for the interactive top-level).

```
const 1
push
acc 0
offsetint 3
```

Figure 4: Byte-code

The `ocaml` top-level and the `ocamlc` compiler provide two undocumented command line options `-dlambda` and `-dinstr` to display the internal representations mentioned above.

2.2.2 The OCaml virtual machine

The OCAML virtual machine (see file `byterun/interp.c`) is an interpreter for the byte-code⁶ produced by `ocamlc` (as described in the previous section). It operates on a stack of OCAML values and five *virtual registers*: the stack pointer `sp`, the accumulator `accu`, the environment pointer `env` (pointing to the current closure, which contains the values for the free variables), the extra arguments counter `extra_args` (for partial applications), and the byte-code program counter `pc`. The virtual machine also deals with byte-code segments (byte-code sequences to be interpreted) and a global data array `caml_global_data` of OCAML values. There is usually only a single byte-code segment, the byte-code sequence in the executable file produced by `ocamlc`, which also contains the marshalled representation of the global data.

OCAML values are either pointers (usually pointing to blocks in the garbage collected heap) or tagged integers. Each block starts with a header containing a tag, the size of the block and the color (used by the garbage collector). Some tags describe special blocks like strings, closures or floating point arrays, but most of them are used for representing sum types. OCAML distinguishes pointers and tagged integers using the least significant bit: if the least significant bit is 1, the value is a tagged integer, with the integer value stored in the remaining 31 or 63 bits, otherwise the value is a pointer.

The OCAML stack, which is – in contrast to the optimizing native code compiler – separate from the native C stack, contains values, byte-code return addresses and extra argument counts, organized into so-called *call frames*. The byte-code is pointed to from either return addresses stored on the stack or closures stored in the heap. The first slot of every closure contains the byte-code address of the closure’s function code; the remaining slots contain the values of the free variables, thereby forming the environment of the closure. Mutually recursive closures – used to implement the `let rec` language construct – may contain more than just one byte-code address, followed by the values of the free variables.

⁶Since each token of the byte-code is a 32-bit word, the byte-code is actually a *word-code*.

The byte-code interpreter `ocamlrun` starts by unmarshalling the global data, loading the byte-code sequence into memory and processing dynamically linked libraries containing external C function primitives referenced from the byte-code file. Once everything is in place, the function `caml_interprete` (see file `byterun/interp.c`) is invoked to interpret the initial byte-code segment (using threaded code [5, 15]), starting from its first byte-code, with an empty stack, and a default environment and accumulator.

2.2.3 The byte-code instruction set

Every byte-code instruction is represented by one or more consecutive 32-bit words. The first word of each byte-code includes the operation (see file `byterun/instruct.h`) and the remaining words include the operands, which are constant integer arguments (usually offsets). Most byte-code instructions operate on the accumulator and the top-most stack elements and produce a result in the accumulator. Some byte-codes include offsets (denoted by p, q, \dots subscripts), which are either encoded within the operation or following the operation as 32-bit words. All byte-code offsets referencing other byte-codes are interpreted relative to their respective positions, so the byte-code forms a *position-independent code*.

Byte-code instructions are classified as follows:

- Stack manipulation instructions: `ACCp` loads `accu` with the value of the p -th top-most stack cell; `PUSH` pushes the accumulator value onto the stack; `POPp` pops the top-most p elements off the stack; `ASSIGNp` places the accumulator value into the p -th top-most stack cell.
- Loading instructions: `CONSTINTp` loads `accu` with the tagged integer p , whereas `ATOMp` loads `accu` with the p -tagged atom⁷.
- Primitive operations: unary `NEGINT` (negation) and `BOOLNOT` (logical negation) operate on `accu`; binary operations `ADDINT`, `SUBINT`, `MULINT`, `DIVINT`, `MODINT`, `ANDINT`, etc. take the first operand in `accu` and the second one popped off the stack, and place the result into the accumulator, where division and modulus also check for 0 and may raise an exception; comparisons `EQ`, `NEQ`, `LTINT`, `LEINT`, etc. similarly place their boolean result into `accu`; `ISINT` tests whether the accumulator value is a tagged integer.
- Environment operations: `ENVACCp` loads `accu` with the p -th slot of the current environment `env`.
- Apply operations are divided into two categories: `APPLYp` creates a new call frame on the OCAML stack and jumps to the called code, whereas `APPTERMp,q` (q denotes the height of the current stack frame) performs a tail-call to the called code. In either case the accumulator contains the closure which is applied to the p top-most arguments on the OCAML stack and becomes the new `env`. Call frames contain

⁷Atoms are pre-allocated 0-sized blocks.

arguments, byte-code return address, previous `env`, and previous `extra_args`; when adding call frames, the stack may grow if necessary.

- Function return instructions: `RETURNp` pops the top-most p values off the stack and returns to the caller; `RESTART` and `GRABp` handle partial application (allocating appropriate closures as necessary).
- Closure allocation instructions: `CLOSUREp,q` allocates a single closure with p variable slots and byte-code at offset q ; `CLOSURERECp,q,k1,...,kp` allocates a mutually recursive closure with p functions k_1, \dots, k_p and q variable slots.
- Allocation instructions: `MAKEBLOCKp,q` creates a p -tagged block of q values (first in `accu`, remaining popped off the stack); `MAKEFLOATBLOCKp` creates an array of floats. The accumulator is loaded with the pointer to the newly allocated block. Every allocation may trigger a garbage collection.
- Field access and modify instructions: `GETGLOBALFIELDp,q` loads `accu` with the q -th field of the p -th global value; `GETFIELDp` loads the accumulator with the p -th field of the block pointed to by `accu`; symmetrically `SETFIELDp` sets the p -th field of `accu` to the value popped off the stack; similarly `GETFLOATFIELDp` and `SETFLOATFIELDp` handle fields in floating point blocks; `GETGLOBALp` and `SETGLOBALp` handle global fields in `caml_global_data`; `GETSTRINGCHAR` and `SETSTRINGCHAR` are used to access and update characters within strings (the index being popped off the stack). All modifying operations have to cooperate with the garbage collector.
- Control instructions include – in addition to the instructions for function application – conditional `BRANCHIFp` and `BRANCHIFNOTp` (depending upon boolean value in `accu`), comparing `BEQp,q`, `BNEGp,q`, ... (comparing q with the integer value in `accu`) and unconditional `BRANCHp` jumps. There is also a `SWITCHp,q,k1,...,kp,k'1,...,k'q` instruction (used to compile the OCAML `match` construct), which tests `accu` and jumps to offset k_{i-1} if `accu` is the tagged integer i or k'_{j-1} if `accu` points to a j -tagged block. Exception and signal handling instructions `PUSHTRAPp`, `POPTRAP`, `RAISE CHECK_SIGNALS` are also control instructions, as is the halting `STOP` instruction, which is the last byte-code of a segment.
- Calling external C primitives: `C_CALLp,q` calls the p -th C primitive with q arguments (first taken from `accu`, remaining popped off the stack). The result is stored into `accu`. Primitives are used for several basic operations, including operations on floating point values.
- Object-oriented operations: `GETMETHOD` retrieves an object method by its index; `GETPUBMET` and `GETDYNMET` fetch the method for a given method tag.
- Debugger related instructions: the OCAML debugger places breakpoints by overwriting byte-code instructions with `BREAK`; these are unsupported by OCAMLJIT2.

2.3 OCamlJit

In this section we briefly describe the design and implementation of the GNU LIGHTNING [6] based OCAMLJIT [31], which inspired many aspects of our present work. The main goal of OCAMLJIT is maximal compatibility with the OCAML byte-code interpreter, its runtime system (including the garbage collector), and its behavior. That means programs running under `ocamljitrun` see the same virtual machine as `ocamlrun`, so utilizing OCAMLJIT is a matter of exchanging `ocamlrun` with `ocamljitrun` when executing OCAML programs.

In order to meet these constraint, OCAMLJIT has to mimic as much as possible the byte-code interpreter.

2.3.1 Implementation

Since LIGHTNING provides only a few registers, the most commonly used OCAML virtual registers (`accu`, `sp`, `env`) are mapped to LIGHTNING registers (hence to machine registers), while other less common OCAML registers are grouped into a *state record* pointed to by a LIGHTNING register.

The core of OCAMLJIT is the `caml_jit_translate` C function, which scans a byte-code sequence and emits equivalent native code. It loops around a big `switch` statement, with one `case` for every byte-code instruction. For example, figure 5 shows the case for the `ANDINT` instruction, which pops a tagged integer off the OCAML stack and does a “bitwise and” with `accu`.

```
case ANDINT:
    /* tmp1 = *sp; ++sp; accu &= tmp1; */
    jit_ldr_p(JML_REG_TMP1, JML_REG_SP);
    jit_addi_p(JML_REG_SP, JML_REG_SP, WORDSIZE);
    jit_andr_l(JML_REG_ACCU, JML_REG_ACCU, JML_REG_TMP1);
    break;
```

Figure 5: `ANDINT` case in `caml_jit_translate`

Byte-code and native-code addresses Total compatibility demands the use of byte-code addresses, in particular within OCAML closures. This means that closure application – a very common operation in OCAML – has to retrieve the byte-code from the closure, find the corresponding native machine code, and jump to it. The byte-code to native machine code address mapping is accomplished by a sparse translation hash-table. An entry is added to this table for every compiled byte-code instruction.

Every byte-code segment has its own native code block, which references a linked list of native code chunks. Such a code chunk is a page-aligned executable memory segment, allocated via the `mmap` system call with read, write and execute permissions.

The byte-code is incrementally translated to native machine code. The native code generation is interrupted when the current native code chunk is filled, a `STOP` instruction

is reached, a configurable number of byte-code instructions has been translated, or when the currently translated byte-code address is already known in the translation hash-table (because a byte-code sequence containing it was previously translated).

Interaction with the runtime The generated native machine code has to interact with the runtime. A common interaction is invoking the garbage collector when space in the minor heap is exhausted during allocations (i.e. `MAKEBLOCK`, `CLOSURE`, etc.). Most of the commonly used runtime interactions are inlined in the generated native code.

Less common interactions, which are too complex to be inlined, are performed by saving the current state of the virtual machine, returning from the virtual machine with a particular `MLJSTATE_*` (passing the required data in the state structure), and letting the `caml_interprete` function do the processing. For example `MLJSTATE_MODIF` is used to modify a heap cell while `MLJSTATE_RAISE` is used to raise an exception (via C).

Performance On x86 machines, `ocamljitrun` typically gives significant speedups with respect to `ocamlrun` of a factor above two. While this is already a quite interesting achievement, there are however some shortcomings to be noted:

1. The compilation overhead is high, even with this naive compilation scheme. This is especially noticable in short running programs. For example, building the OCAML standard library is nearly three times slower with `ocamljitrun` than with `ocamlrun`. There are various points that add to this slow translation speed; for example, adding every (byte-code, native-code) address pair to the translation hash-table takes a significant amount of the translation time – actually more than a fourth of it [31].
2. The limited register set provided by GNU LIGHTNING prevents efficient register usage for modern targets like x86-64, ARM or PowerPC, which offer more than just 8 general purpose registers. In addition, the *portability at the lowest level* approach of GNU LIGHTNING prevents interesting target specific optimizations.
3. The naive compilation scheme (with peephole optimizations) is limited in efficiency. There is almost no way to achieve performance close to the optimizing native code compiler `ocamlopt`.

The present work is part of a research project to overcome these problems.

2.4 Runtime code generators

Several runtime code generators are available today, including but not limited to GNU LIGHTNING [6], LLVM [22, 23] and ASMJIT [21]. These projects provide building blocks for JIT engines using different levels of abstraction.

GNU LIGHTNING, which was recently ported to the x86-64 architecture⁸, provides users with a *portability at the lowest level* approach. To gain reasonable portability, one has to

⁸Also known as *AMD64 architecture*, but we prefer the vendor-neutral term.

limit itself to the least common denominator w.r.t. instruction sets and registers available on the target platforms. This is a great concept for rapid prototyping, but it is bound to fail in the long run.

LLVM⁹ provides an internal code representation, which describes a program using an abstract RISC-like instruction set, and offers C/C++ APIs for code generation and Just-In-Time compilation (currently supporting x86, x86-64 and PowerPC). But this comes at a cost: In a simple OCAMLJIT2 prototype based on LLVM, we have measured significant compilation overhead; short running programs (like `ocamlc` applied to a small to medium sized `*.ml` file) were three to four times slower than running the same program with the byte-code interpreter. Similar results were observed by other projects using LLVM for Just-In-Time compilation. For example, the Mono project¹⁰ reported impressive speedups for long running, computationally intensive applications, but at the cost of increased compilation time and memory usage.

The ASMJIT project aims to provide a nice API to emit x86 and x86-64 code at runtime (support for ARM processors is planned) using a well-designed C++ API. Our first fully working prototype used ASMJIT to generate x86-64 native code. While this worked out very well, we came to the conclusion that we did not actually need any of the features provided by ASMJIT, except for the code emitter functions, and we found mixing C++ code with the somewhat ancient OCAML C code a rather painful experience. We ended up writing our own C preprocessor macros replacing the code emitter functions provided by ASMJIT.

3 Design and implementation

We aim to provide almost the same amount of compatibility with the OCAML byte-code interpreter as provided by OCAMLJIT, although not at all costs. At each point where we have to either adjust the runtime or add costly work-arounds within the JIT engine, we choose to adjust the runtime. Nonetheless OCAMLJIT2 should be able to handle any byte-code sequence, and the heap and stack should be the same as with the OCAML byte-code interpreter.

Our design goals are therefore similar to those of OCAMLJIT:

- The runtime (i.e. the garbage collector and the required C primitives) should be (roughly) the same as in the byte-code interpreter.
- The heap is the same, in particular all OCAML values keep exactly the same representation (i.e. the code pointer inside a closure is still a byte-code pointer, not a native code pointer). The relevant parts of the byte-code runtime (i.e. the serialization mechanism) are reused¹¹.

⁹<http://llvm.org/>

¹⁰<http://www.mono-project.com>

¹¹Especially hashing and serializing closures or objects (with their classes) give the same result in OCAMLJIT2 as in the OCAML byte-code interpreter.

- C programs embedding the OCAML byte-code interpreter should work with OCAMLJIT2 as well, without the need to recompile the C program¹².
- OCAMLJIT2 should execute programs at least as fast as the byte-code interpreter, even for short running programs. In particular, the compilation time must not be noticeably longer when evaluating phrases within the `ocaml` top-level.

A program running with OCAMLJIT2 should therefore see the same virtual machine as seen when running with the byte-code interpreter. It should not be able to tell whether it is being interpreted or JIT compiled, except by measuring the execution speed.

The relevant files are `byterun/jit.c`, which contains the actual JIT engine, `byterun/jit_rt_amd64.S` containing the x86-64 runtime setup and helper routines for the JIT generated machine code, and `byterun/jx86.h`, which contains the x86-64 (and x86) code emitter preprocessor macros.

3.1 Virtual machine design

This section describes the mapping of the OCAML virtual machine to the x86-64 architecture, in particular the mapping of the virtual machine registers to the available physical registers. This is inherently target specific, in contrast to most of the other ideas we present.

The x86-64 architecture [1, 20] provides 16 general purpose 64-bit registers `%rax`, `%rbx`, `%rcx`, `%rdx`, `%rbp`, `%rsp`, `%rdi`, `%rsi`, `%r8`, ..., `%r15`, as well as 8 80-bit floating-point registers (also used as MMX registers) and 16 128-bit SSE2 registers `%xmm0`, ..., `%xmm15`. The System V ABI for the x86-64 architecture [28], implemented by almost all operating systems running on x86-64 processors¹³, mandates that registers `%rbx`, `%rbp`, `%rsp` and `%r12` through `%r15` belong to the calling function and the called function is required to preserve their values. The remaining registers belong to the called function.

Since the native machine code generated by the JIT engine will have to call C functions (i.e. primitives of the byte-code runtime) frequently, it makes sense to make good use of callee-save registers to avoid saving and restoring too many aspects of the machine state whenever a C function is called. We therefore settled on the following assignment for our current implementation:

- `accu` goes into `%rax` for various reasons: typically `accu` does not need to be preserved across C calls, but is usually assigned the result of C function calls, which is then already available in `%rax` [28]; additionally, many computations are performed on `accu` and several x86-64 instructions have shorter encodings when used with `%rax`.
- `extra_args` goes into `%r13`. One difference to the byte-code interpreter is that `%r13` contains the number of extra arguments as tagged integer, in order to avoid the conversions when storing it to and loading it off the OCAML stack.

¹²We therefore need to ensure that the API of OCAMLJIT2's `libcamlrn_shared.so` match exactly the API of OCAML's `libcamlrn_shared.so`.

¹³With the notable exception of Win64, which uses a quite different ABI.

- The environment pointer `env` goes into `%r12`.
- The stack pointer `sp` goes into `%r14`.
- The byte-code program counter `pc` is not needed anymore.

We use `%r15` to cache the value of the minor heap allocation pointer `caml_young_ptr`, similar to `ocaml_opt`. This greatly speeds up allocations of blocks in the minor heap, which is a common operation. The remaining registers – except for `%rsp`, of course – are available to implement the native machine code for the byte-code instructions.

There are several other target specific issues to resolve besides the assignment of registers. For example, we need to ensure that the native stack is always aligned on a 16-byte boundary prior to calling C functions. But those issues are beyond the scope of this document.

The setup and helper routines that form the runtime of the JIT engine are in the file `byterun/jit_rt_amd64.S`. The most important is `caml_jit_rt_start`, which starts the virtual machine by setting up the runtime environment, loading the appropriate registers, and jumping to the first byte-code instruction (via a byte-code trampoline). This is the only JIT runtime function that may be called from C, all other functions may only be called from inside the generated native code. `caml_interprete` invokes `caml_jit_rt_start` to execute the byte-code instruction sequence given to it (using a demand driven compilation scheme, described below).

When the execution of the byte-code instruction sequence is done (i.e. the translated `STOP` instruction is reached) or an exception is thrown and not caught within the byte-code sequence, `caml_jit_rt_stop` is invoked, which undoes the runtime setup of `caml_jit_rt_start` and returns to the calling C function (i.e. `caml_interpret`).

3.2 Address mapping

As previously described, OCAMLJIT uses a sparse hash-table to map byte-code addresses to native machine code addresses. Whenever the virtual machine needs to jump to a byte-code pointer (i.e. during closure application), it has to lookup the corresponding native machine code in the hash-table, and jump to it (falling back to the JIT compiler if no native code address is recorded).

Since closure application is a somewhat common operation in OCAML, this indirection via a hash-table does not only complicate the implementation, but also decreases the performance of the virtual machine. In addition, constructing the hash-table during JIT compilation takes a significant amount of the translation time. To make matters worse, OCAMLJIT adds an entry for every translated (byte-code, native-code) address pair to the hash-table, even though the vast majority of these entries is never used.

Using a hash-table for the address mapping was therefore not an option for OCAMLJIT2. While looking for another way to associate native code with byte-code addresses, we found that the threaded byte-code interpreter is faced with a similar problem, whose solution may also be applicable in the context of a JIT engine. When threading is enabled for the

byte-code interpreter¹⁴, prior to execution, every byte-code segment is preprocessed by the C function `caml_thread_code`, which replaces each instruction opcode in the byte-code sequence with the offset of the C code implementing the instruction relative to some well-defined base address. Executing a byte-code instruction is then a matter of determining the C code address from the offset in the instruction opcode and the well-defined base address¹⁵, and jumping to that address. The threaded byte-code sequence itself therefore provides a mapping from byte-code to native machine code addresses.

We have found that a similar approach is applicable in the context of a JIT engine. Whenever a given byte-code instruction is JIT compiled to native code, we replace the instruction opcode word within the byte-code sequence by the offset of the generated native code relative to the base address `caml_jit_code_end`. When jumping to a byte-code address – i.e. during closure application – we just read the offset from the instruction opcode field, add it to `caml_jit_code_end`, and jump to that address. This trampoline code for x86-64 is shown in figure 6 (`%rdi` contains the address of the byte-code to execute and `%rdx` is a temporary register, not preserved across byte-code jumps).

```
movslq (%rdi), %rdx
addq   caml_jit_code_end(%rip), %rdx
jmpq   *%rdx
```

Figure 6: Byte code trampoline (x86-64)

Care must be taken on 64-bit targets, as the addressable range is limited to $-2^{31} \dots 2^{31} - 1$ bytes, since the width of the instruction opcode field is only 32 bits. Our current approach allocates a fixed $2^{27} - x$ bytes code area¹⁶ (using `mmap`) with `caml_jit_code_end` pointing to the end of the code area. The offsets are therefore values in the range $-2^{27} + x$ to -1 .

3.3 Demand driven compilation

Byte-code is incrementally translated to native machine code. Whenever the JIT engine detects a byte-code instruction, which has not been translated to native machine code yet, it invokes the C function `caml_jit_compile`. `caml_jit_compile` then translates the byte-code block starting at the given byte-code offset and all byte-code blocks, which appear as branch targets. These *byte-code blocks* are basic blocks within the OCAML byte-code, terminated by branch, raise, return and stop instructions. Translation of a byte-code block also stops once an instruction is hit, which was already translated, and the compiler simply generates a jump to the previously translated native machine code.

With the address mapping described above, this incremental, *on-demand* compilation scheme was very easy to implement. Given that all offsets to native machine code are negative values and all byte-code instruction opcodes are positive values between 0 and

¹⁴This requires the byte-code interpreter to be compiled with `gcc` at the time of this writing.

¹⁵On 32-bit targets the base address is 0, hence the offset is already the address, while on 64-bit targets the offset is usually added to the base address.

¹⁶Where x is a small amount of memory reserved for special purposes.

145, it is easy to distinguish already translated and not yet translated byte-code instructions. Now one could alter the byte code trampoline in figure 6 and add a test to see whether (`%rdi`) contains a positive value, and if so invoke the `caml_jit_compile` function. We did however discover that such a test is unnecessary if we ensure that a jump to an address in the range `0...145` relative to `caml_jit_code_end` automatically invokes the `caml_jit_compile` function. This was surprisingly easy to accomplish on x86-64 (and x86); we simply reserve space after `caml_jit_code_end` for 145 `nop` instructions and trampoline code which invokes `caml_jit_compile` and jumps to the generated code afterwards. The byte-code compile trampoline for x86-64 is shown in figure 7.

```

movq  %rax, %rbx
call  caml_jit_compile(%rip)
xchgq %rbx, %rax
jmpq  *%rbx

```

Figure 7: Byte-code compile trampoline (x86-64)

Now whenever a jump to a not yet translated byte-code instruction is performed, the byte-code trampoline will jump to one of the `nop` instructions and fall through the following `nop`'s to the byte-code compile trampoline. At this point `%rax` contains the `accu` value, `%rdi` contains the byte-code address and current state of the virtual machine is stored in global variables and callee-save registers (see section 3.1). Now the compile trampoline preserves `%rax` in the callee-save register `%rbx` and invokes `caml_jit_compile` with the byte-code address to compile in `%rdi`. Once `caml_jit_compile` returns, `%rax` contains the address of the generated native code, so we save that address to `%rbx`, restore the previous value of `%rax`, and jump to the native machine code.

3.4 Main compile loop

In the previous section we have already briefly described the `caml_jit_compile` function, which performs the actual translation of byte-code instructions to native machine code. This section gives a detailed overview of the translation process.

Whenever the byte-code compile trampoline invokes `caml_jit_compile`, it translates the byte-code block starting at the byte-code instruction that triggered the translation (usually the beginning of a function, i.e. a `GRAB` instruction) and continues until a terminating instruction is found (i.e. `STOP`, `RETURN`, etc.) or an already translated instruction is reached. During this process, all referenced byte-code addresses, which have no associated native machine code yet, are collected in a list of pending byte-code blocks. Once the translation of a byte-code block is done, `caml_jit_compile` continues with the next pending block, and repeats until all pending blocks are translated. It is trivial to see that this process terminates, since every block is translated at most once; as soon as it is translated, the native machine code address is known and it will not be added to the pending list again.

The actual translation of byte-code instructions is done in a large `switch` statement, which is quite similar to the non-threaded byte-code interpreter, except that we do not interpret the opcodes, but generate appropriate native machine code. Figure 8 shows the case for the `ANDINT` instruction on x86-64.

```
case ANDINT:
    jx86_andq_reg_membase(cp, JX86_RAX, JX86_R14, 0);
    jx86_addq_reg_imm(cp, JX86_R14, 8);
    break;
```

Figure 8: `caml_jit_compile` `ANDINT` case (x86-64)

The generated native code for `ANDINT` on x86-64 is shown in figure 9. The `andq` instruction performs the “bitwise and” of `accu` (located in `%rax`) and the top-most stack cell (located in `0(%r14)`), storing the result into `%rax`. The `addq` instruction pops the top-most cell off the stack (the OCAML stack grows towards lower addresses).

```
andq 0(%r14), %rax
addq $8, %r14
```

Figure 9: `ANDINT` native code (x86-64)

The translation of most other byte-code instructions is just as straight-forward as the `ANDINT` case shown above. There are however a few byte-code instructions that required special processing, most notably branch instructions. These instructions are used to change the flow of control within a byte-code segment. Whenever a `BRANCH` instruction is hit, we check whether the branch target is already translated (i.e. whether the opcode of the instruction contains a native code offset). If so, we simply generate a `jmp` to the known native code address. Otherwise we generate a *forward jump* (a `jmp` instruction with no known target), record a reference from the target byte-code to the forward jump¹⁷, and append the target byte-code to the pending list (if it wasn’t already listed there). The translation of conditional branch and `SWITCH` instructions follows a similar scheme. Since `caml_jit_compile` loops until all pending blocks are translated, we are sure that all forward jumps will have been patched once `caml_jit_compile` returns.

Recording a reference from a target byte-code to a forward jump is done by threading the jump templates to the target byte-code: The current value of the target byte-code opcode is stored into the forward jump and the target byte-code opcode field is set to point to the address of the forward jump. This is illustrated in figure 10, which shows two forward jumps f_1 and f_2 threaded to a byte-code address b_1 , which contains a not yet translated `PUSH` instruction.

The `caml_jit_compile` driver loop works by allocating a native code address, patching

¹⁷The term *forward jump* may be misleading, since the jump does not need to be forward within the byte-code segment.

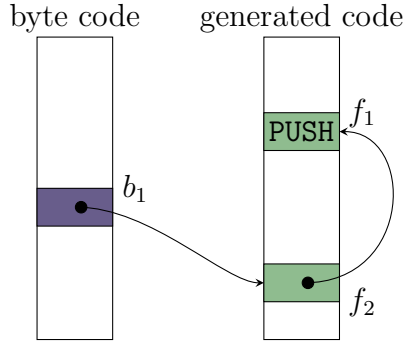


Figure 10: Threading of forward jumps

all forward jumps threaded to the current byte-code address¹⁸, replacing the instruction opcode of the current byte-code with the offset to the native machine code, and generating the native code for the byte-code instruction at the allocated address. This way all byte-code instructions within the translated blocks contain the offsets of the appropriate native machine code once `caml_jit_compile` returns.

3.5 Optimizations

We have applied several optimizations to improve the performance of the generated native machine code. In this section we describe two important optimizations that led to significant performance improvements.

3.5.1 Stack pointer updates

The compilation scheme described above is similar to the one implemented by OCAMLJIT in that it generates a lot of unnecessary stack pointer updates. For every instruction that pushes values to or pops values off the stack an appropriate update of the stack pointer is generated. For example, the byte-code sequence `PUSH ACC4 ADDINT` generates two stack pointer updates, one decrement by `PUSH` and one increment by `ADDINT`, which are unnecessary as the stack pointer will point to the initial stack cell again after executing the three instructions.

In order to avoid these unnecessary updates, we introduced a *stack offset* into the compilation loop, which identifies the current top of stack relative the stack pointer in the hardware register. Whenever the real stack pointer must reside in the hardware register (i.e. at the end of a byte-code block), we emit an instruction to add the stack offset to the stack pointer, and reset the stack offset to 0.

This required changes to the address handling in the compilation loop: Byte-code addresses can only be mapped to native code addresses if the stack offset is 0 at this point

¹⁸We intentionally omitted the implementation details of threading and patching, and refer to the interested reader to the source code in the file `byterun/jit.c`.

in code generation, and when patching forward jumps, the stack offset must be flushed first. Now not every translated byte-code instruction gets mapped to a native code address, which may lead to duplicate translation of the same byte-code instruction. However this duplicate translation happens seldomly and affects mostly backward branches, which are usually loop branches jumping to a `CHECK_SIGNALS` byte-code instruction. So we can almost reduce the probability of duplicate translation to zero by ensuring that the stack offset is flushed prior to generating code for `CHECK_SIGNALS` instructions.

This way we were able to avoid nearly all unnecessary stack pointer updates, which also reduced the size of the generated native machine code by 4 – 17% in our tests.

3.5.2 Floating point operations

Most floating point operations are implemented as C primitives in the byte-code interpreter. Calling C primitives from the generated native code comes at a certain cost – i.e. making the stack and minor allocation pointer available to C code, and reloading it once the C function returns – which significantly decreases the performance of programs using a lot of floating point operations. We have therefore decided to inline various floating point primitives (`caml_add_float`, `caml_sub_float`, etc.), which speed up certain floating point benchmarks by a factor of 1.7.

4 Performance

The performance measurements presented here are done on a MacBook with an Intel Core 2 Duo 2.4 GHz CPU (3 MiB L2 Cache), and 4 GiB 1067 MHz DDR3 RAM, running Mac OS X 10.6.4. The C compiler is `gcc-4.2.1` (Apple Inc. build 5664). The OCAML compiler is 3.12.0. The OCAMLJIT2 code is the tagged revision `ocamljit2-2010-tr1`, compiled with `gcc` optimization level `-O3`.

4.1 Execution time speedup

We measured only the total execution time, as this is the significant time. Table 1 gives some timings (combined system and user CPU time, in seconds), comparing the byte-code interpreter time t_{byt} to the OCAMLJIT2 time t_{jit} , and the time t_{opt} taken by the same program compiled with the optimizing native code compiler `ocamlopt`. It also lists the relative speedups $\sigma_{byt}^{jit} = \frac{t_{byt}}{t_{jit}}$, $\sigma_{jit}^{opt} = \frac{t_{jit}}{t_{opt}}$, and $\sigma_{opt}^{jit} = \frac{t_{byt}}{t_{opt}}$, where bigger values are better, and a value less than 1 would represent a regression.

The commands listed in the first 11 rows are test programs from the `testsuite/tests` folder of the OCAML 3.12.0 distribution. They do more or less represent typical OCAML applications. The remaining rows are invocations of the `ocamlc` and `ocamlopt` compilers, where `stdlib/*.ml` means all source files from the `stdlib` folder of the OCAML 3.12.0 distribution that do not require special compilation arguments, like `-nopervasives` for `stdlib/pervasives.ml*` or `-nolabels` for `stdlib/*Labels.ml*`.

invocation command	time (cpu sec.)			speedup			notes
	t_{byt}	t_{jit}	t_{opt}	σ_{byt}^{jit}	σ_{jit}^{opt}	σ_{byt}^{opt}	
almabench	27.83	10.05	4.47	2.77	2.25	6.22	number crunching
almabench.unsafe	27.74	9.76	4.35	2.84	2.24	6.38	number crunching (no bounds check)
bdd	8.50	1.93	0.68	4.41	2.86	12.60	binary decision diagram
boyer	4.34	1.68	1.06	2.59	1.59	4.11	term processing
fft	5.72	2.17	0.64	2.63	3.39	8.94	fast fourier transformation
nucleic	14.78	4.10	0.80	3.61	5.13	18.52	floating point
quicksort	6.78	1.27	0.24	5.34	5.34	28.48	array quicksort
quicksort.unsafe	4.07	0.93	0.20	4.36	4.79	20.86	array quicksort (no bounds check)
soli	0.17	0.03	0.01	5.09	3.40	17.30	tiny solitaire
soli.unsafe	0.14	0.02	0.01	6.27	2.75	17.25	tiny solitaire (no bounds check)
sorts	19.29	7.13	3.71	2.71	1.92	5.19	various sorting algorithms
ocamlc -help	0.01	0.01	0.00	1.14	7.00	8.00	(short execution)
ocamlc -c format.ml	0.17	0.08	0.05	2.16	1.65	3.57	
ocamlopt -c format.ml	0.65	0.28	0.19	2.31	1.46	3.38	
ocamlc -c stdlib/*.ml	2.03	0.95	0.67	2.14	1.42	3.04	byte-compile stdlib/*.ml
ocamlopt -c stdlib/*.ml	5.04	2.36	1.70	2.14	1.39	2.97	native-compile stdlib/*.ml

Table 1: Running time and speedup

Figure 11 highlights the speedup of OCAMLJIT2 w.r.t. the byte-code interpreter. As expected, really short executions (i.e. **ocamlc -help**) do not benefit from the Just-In-Time compilation, but thanks to our design, their running time is on par with the byte-code interpreter (in contrast to OCAMLJIT, where the running time was worse). Other invocations are speed up by a factor of 2.1 to 6.3.

Floating point intensive programs like **almabench**, **fft** and **nucleic** spend a large amount of time collecting garbage, which is due to the fact that we still allocate a heap block for each and every floating point value used during execution. We intend to further optimize the performance here using better instruction selection and register allocation.

In particular short running programs, like invocations of **ocamlc** with small to medium sized ***.ml** files, which caused noticable slow-downs with OCAMLJIT, benefit from OCAMLJIT2. For example, building the full OCAML distribution – using **make world opt** in a clean source tree – takes only 153.8s¹⁹ with OCAMLJIT2, while OCAML requires 261s. This may not seem significant upon first sight, but it is indeed a nice improvement for a task that is mostly I/O bound and involves numerous invocations of the C compiler, the assembler and the linker, which aren’t affected by the use of OCAMLJIT2.

4.2 Branch prediction accuracy

Efficient virtual machine interpreters – like the OCAML byte-code interpreter – perform a large number of indirect jumps (up to 11% of all executed instructions in our benchmarks). Even though branch target buffers (BTBs) in modern CPUs and various interpreter techniques [8] can greatly reduce the number of mispredicted branches, low branch prediction

¹⁹Measured in combined user and system CPU time.

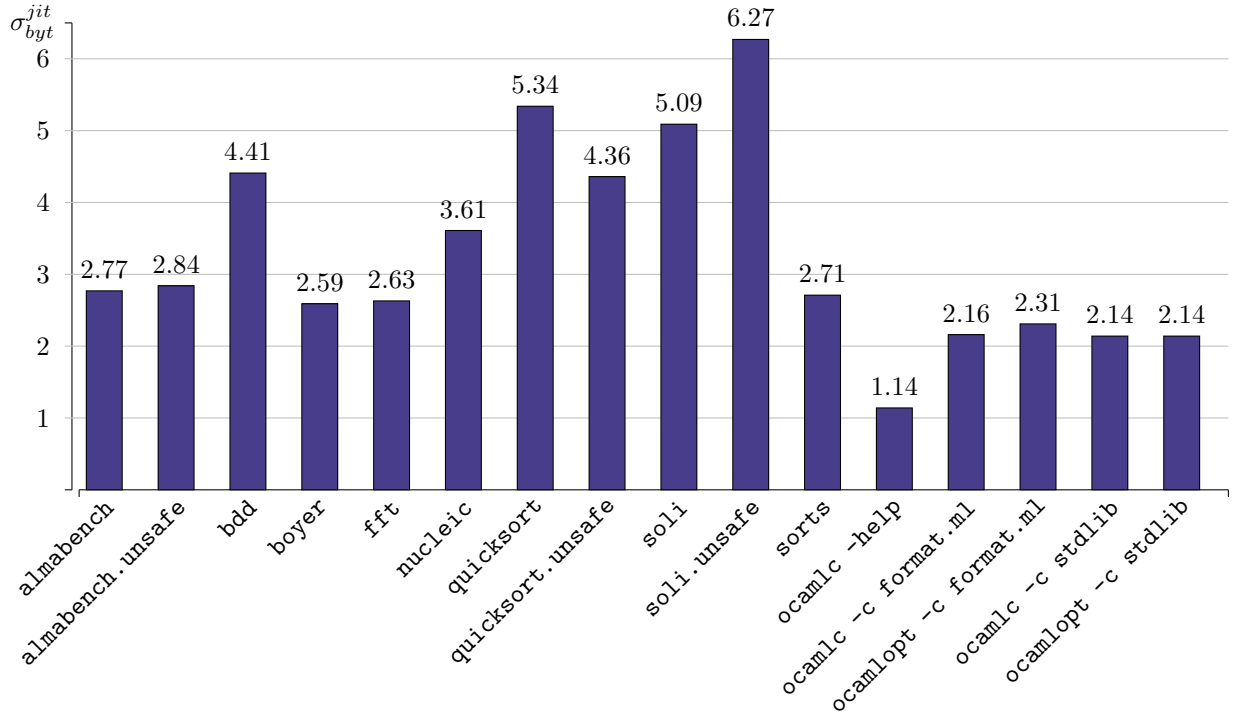


Figure 11: Speedup w.r.t. the byte-code interpreter

accuracy is still a major bottleneck in virtual machine interpreters [14].

Virtual machines based on Just-In-Time compilers are able to reduce almost all indirect branches and thereby increase the branch prediction accuracy to nearly 98 – 99% in most cases. A few indirect branches cannot be eliminated, for example closure application in a virtual machine for a functional language needs an indirect branch (unless the closure’s code pointer is known at compile time). Figure 12 compares the branch prediction accuracy of OCAML and OCAMLJIT2 in various benchmarks on the Intel Core 2 Duo.

4.3 Comparison with OCamlJit

Unfortunately we were unable to get OCAMLJIT running with OCAML 3.12.0 and the most recent revision of GNU LIGHTNING, which supports x86-64 processors, so we are unable to directly compare the running times of OCAMLJIT and OCAMLJIT2 on identical hardware.

We can only compare the relative speedups achieved on different architectures (x86 in case of OCAMLJIT and x86-64 in case of OCAMLJIT2), which seem to indicate that we are doing quite well.

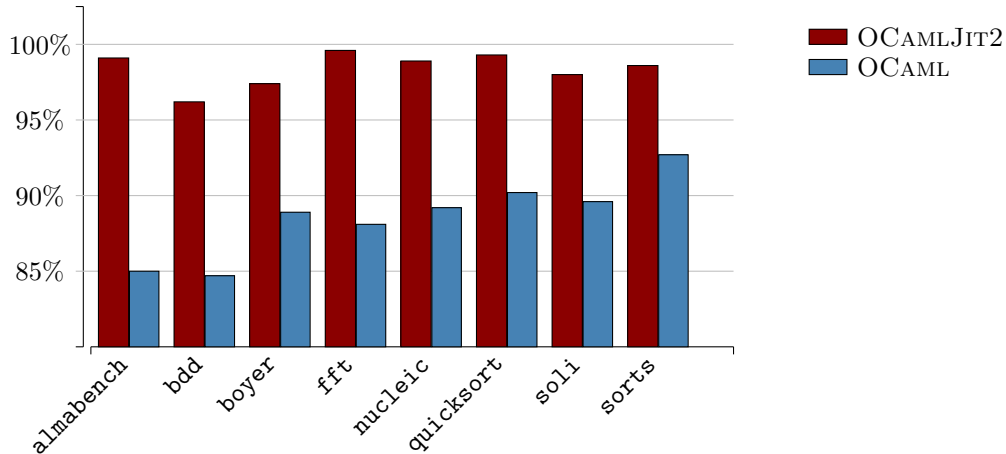


Figure 12: Branch prediction accuracy

5 Related work

Just-In-Time compilation is an active research field [3, 4]. A lot of work has been done on efficient dynamic compilation in the field of the Java Virtual Machine (JVM) and the Common Language Runtime (CLR), which is part of the .NET framework. JIT compilation of Java or .NET byte-code is usually driven by first interpreting the byte-code, collecting profiling information, and selecting the appropriate methods to optimize by inspecting the collected profiling data [9, 33].

Both JVM and CLR depend on profile guided optimizations for good performance, because their byte-code is slower than the OCAML byte-code. This is mostly because of the additional overhead that comes with exception handling, multi-threading, class loading and object synchronization. But it is also due to the fact that both the JVM byte-code [27] and the Common Intermediate Language (CIL) [10] instruction sets are not as optimized as the OCAML byte-code.

More recent examples include the Lua programming language [18, 19], a powerful, fast, lightweight, embeddable scripting language, accompanied by LuaJIT [29], which combines a high-performance interpreter with a state-of-the-art Just-In-Time compiler.

Another recent example is the Dalvik virtual machine²⁰ [7], which drives applications running on the Android mobile operating system²¹. As of Android 2.2, Dalvik features a Just-In-Time compiler, which increased performance of several applications by a factor of 2 to 4.

During the last decades a lot of work has been done on improving the performance of interpreters and native code compilers for other stack-based languages like FORTH [12], in particular compiling FORTH to efficient native code for register machines [11, 13, 16, 17].

²⁰<http://dalvikvm.com/>

²¹<http://www.android.com/>

6 Future work

Although the current implementation already performs quite well in most cases, it also has several drawbacks, which will be addressed in the future.

First of all, instruction selection, instruction scheduling and register allocation [2] in the current implementation are far from optimal. For example most of the generated code uses less than a fourth of the registers provided by the x86-64 processor. This is due to the simple compilation scheme.

The code memory management is another open issue. Right now code memory is managed by allocating in one large memory chunk. Unused code memory is never released or reused, which may cause trouble for METAOCAML [32] or long-lived interactive top-level sessions. This could be addressed by adding support for garbage collected byte-code segments to the OCAML runtime, which would then also take care of native machine code segments.

We also plan to improve the portability of the JIT engine. While it is more or less straight-forward to port the current code to the x86 architecture (trivial in most cases, but nevertheless a time consuming task), it would take considerable amounts of time to port the JIT engine to a different architecture like PowerPC, Sparc or ARM.

In the future, we may even be able to get the JIT engine on par with the optimizing native code compiler `ocamlopt` w.r.t. execution speed, which would offer the possibility to drop `ocamlopt` and its runtime. This would improve the maintainability of OCAML, because then only a single compiler and a single runtime would have to be maintained – plus the JIT engine, of course. However this is an optional goal and by no means a requirement, especially since OCAML’s optimizing native code compiler is also of educational interest.

7 Conclusion

The OCAMLJIT2 implementation presented here did achieve some significant speedups (at least twice as fast as the byte-code interpreter in all relevant comparisons), but there are various open issues to address. In particular the suboptimal instruction selection, instruction scheduling and register allocation will have to be addressed to get OCAMLJIT2 on par with the optimizing native code compiler and JIT engines available for other programming languages and runtime environments.

Acknowledgements

We would like to thank Simon Meurer and Christian Uhrhan for their careful proof-reading.

References

- [1] Advanced Micro Devices, Inc. *AMD64 Architecture Programmer's Manual Volume 1: Application Programming*, Nov 2009. <http://developer.amd.com/documentation/guides/Pages/default.aspx>.
- [2] A. V. Aho, M. S. Lam, R. Sethi, and J. D. Ullman. *Compilers: Principles, Techniques, and Tools*. Addison Wesley, 2nd edition, August 2006.
- [3] M. Arnold, S. J. Fink, D. Grove, M. Hind, and P. F. Sweeney. A Survey of Adaptive Optimization in Virtual Machines. In *Proceedings of the IEEE, 93(2), 2005. Special issue on program generation, optimization, and adaptation*, 2004.
- [4] J. Aycock. A Brief History of Just-In-Time. *ACM Computing Surveys*, 35(2):97–113, 2003.
- [5] J. R. Bell. Threaded code. *Communications of the ACM*, 16(6):370–372, 1973.
- [6] P. Bonzini et al. GNU Lightning library. <http://www.gnu.org/software/lightning/>, 2010.
- [7] D. Bornstein. Dalvik VM Internals, May 2008. <http://sites.google.com/site/io/dalvik-vm-internals/>.
- [8] K. Casey, M. A. Ertl, and D. Gregg. Optimizing indirect branch prediction accuracy in virtual machine interpreters. *ACM Trans. Program. Lang. Syst.*, 29, October 2007.
- [9] J. Cuthbertson, S. Viswanathan, K. Bobrovsky, A. Astapchuk, and E. K. U. Srinivasan. A Practical Approach to Hardware Performance Monitoring Based Dynamic Optimizations in a Production JVM. In *CGO '09: Proceedings of the 7th annual IEEE/ACM International Symposium on Code Generation and Optimization*, pages 190–199, Washington, DC, USA, 2009. IEEE Computer Society.
- [10] ECMA International. *Standard ECMA-335 - Common Language Infrastructure (CLI)*, 4th edition, June 2006. <http://www.ecma-international.org/publications/standards/Ecma-335.htm>.
- [11] M. A. Ertl. A New Approach to Forth Native Code Generation. In *EuroForth '92 Conference Proceedings*, pages 73–78, Southampton, England, 1992.
- [12] M. A. Ertl. A portable Forth engine. In *EuroForth '93 Conference Proceedings*, Mariánské Lázně (Marienbad), 1993.
- [13] M. A. Ertl. *Implementation of Stack-Based Languages on Register Machines*. PhD thesis, Technische Universität Wien, Austria, 1996.

- [14] M. A. Ertl and D. Gregg. The Behavior of Efficient Virtual Machine Interpreters on Modern Architectures. In *Proceedings of the 7th International Euro-Par Conference Manchester on Parallel Processing*, Euro-Par '01, pages 403–412, London, UK, 2001. Springer-Verlag.
- [15] M. A. Ertl and D. Gregg. The Structure and Performance of Efficient Interpreters. *The Journal of Instruction-Level Parallelism*, 5:1–25, Nov 2003.
- [16] M. A. Ertl and C. Pirker. The Structure of a Forth Native Code Compiler. In *EuroForth '97 Conference Proceedings*, pages 107–116, Oxford, 1997.
- [17] M. A. Ertl and C. Pirker. Compilation of Stack-Based Languages. Final report to FWF for research project P11231, Institut für Computersprachen, Technische Universität Wien, 1998.
- [18] R. Ierusalimsky, L. H. de Figueiredo, and W. Celes. The evolution of lua. In *HOPL III: Proceedings of the 3rd ACM SIGPLAN Conference on History of Programming Languages*. ACM Press, 2007.
- [19] R. Ierusalimsky, L. H. de Figueiredo, L. Henrique, F. Waldemar, and W. C. Filho. Lua - an Extensible Extension Language, 1996.
- [20] Intel Corporation. *Intel 64 and IA-32 Architectures Software Developer's Manual Volume 1: Basic Architecture*, June 2010. <http://www.intel.com/products/processor/manuals/>.
- [21] P. Kobalick. AsmJit. <http://code.google.com/p/asmjit/>, 2010.
- [22] C. Lattner. LLVM: An Infrastructure for Multi-Stage Optimization. Master's thesis, Computer Science Dept., University of Illinois at Urbana-Champaign, Urbana, IL, Dec 2002.
- [23] C. Lattner and V. Adve. LLVM: A Compilation Framework for Lifelong Program Analysis & Transformation. In *Proceedings of the 2004 International Symposium on Code Generation and Optimization (CGO'04)*, Palo Alto, California, Mar 2004.
- [24] X. Leroy. The ZINC Experiment: An Economical Implementation of the ML Language. Technical Report 117, INRIA, Feb 1990.
- [25] X. Leroy, D. Doligez, A. Frisch, J. Garrigue, D. Remy, J. Vouillon, et al. The OCAML language (version 3.12.0). <http://caml.inria.fr/ocaml/>, 2010.
- [26] X. Leroy et al. The Caml Light language (version 0.75). <http://caml.inria.fr/caml-light/>, 2002.
- [27] T. Lindholm and F. Yellin. *The JavaTM Virtual Machine Specification*. Prentice Hall PTR, 2nd edition, April 1999.

- [28] M. Matz, J. Hubicka, A. Jaeger, and M. Mitchell. *System V Application Binary Interface: AMD64 Architecture Processor Supplement (Draft Version 0.99.5)*, Sep 2010. <http://www.x86-64.org/documentation.html>.
- [29] M. Pall. LuaJIT, 2010. <http://luajit.org/>.
- [30] D. Remy. Using, Understanding, and Unraveling the OCaml Language From Theory to Practice and Vice Versa. In G. Barthe et al., editors, *Applied Semantics*, volume 2395 of *Lecture Notes in Computer Science*, pages 413–537, Berlin, 2002. Springer-Verlag.
- [31] B. Starynkevitch. OCAMLJIT a faster Just-In-Time Ocaml Implementation, 2004.
- [32] W. Taha, C. Calcagno, X. Leroy, E. Pizzi, E. Pasalic, J. L. Eckhardt, R. Kaiabachev, O. Kiselyov, et al. MetaOCaml - A compiled, type-safe, multi-stage programming language, 2006. <http://www.metaocaml.org/>.
- [33] K. Vaswani and Y. N. Srikant. Dynamic recompilation and profile-guided optimisations for a .NET JIT compiler. *IEEE Proceedings - Software*, 150(5):296–302, 2003.