Binary Translation

G. Lettieri

22 Oct. 2014

1 Introduction

The idea of *binary translation* is to first translate the guest code into the equivalent host code for the virtual machine, and then jump at the translated code. If the translated code is kept in a cache and reused whenever the the guest is trying to execute it again, the cost of decoding the guest instructions is thus amortized. Moreover, the translated code can be optimized during the translation, since our emulator now looks at more than a single guest instruction at a time. This strategy generally brings great speedups w.r.t. the simpler emulation we have already seen, where each guest instruction is fetched, decoded and emulated in isolation, and this is done every time the guest tries to execute it. Apart from this, our emulator is again a normal, unprivileged program running on the host system, relying on the host operating system for the management of its resources. What we are doing is simply to replace the CPU loop with a more sophisticated one. In particular, the considerations about I/O, virtual memory and multi-threading are essentially the same as before.

Typically, the translation of guest code is not performed all at once, but in smaller units called *Dynamic Basic Blocks* (DBB for short). A DBB starts with an instruction which is the target of a jump (including the jump to the entry point of the program) and includes all the instructions that follow, stopping immediately after the first branch or jump instruction. One reason for using DBBs is that it is otherwise very difficult to identify all the code in the guest memory, since code looks just like data. DBBs start at instructions that the emulated CPU is actually trying to fetch after a jump, and therefore we rest assured that the corresponding bytes in the emulated memory must be interpreted as an instruction. Moreover, as long as the fetched instruction is not a jump (and we don't need to execute the instruction to know this), we are sure that it is followed by another instruction, and so on, until we found a branch or a jump. At unconditional jumps we stop, because we don't know whether the bytes that follow them are for code or for data (the emulated CPU is not going to execute them, for what we now). At branches (conditional jumps) we also stop, because it is possible that one of the two branches may never be taken, and we don't know if this is actually the case. The bytes that live at a dead branch may not be code at all. DBBs allow us to only translate code that the guest CPU is actually going to execute.

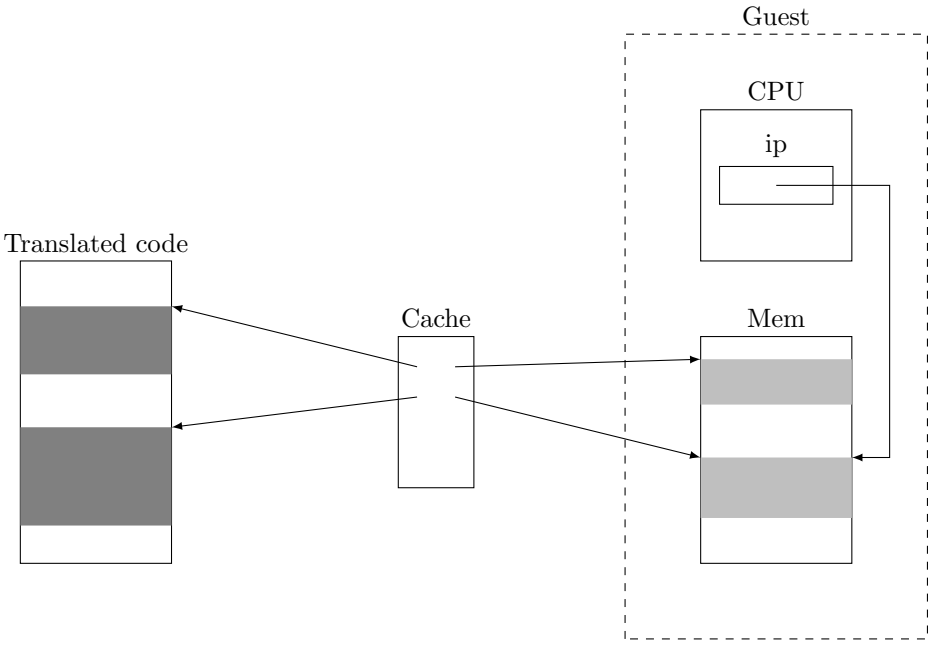

Figure 1: Data structures used in binary translation.

We call Translated Block (TB) the translation of a DBB. DBBs contain target instructions in the target memory (and in its emulation in the virtual machine), while TBs contain host instructions. A TB is identified by the guest address of the first instruction in the corresponding DBB so that, after the execution of a TB, we can use the current value of the guest instruction pointer to find the next TB to execute.

The CPU loop becomes something like

```
for (;;) {
        tb = find_in_cache(CPU->ip);
        if (!tb) {
                tb = translate(CPU->ip);
                add_to_cache(tb);
        }
        exec(tb, env);
}
```

Where CPU is the usual descriptor for the emulated CPU, env is some data structure containing all the information on the state of the guest (including the CPU, but also memory), and tb is a pointer to a TB descriptor.

Fig. 1 shows the main data structures used in binary translation. The guest system is represented by the usual data structures implementing the target

CPU and memory. The light gray areas in the Mem data structure represent dynamic basic blocks. The cache binds together each (currently translated) dynamic basic block with its translation (dark gray areas). Translations are kept in a memory area suitable for host execution (e.g., allocated via mmap()).

2 Optimizations

Recall from the first lecture the basic property of emulation/virtualization: we take snapshots of the guest and target states and guarantee that each guest snapshot is the representation of the corresponding target snapshot; on the other end, we do *not* guarantee any correspondence of guest and target states *between* two snapshots. In binary translation, snapshots are taken just before the execution of each dynamic basic block. During the execution of each basic block we are free to rearrange, omit and otherwise optimize the operations, provided that we always obtain the correct snapshot in the end.

There are several optimizations that the binary translator can use during the translation of a dynamic basic block. Here we mention only the most important.

2.1 Constant propagation

Constant propagation consists in replacing a register or memory operand with an immediate operand. This can be done whenever the content of the register or memory operand is known to the optimizer. E.g.,

```
movl $0, %eax
...
incl %eax
movl %eax, %ebx
```

If there are not other updates to **%eax** between the inizialization and the increment, then the optimizer can replace the last instruction with **movl** $1, **%ebx**.

2.2 Dead code elimination

This optimization consists in removing (host) code that cannot affect the state. A typical example is the removal of the code that computes the condition codes (e.g., the contents of the EFLAGS register in IA32). Assume, for example, that the guest code contains something like

```
ADDL %EAX, %EBX
SUBL $1, %EBX
```

The flags computed by the first addition will be overwritten by the following subtraction, so there is no need to compute them.

Another example is the value of the guest instruction pointer register: this is updated after the execution of each guest instruction, but its value is only used

during a relative jump (to sum it to the jump offset) or a subroutine call (e.g., to store it on the stack). The translated code can avoid computing the value of the guest instruction pointer and only update it at the end of the dynamic basic block. In the example translation of the previous lecture, we already used this optimization: in that case, the guest instruction pointer was updated only at the end, by loading it from the guest stack during the emulation of the **RET** instruction.

2.3 Register allocation

Instead of repeatedly reading and storing the guest registers from the guest CPU data structure in memory, we can allocate some host register to store the content of a guest register for the entire duration of a dynamic basic block. At the beginning of the translated block we load the contents of the guest register into the selected host register, we use the host register for all the translation, then we store the value of the host register back into the guest register at the end of the block. Note that registers are generally faster than memory, even than fist level caches. A modern CPU may need 4 cycles to access its cache, but registers are always read and written in one cycle.

There is no need to permanently map a guest register to the same host register. The mapping can be optimized on a per-block basis. However, imporant registers that are used very often in most blocks may be always kept in the same host register, thus saving the loads and stores at the beginning of each translated block. In IA32 this may be the case for the guest ESP register, for example.

2.4 Lazy condition code computation

Even if we can omit most condition code computations using dead code elimination, we still have to compute the final value of the flags register ad the end the block. This is because we do not know if the next block will need it. However, we can do better: instead of computing the flags, we store the operands and the result of the last guest instruction that would have updated them. We will use these values to compute the flags later, but only if we fetch a guest instruction that needs them. In most cases we will never compute the flags, since we will first fetch a guest instruction that overwrites them.

2.5 Translated block chaining

In the standard binary translation algorithm we return to the the main loop after the execution of each translated block; then, we use the current value of the guest CPU instruction pointer to look up the translation cache, eventually translating a new dynamic basic block in case of miss. Once we have the pointer to the new translation block, we can patch the previous block so that, the next time it will be executed, it will directly jump to the new block, instead of returning to the main loop. This can be done even if a block ends with a

conditional jump: the block will jump to one of two possible next translation blocks.

3 Problems

Depending on the target architecture, the binary translator also have to solve a set of complex problems, in order to correctly emulate the target states.

3.1 Handling interrupts

The target will check for interrupts after the execution of each instruction. The binary translator can emulate this by inserting the code that checks for interrupts after the translation of each guest instruction.

However, since there is generally no guarantee on the exact timing of interrupt arrival, we can adopt a more efficient strategy and only check for interrupts whenever the host returns to the main loop. This is acceptable if basic block chaining is not used, since each translation block will contain no loops. If chaining is used, however, the emulator may spend an unbounded amount of time in the exec() function, without returning to the main loop. In this case, chaining must be disabled whenever there is a pending interrupt This is essentially equivalent to putting the interrupt checking code at the end of each translation block, before the jump to the next block.

Note that this optimization may have a confusing effect: since we are artificially disabling interrupts for the duration of a dynamic basic block, we may be hiding synchronization errors in the guest code.

3.2 Handling faults

The problem with faults is that they may occur in the middle of a dynamic basic block and, unlike interrupts, they typically cannot be delayed. We can implement this with a lonjmp() to the main loop, where the corresponding setjmp() will then read the guest interrupt table and change the guest instruction pointer accordingly. The problem, however, is that the guest fault handler may need the contents of all guest registers *at the time of the fault*, while the translated code has been optimized assuming that the guest state was needed only at the end of each block. Since we certainly do not want to remove optimizations, we need to reconstruct the exact state of the guest whenever a fault is generated. This is still good performance-wise, since faults are rare.

The complexity of the guest state reconstruction depends on the optimizations that have been used during the translation of the block.

- If we have used register allocation, we can store a table with the mapping between guest and host registers together with each translated block. If a fault is generated during the execution of the block, we can use the table to update the contents of the guest registers before jumping to the guest fault handler.

- We have to consider possible faults when we apply dead code elimination. Essentially, faults may insert code between two instructions, thus creating users for otherwise dead values. If we can identify before hand all the instructions that may cause a fault (e.g., division by zero can only be generated by a division), then we can make sure that the guest state is updated before the execution of any such instruction.

- In the most complex cases we may need to return to a known consistent state and *switch to emulation* until we hit the fault again. By "switch to emulation" we mean that our binary translator must also include an emulator (like the one we studied in the first lectures) and, when needed, it must be able to temporarily stop translating dynamic basic blocks and start fetching and executing one guest instruction at a time.

3.3 Virtual memory

In the example translation of the previous lecture we also omitted to consider the guest virtual memory. In a complete translation, we need to translate every guest address before using it to access the guest memory. This can be done by inserting the code for the translation in all the guest instructions that have memory operands. The code to be added is essentially the same code that we already considered when we talked about virtual memory in emulators, and can benefit from the same optimizations (e.g., a software TLB). If a page fault needs to be generated, we need to apply all the considerations that we have already discussed for the general fault case.

3.4 Self modifying code

The most complex problem is guest code that modifies itself. Assume first that code in one block B_1 modifies code in another block B_2. If we have already translated B_2, we need to invalidate the translation. First, we have to detect the fact that the B_2 has been altered. Since code is just data somewhere in guest memory, *any* guest write to memory may modify code. We have two main tecniques to detect code modification:

- (software only) remember in some data structure the range of guest addresses containing code that we have already translated and look up the destination address before each guest memory write (e.g., we can keep a bitmap with one bit for each page);

- (with the help of the host hardware) write-protect the guest memory that contains code already translated, then invalidate the relavant cache entries on page fault.

For the latter teqnique, remember that the binary translator is generally an unprivileged program running on some OS, so it does not have direct access to page tables and it cannot directly intercept page faults. However, in Unix-like system, we can proceed as follows:

1. Allocate also the guest memory using mmap() (this is necessary to be able to use mprotect() below);

2. use the signal () system call to intercept the SIGSEGV signal;

3. whenever we translate a block, use mprotect() to ask the kernel to remove the PROT_WRITE permission from the corresponding pages in guest memory;

Now, when our process will try to write into the protected pages, the kernel will send it a SIGSEGV signal. Normally, this signal causes the abnormal termination of the process, returning to the shell which then prints "Segmentation fault". However, in step 2, we have asked the kernel that we want to handle SIGSEGV by our own. All we need to do is to call signal () with the number of the signal we are instered in (i.e., SIGSEGV) and a pointer to a function of our own. Whenever the process receives the selected signal, it executes our function instead of terminating. We can use this function to invalidate the translation cache.

The most complex case occurs when block B tries to modify *itself*. Now we cannot simply invalidate the translated block, since it is the very code we are executing. Probably the best thing we can do is to switch to emulation as soon as we detect the write, and continue to emulate one instruction at a time until we get out of the block.