

Interpreters and virtual machines

Advanced Compiler Construction
Michel Schinz – 2024-04-25

Interpreters

Interpreters

An **interpreter** is a program that executes another program, which could be represented as:

- raw text (source code), or
- a tree (AST of the program), or
- a linear sequence of instructions.

Pros of interpreters:

- no need to compile to native code,
- simplify the implementation of programming languages,
- often fast enough on modern CPUs.

Text-based interpreters

Text-based interpreters directly interpret the textual source of the program. Seldom used, except for trivial languages where every expression is evaluated at most once (no loops/functions). Plausible example: a calculator, evaluating arithmetic expressions while parsing them.

Tree-based interpreters

Tree-based interpreters walk over the abstract syntax tree of the program to interpret it.

Better than string-based interpreters since parsing and analysis is done only once.

Plausible example: a graphing program, which repeatedly evaluates a function supplied by the user to plot it.

(Also, all the interpreters included in the L₃ compiler are tree-based.)

Virtual machines

Virtual machines

Virtual machines resemble real processors, but are implemented in software. They take as input a sequence of instructions, and often also abstract the system by:

- managing memory,
- managing threads,
- managing I/O,
- etc.

Used in the implementation of many important languages, e.g. SmallTalk, Lisp, Forth, Pascal, Java, C#, etc.

Why virtual machines?

Since the compiler has to generate code for some machine, why prefer a virtual over a real one?

- for portability: compiled VM code can be run on many actual machines,
- for simplicity: a VM is usually more high-level than a real machine, which simplifies the task of the compiler,
- for simplicity (2): a VM is easier to monitor and profile, which eases debugging.

Virtual machines drawbacks

Virtual machines have one drawback: performance.

Why?

- interpretation overhead (fetching/decoding, etc.).

Mitigations:

- compile the (hot parts) of the program being interpreted,
- adapt optimization on program behavior.

Kinds of virtual machines

Two broad kinds of virtual machines:

- **stack-based VMs** use a stack to store intermediate results, variables, etc.
- **register-based VMs** use a limited set of registers for that, like a real CPU.

What's best?

- for compiler writers: stack-based is easier (no register allocation),
- for performance: register-based *can* be better.

Most widely-used virtual machines today are stack-based (e.g. the JVM, .NET's CLR, etc.) but a few recent ones are register-based (e.g. Lua 5.0).

Virtual machine input

Virtual machines take as input a program expressed as a sequence of instructions:

- each instruction is identified by its **opcode** (**operation code**), a simple number,
- when opcodes are one byte, they are often called **byte codes**,
- additional arguments (e.g. target of jump) appear after the opcode in the stream.

VM implementation

Virtual machines are implemented in much the same way as a real processor:

1. the next instruction to execute is fetched from memory and decoded,
2. the operands are fetched, the result computed, and the state updated,
3. the process is repeated.

VM implementation

Which language are used to implement VMs?

Today, often C or C++ as these languages are:

- fast,
- at the right abstraction level,
- relatively portable.

Moreover, GCC and clang have an extension that can be used to speed-up interpreters.

Implementing a VM in C

```
typedef enum {
    add, /* ... */
} instruction_t;

void interpret() {
    static instruction_t program[] = { add /* ... */ };
    instruction_t* pc = program;
    int* sp = ...; /* stack pointer */
    for (;;) {
        switch (*pc++) {
            case add:
                sp[1] += sp[0];
                sp++;
                break;
                /* ... other instructions */
            }
        }
    }
}
```

Optimizing VMs

The basic, switch-based implementation of a virtual machine just presented can be made faster using several techniques:

- threaded code,
- top of stack caching,
- super-instructions,
- JIT compilation.

Threaded code

Threaded code

In a `switch`-based interpreter, two jumps per instruction:

- one to the branch handling the current instruction,
- one from there back to the main loop.

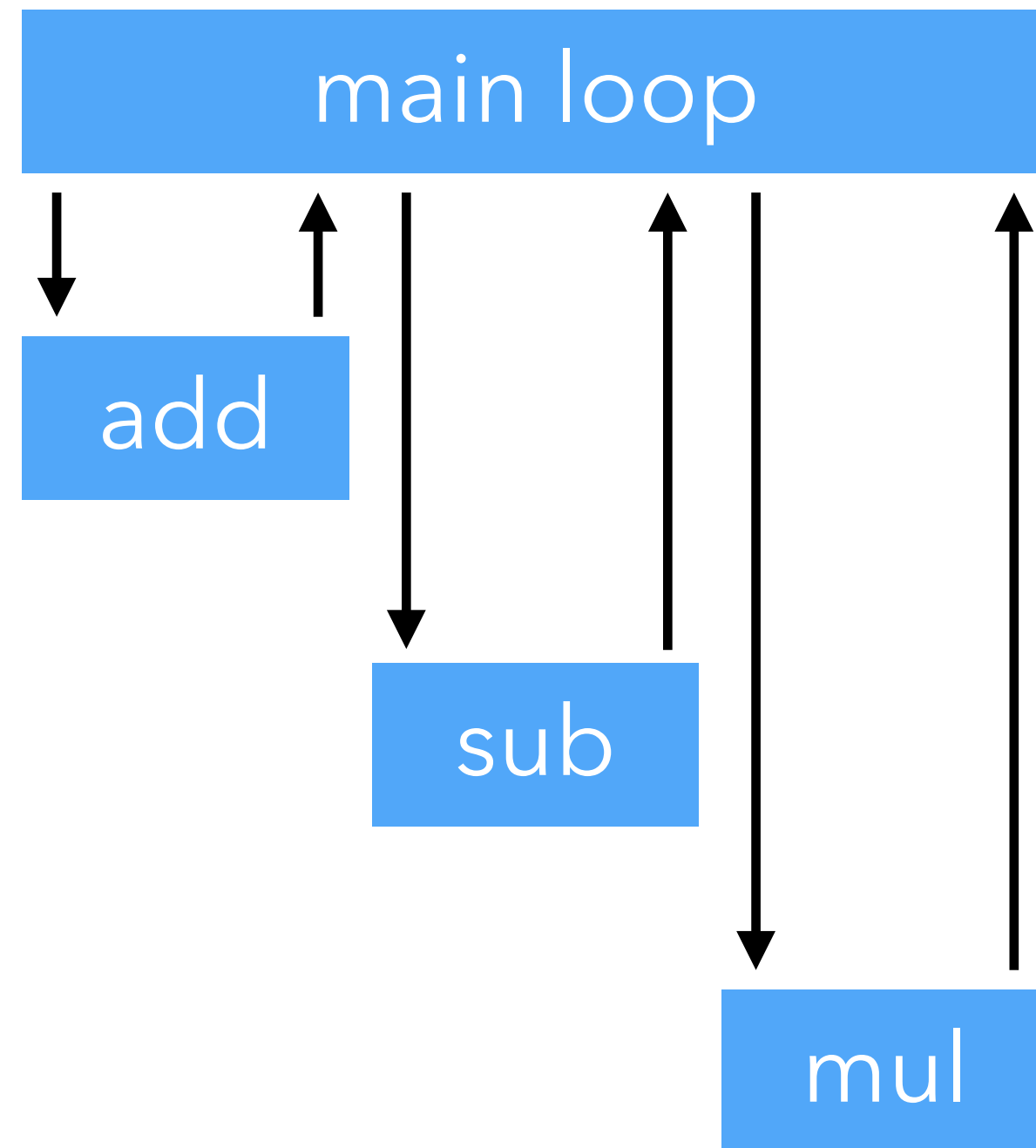
The second one should be avoided, by jumping directly to the code handling the next instruction.

This is the idea of **threaded code**.

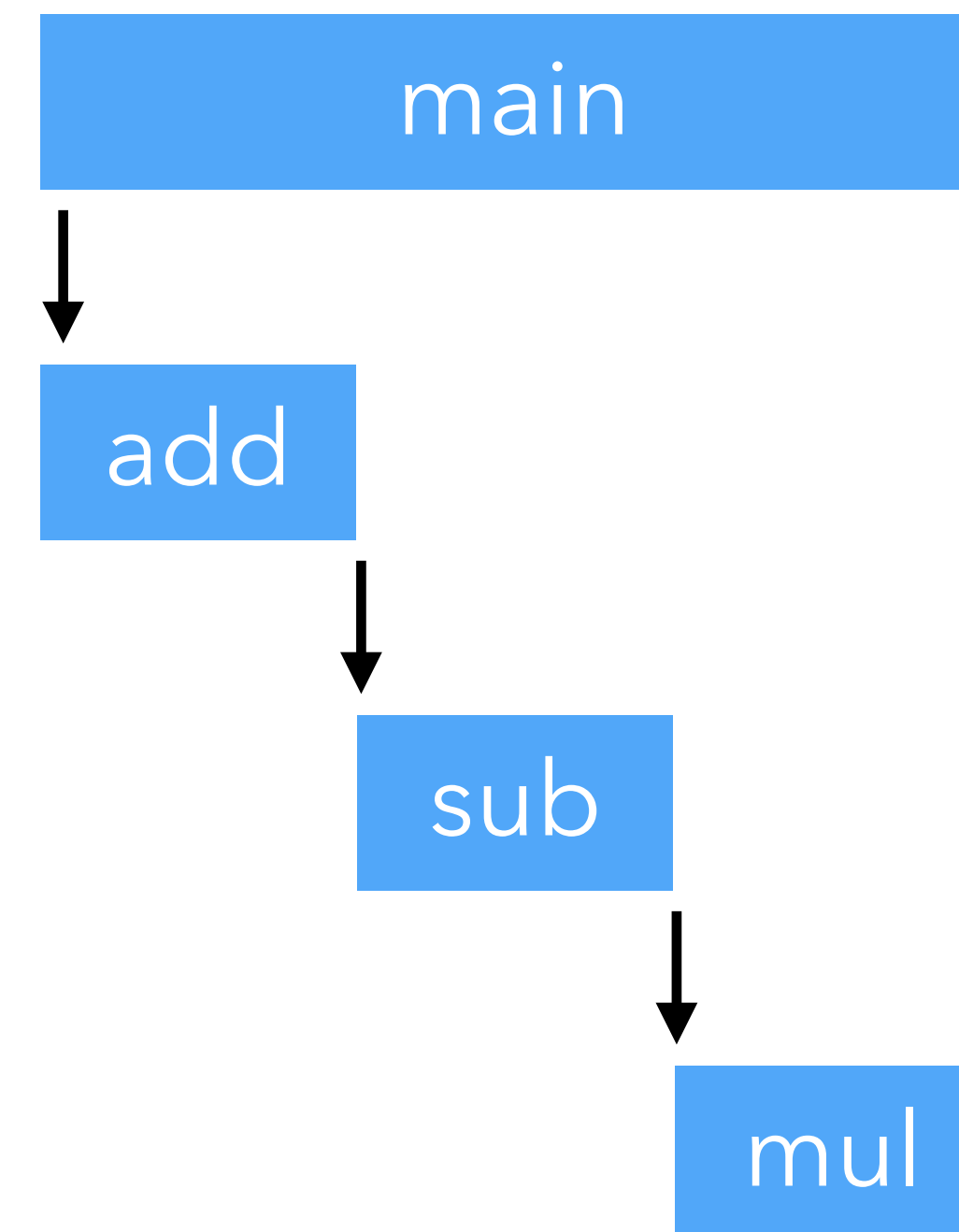
Switch vs threaded

Program: add sub mul

switch-based



Threaded



Implementing threaded code

Two main variants of threading:

1. **indirect threading**, where instructions index an array containing pointers to the code handling them,
2. **direct threading**, where instructions are pointers to the code handling them.

Pros and cons:

- direct threading has one less indirection,
- direct threading is expensive on 64 bits architectures (one opcode = 64 bits).

Threaded code in C

Threaded code represents instructions using code pointers.

How can this be done in C?

- in standard (ANSI) C, with function pointers (requires tail-call elimination),
- with GCC or clang, with label pointers (does not require tail-call elimination).

Direct threading in ANSI C

Direct threading in ANSI C:

- one function per VM instruction,
- the program is a sequence of function pointers,
- each function ends with code to handle the next instruction.

Easy but requires tail-call elimination!

Direct threading in ANSI C

```
typedef void (*instruction_t)(void*, int*);

static void add(void* pc0, int* sp) {
    instruction_t* pc = pc0;
    sp[1] += sp[0];
    sp += 1;
    pc += 1;
    (*pc)(pc, sp); /* handle next instruction */
}

/* ... other instructions */

static instruction_t program[] = { add, /* ... */ };

void interpret() {
    int* sp = ...;
    instruction_t* pc = program;
    (*pc)(pc, sp); /* handle first instruction */
}
```

Direct threading in ANSI C

Major problem of direct threading in ANSI C:

- stack overflow in the absence of tail call elimination.

With compilers that do not do TCE, the only option is to use trampolines (or similar), which is very slow!

Conclusion: direct threading in ANSI C is only realistic with a compiler that eliminates tail calls.

Direct threading with GCC

Direct threading with GCC or clang:

- one *block* per VM instruction,
- the program is a sequence of *block* pointers,
- each function ends with code to handle the next instruction.

This requires a non-standard extension called *labels as values* (basically, label pointers).

Direct threading with GCC

label as value

```
void interpret() {  
    void* program[] = { &&l_add, /* ... */ };  
  
    int* sp = ...;  
    void** pc = program;  
    goto **pc; /* jump to first instruction */
```

computed goto

```
l_add:  
    sp[1] += sp[0];  
    ++sp;  
    goto **(++pc); /* jump to next instruction */  
  
/* ... other instructions */  
}
```

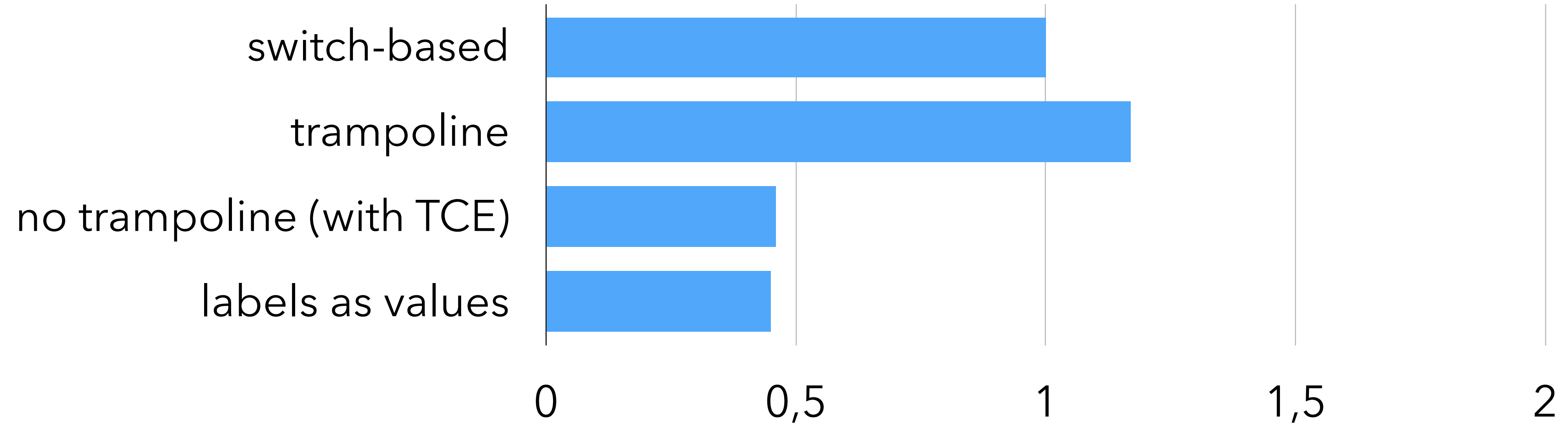
Threading benchmark

Benchmark: 500'000'000 iterations of a loop

Processor: 2.0 GHz Intel Core i5

Compiler: clang 13.1.6

Optimization settings: -O3



Top-of-stack caching

Top-of-stack caching

In a stack-based VM, the stack is typically represented as an array in memory, accessed by almost all instructions.

Idea:

- store topmost element(s) in registers.

However:

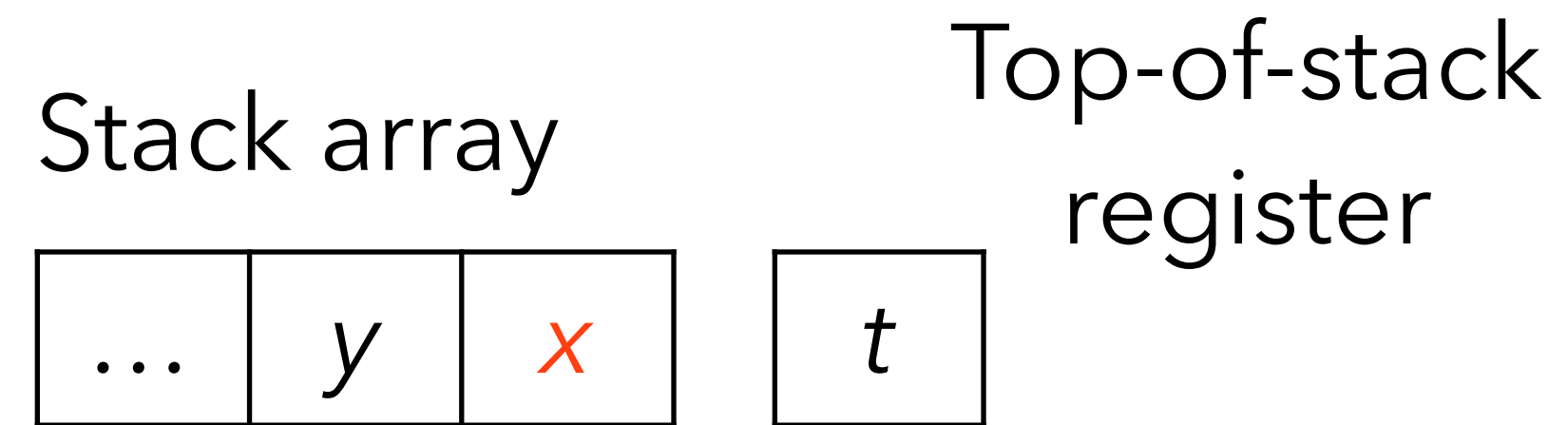
- storing a fixed number of topmost elements is not a good idea!

Therefore:

- store a variable number of topmost elements, e.g. at most one.

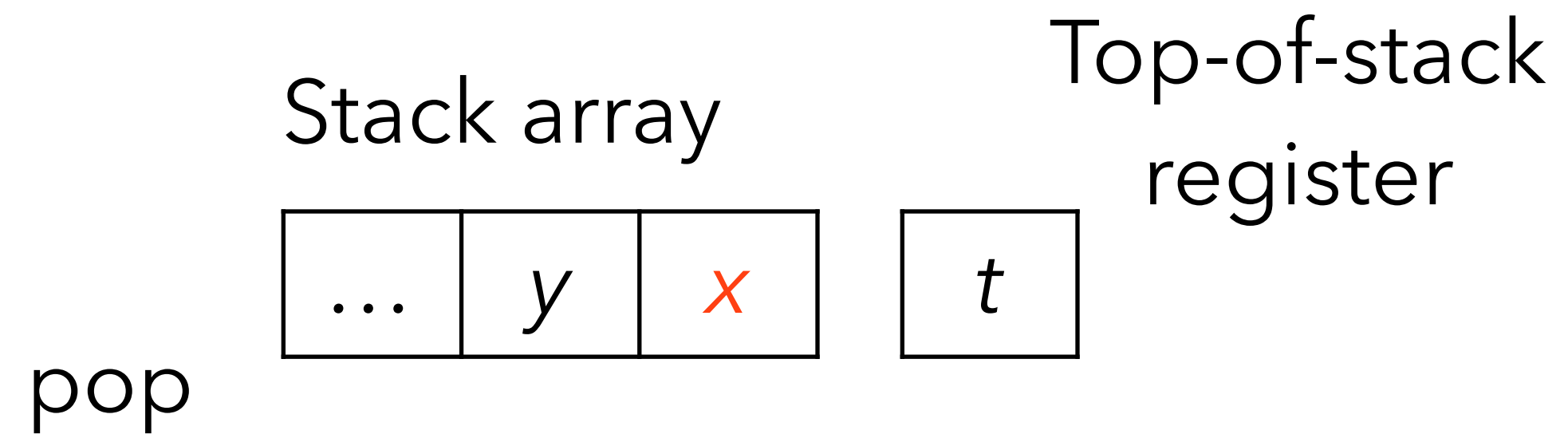
Top-of-stack caching

The top element is always cached:



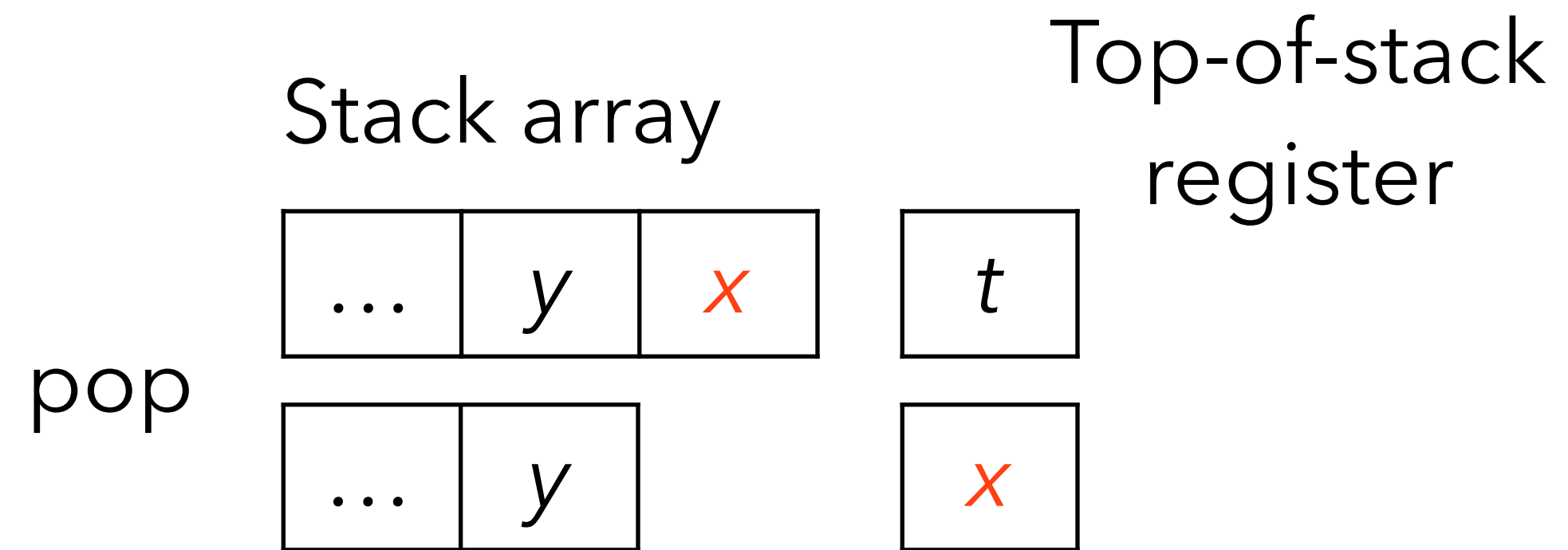
Top-of-stack caching

The top element is always cached:



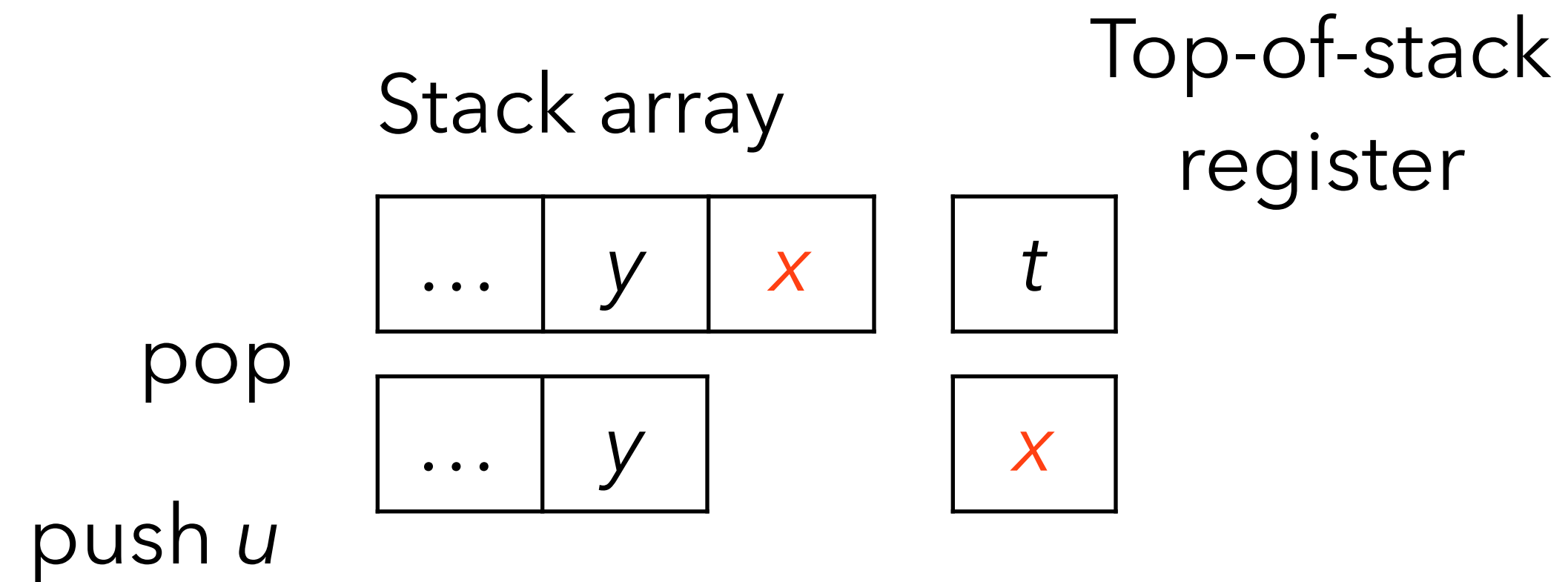
Top-of-stack caching

The top element is always cached:



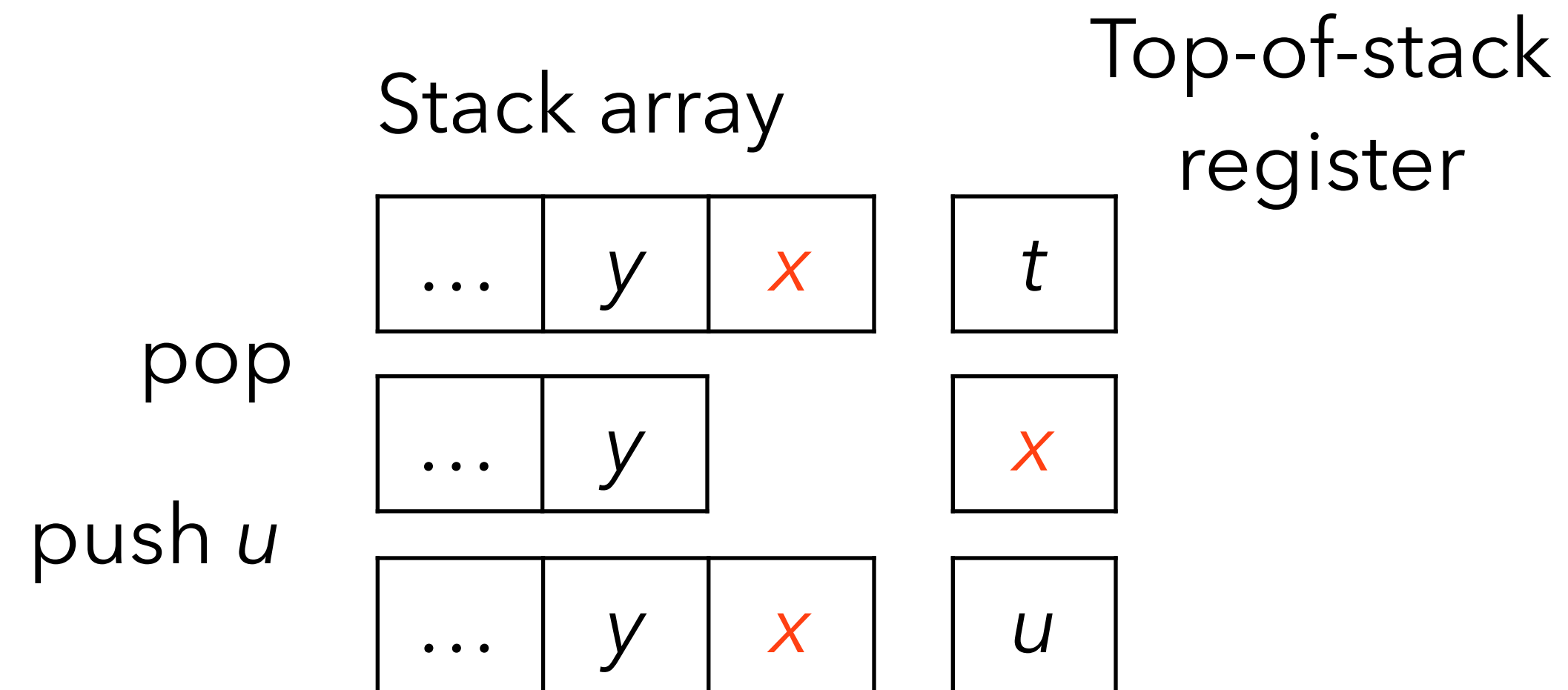
Top-of-stack caching

The top element is always cached:



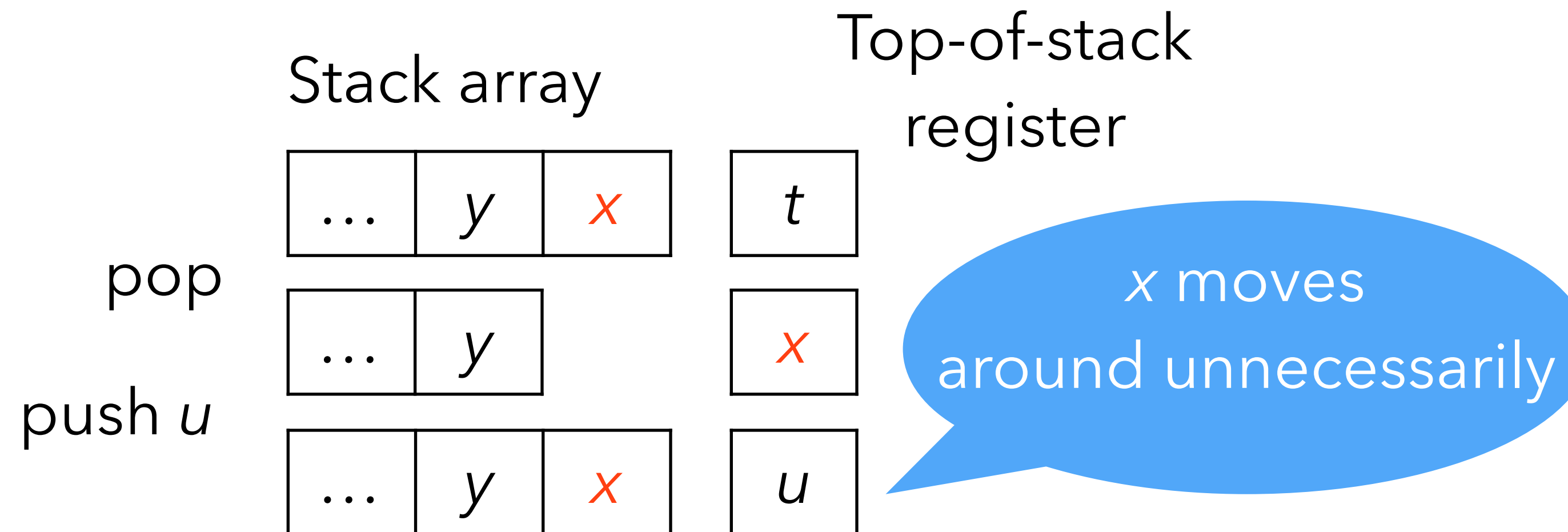
Top-of-stack caching

The top element is always cached:



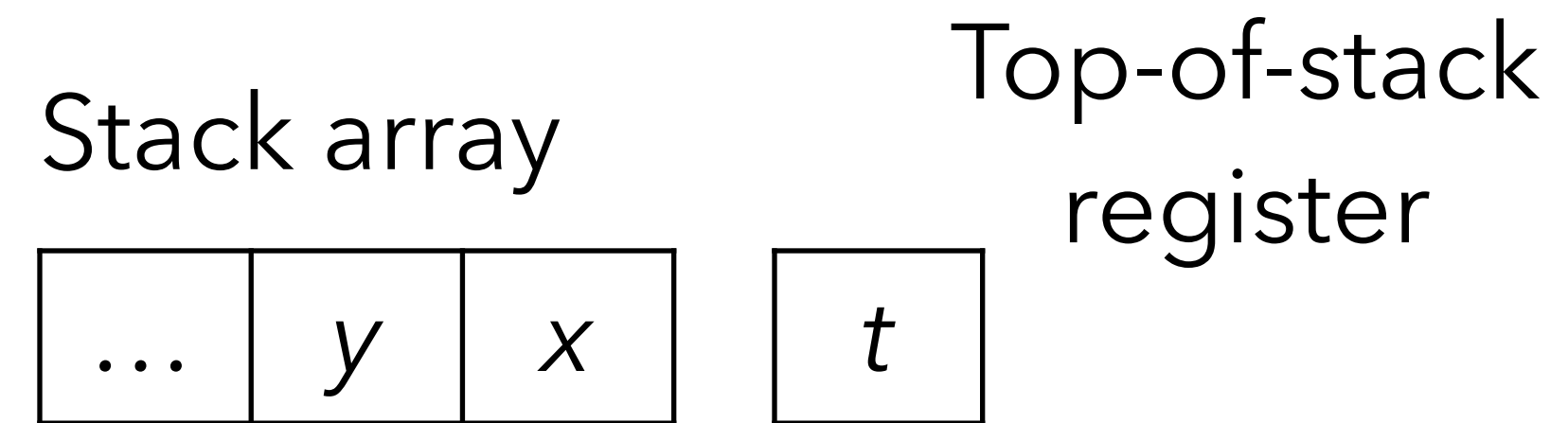
Top-of-stack caching

The top element is always cached:



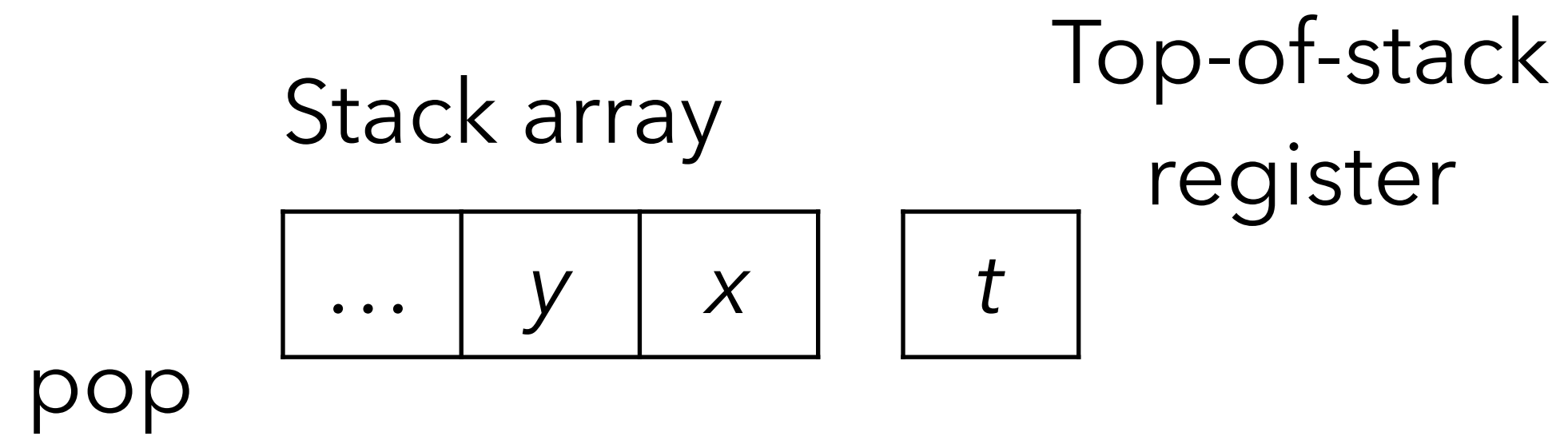
Top-of-stack caching

Either 0 or 1 top-of-stack element is cached:



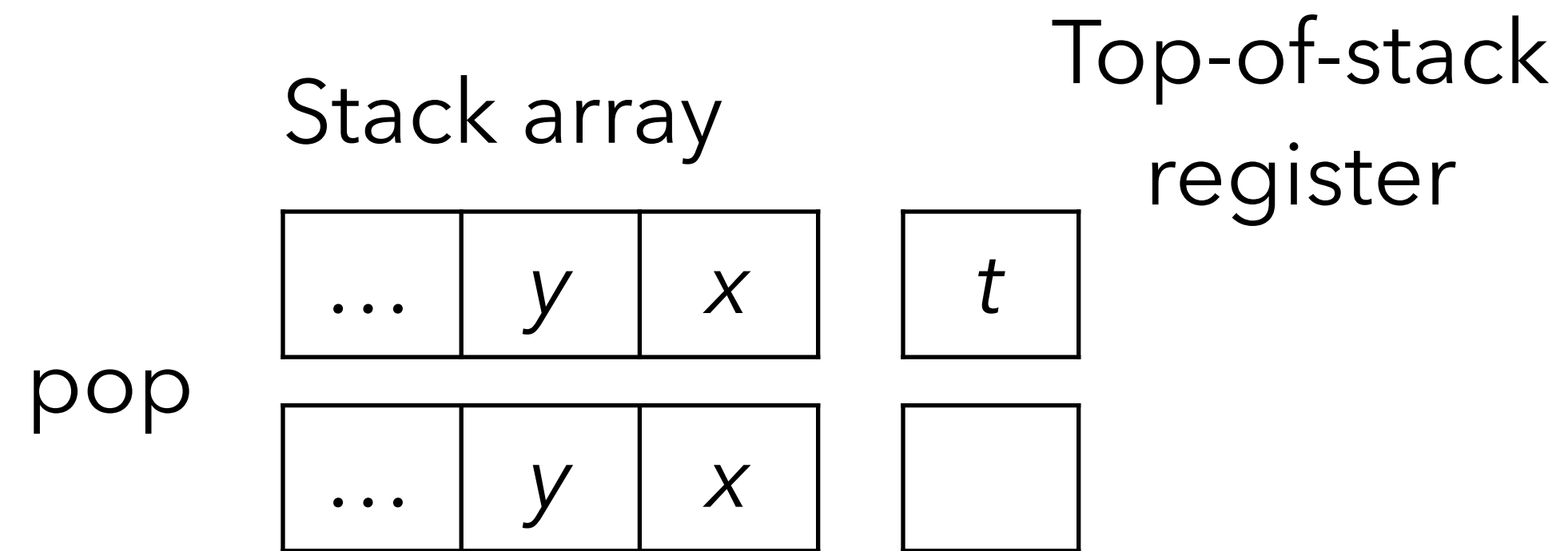
Top-of-stack caching

Either 0 or 1 top-of-stack element is cached:



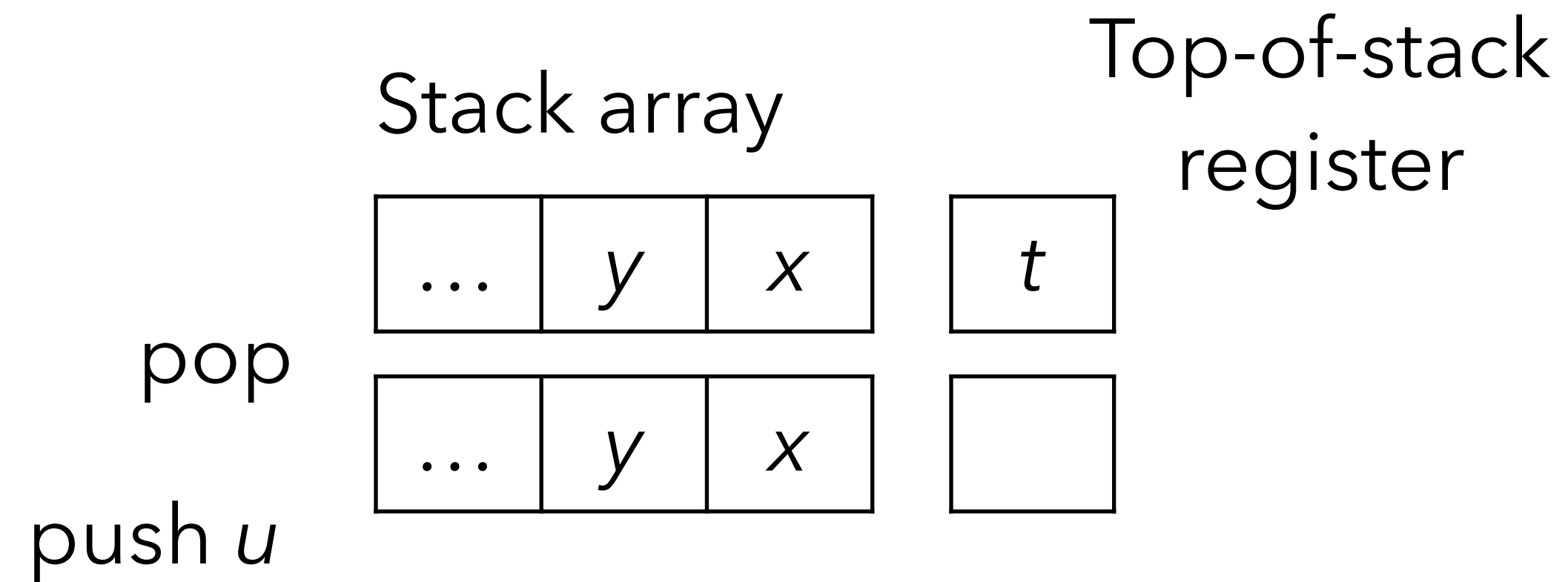
Top-of-stack caching

Either 0 or 1 top-of-stack element is cached:



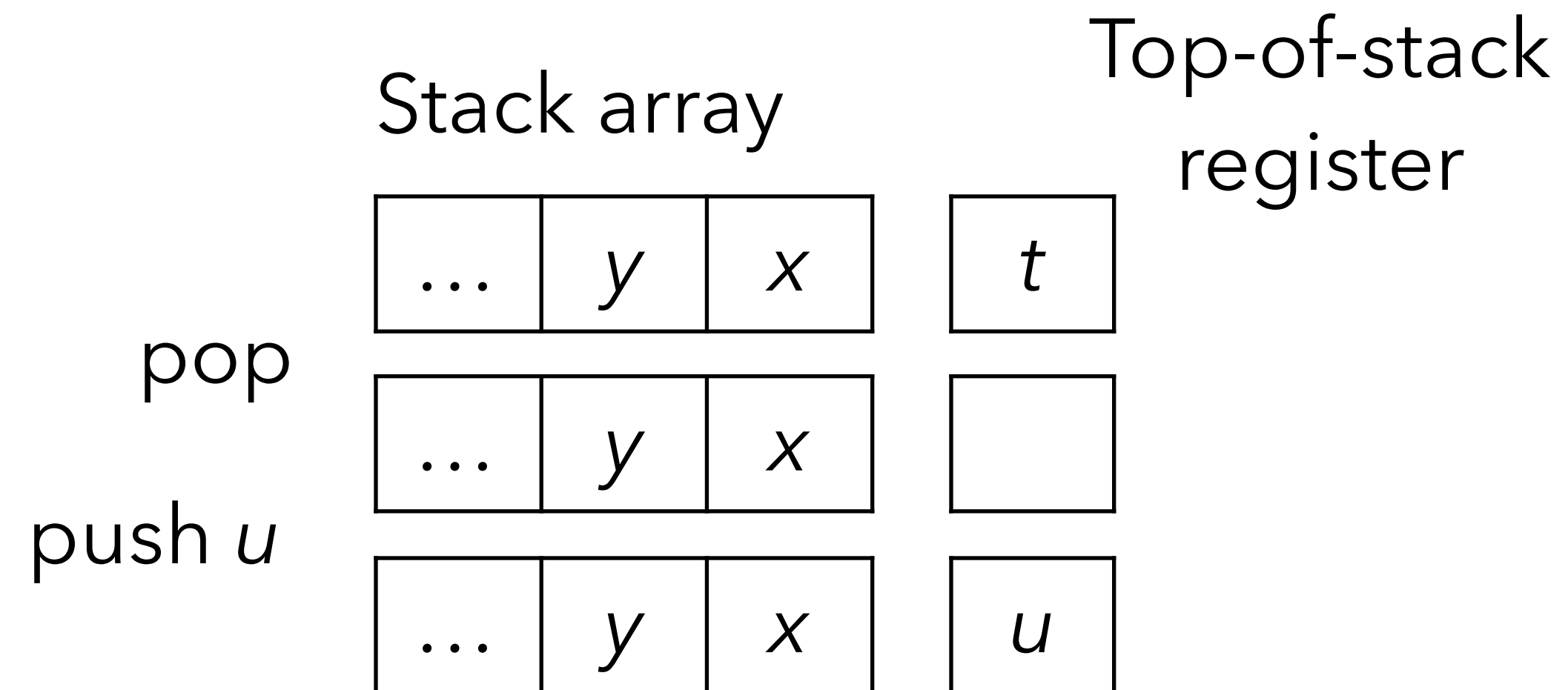
Top-of-stack caching

Either 0 or 1 top-of-stack element is cached:



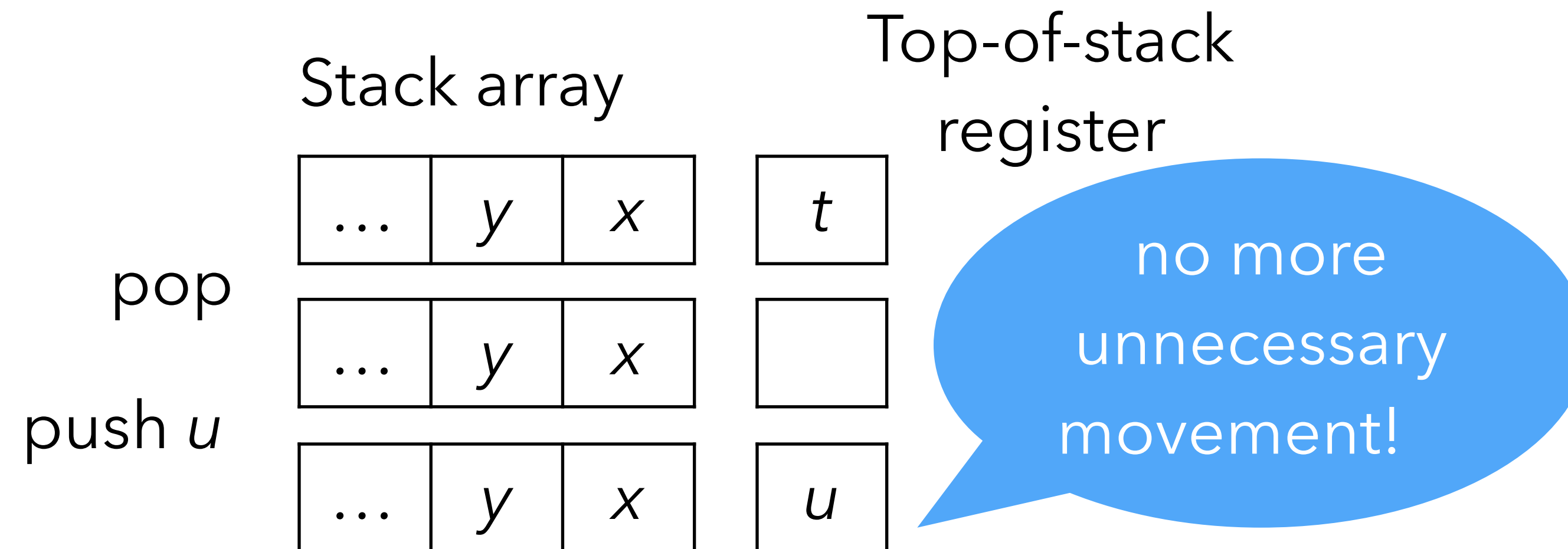
Top-of-stack caching

Either 0 or 1 top-of-stack element is cached:



Top-of-stack caching

Either 0 or 1 top-of-stack element is cached:



Top-of-stack caching

Beware: caching a variable number of stack elements means that every instruction must have one implementation per **cache state** (number of stack elements currently cached)

E.g., when caching at most one stack element, the add instruction needs the following two implementations:

State 0: no elements in reg.

```
add_0:
    tos = sp[0]+sp[1];
    sp += 2;
    // go to state 1
```

State 1: top-of-stack in reg.

```
add_1:
    tos += sp[0];
    sp += 1;
    // stay in state 1
```

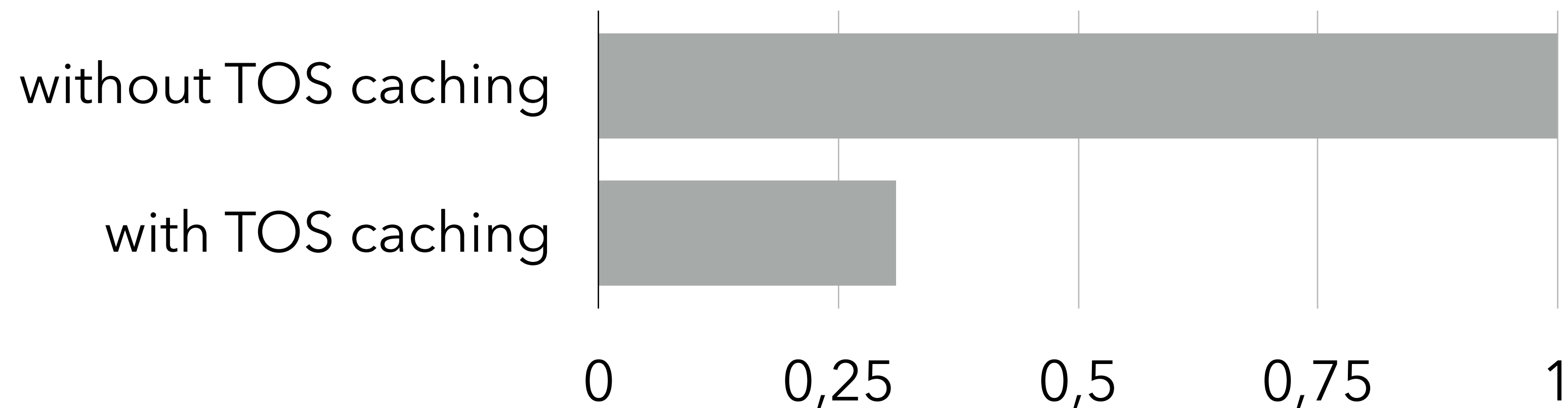
Benchmark

Benchmark: sum first 200'000'000 integers

Processor: 2.0 GHz Intel Core i5

Compiler: clang 13.1.6

Optimization settings: -O3



Super-instructions

Static super-instructions

Observation:

instruction dispatch is expensive in a VM.

Conclusion:

group several instructions into **super-instructions**.

Idea:

- use profiling to determine which sequences should be transformed into super-instructions,
- modify the the instruction set of the VM accordingly.

E.g., if mul, add appears often in sequence, combine the two in a single madd (multiply and add) super-instruction.

Dynamic super-instructions

Super-instructions can also be generated at run time, to adapt to the program being run.

This is the idea of **dynamic super-instructions**.

Pushed to its limits: generate one super-instruction per basic-block.

L₃VM

L₃VM

L₃VM is the VM of the L₃ project. Main characteristics:

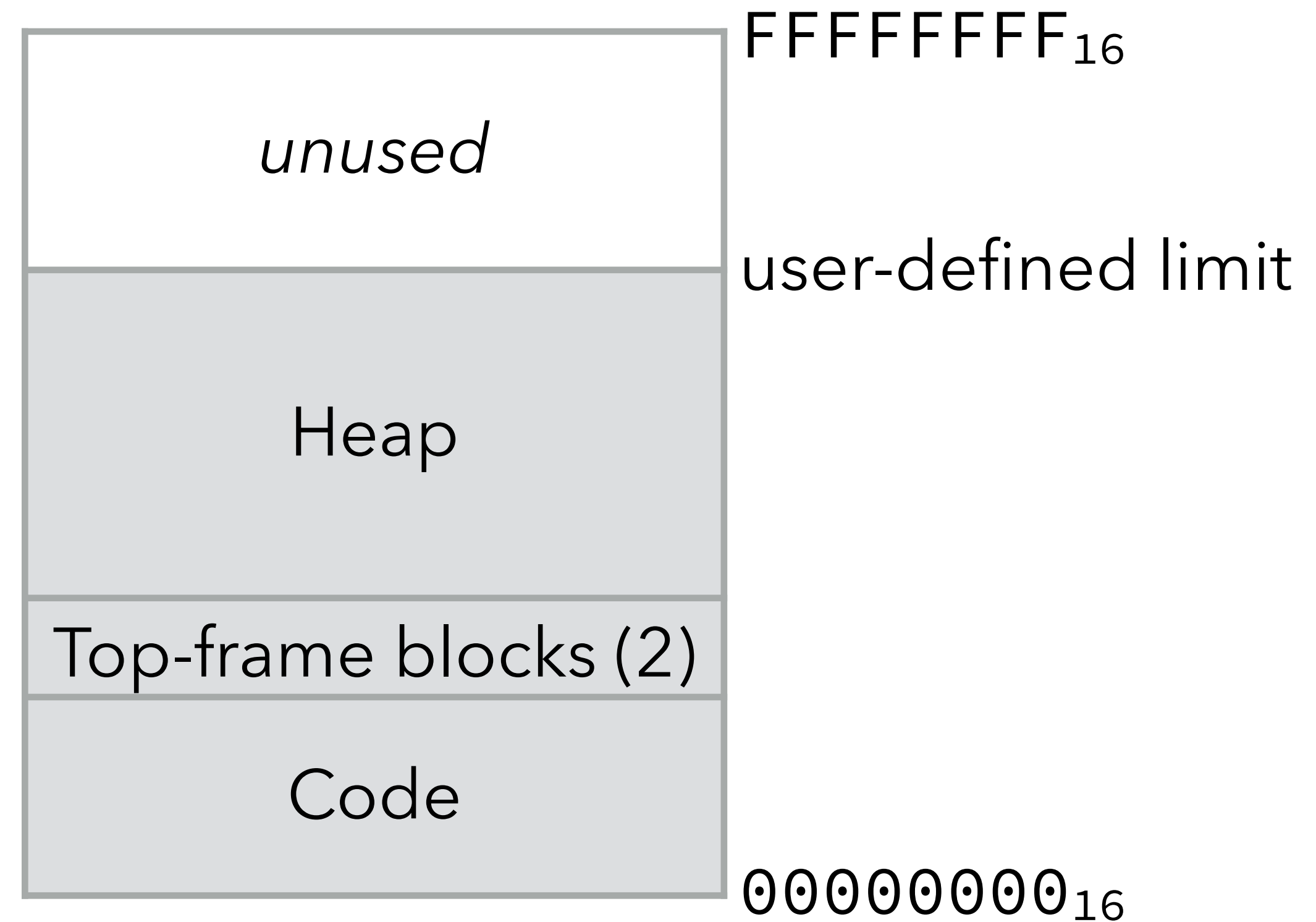
- it is a 32 bits VM:
 - (untagged) integers are 32 bits,
 - pointers are 32 bits,
 - instructions are 32 bits,
- it is register-based (with an unconventional notion of register),
- it is simple: only 32 instructions.

Memory

A single 32-bit address space is used to store code and heap.

Code is stored starting at address 0, the rest is used for the heap and the top-frame blocks.

(Note: L₃VM addresses are not the same as those of the host).



Registers

Strictly speaking, L₃VM has only two registers:

- the **program counter** (PC), containing the address of the instruction being executed,
- the **frame pointer** (FP), containing the address of the activation frame of the current function.

The frame of the current function always resides in one of the two **top-frame blocks**, so FP always points to one of them. Most of that block's slots contain the values manipulated by instructions and are therefore referred to as "registers".

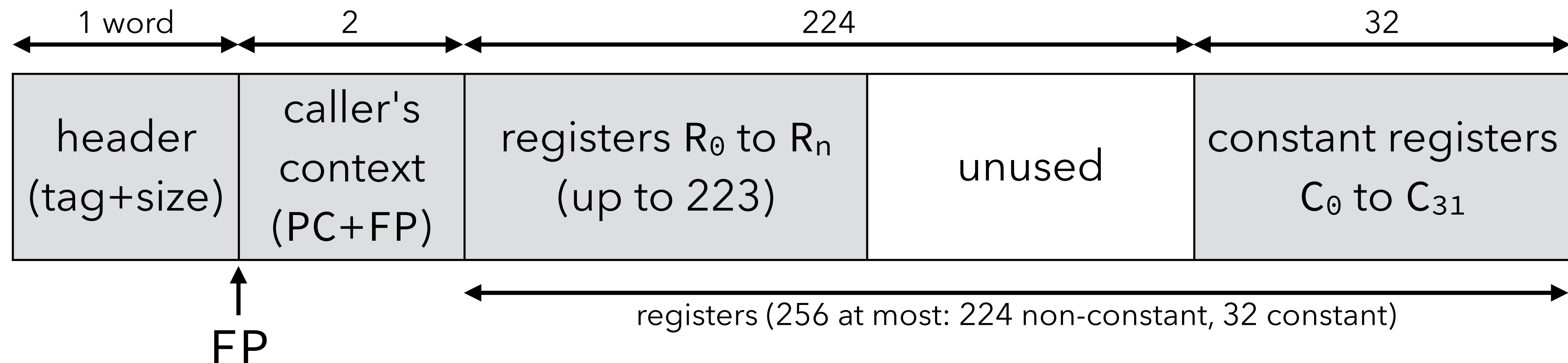
Top-frame blocks

One of the two top-frame blocks contains the frame of the current function.

The other contains either:

- nothing, or
- (some of) the arguments of a function about to be called, or
- the frame of the caller.

Both are laid out in memory as follows:



Non-tail call and return

When a function (the caller) wants to call another function (the callee), it:

- frees the other top-frame block to use it as the callee's frame – see later,
- stores the callee's arguments in the first register slots of its frame,
- does the actual call, which:
 - saves the PC/FP of the caller in slots 0/1 of callee's frame,
 - makes the PC point to the callee's first instruction,
 - makes the FP point to the callee's frame.

When a function wants to return, it:

- ensures that the frame of the caller is in one of the top-frame blocks,
- makes the PC point to the saved return address,
- makes the FP point to the caller's frame.

Top-frame eviction

If a function wants to call another function and the other top-frame block contains the frame of its own caller, then:

- it saves the caller's frame into a heap-allocated block,
- it adjusts its own pointer to it so that it refers to that block.

The other top-frame block is then free to be used to store the callee's frame.

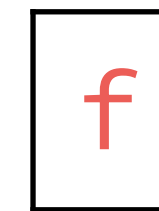
Consequently, during a return, the frame of the caller might have to be copied back from the heap to one of the top-frame blocks.

Non-tail calls and returns

top frames

heap

call stack

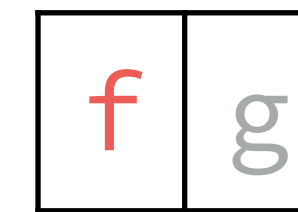
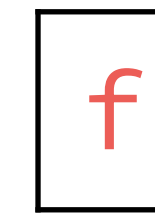


Non-tail calls and returns

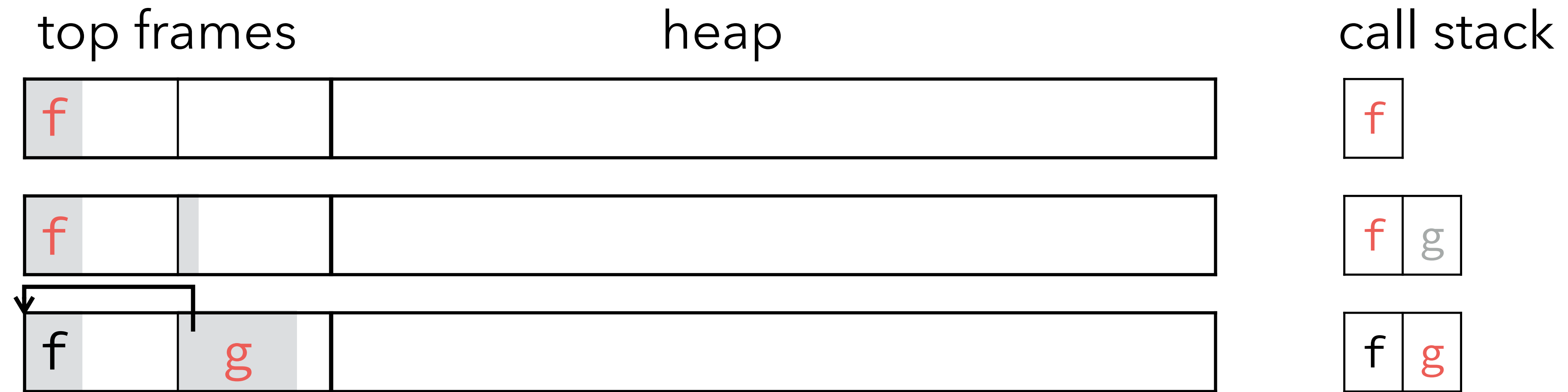
top frames

heap

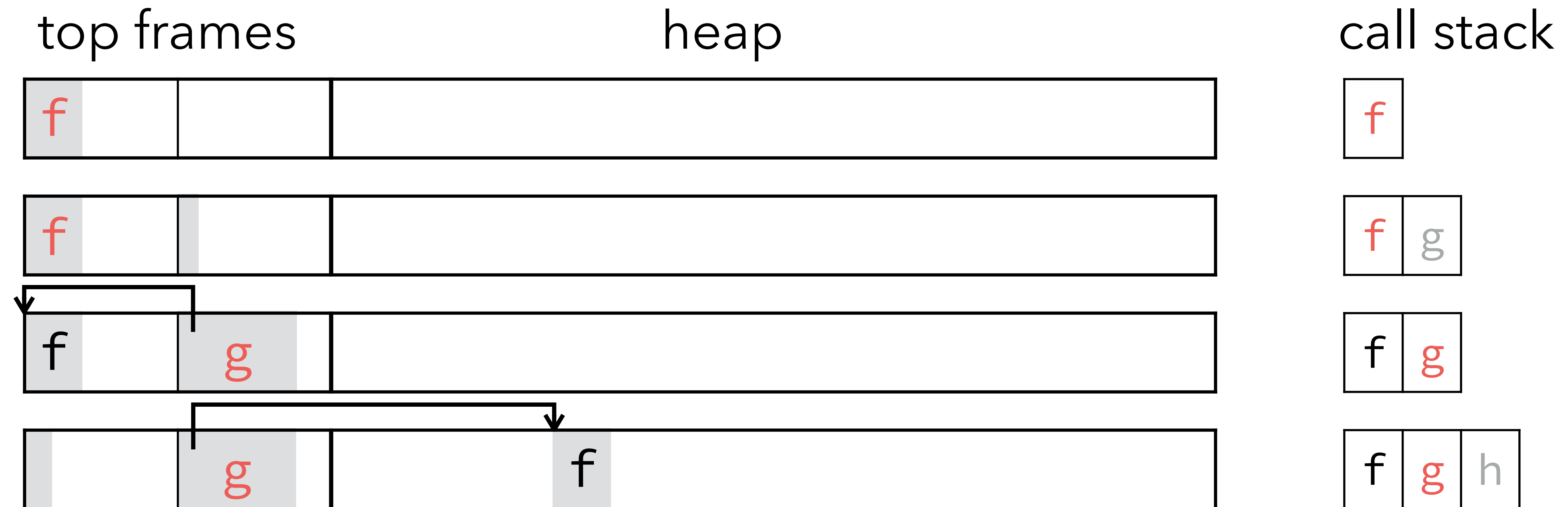
call stack



Non-tail calls and returns



Non-tail calls and returns

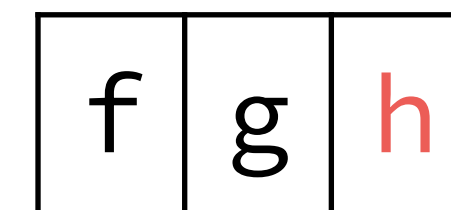
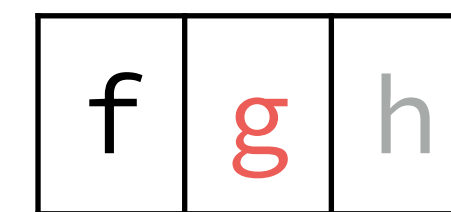
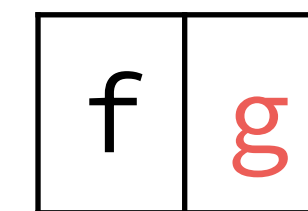
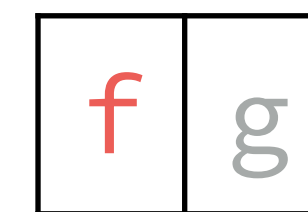
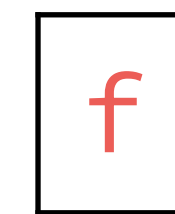
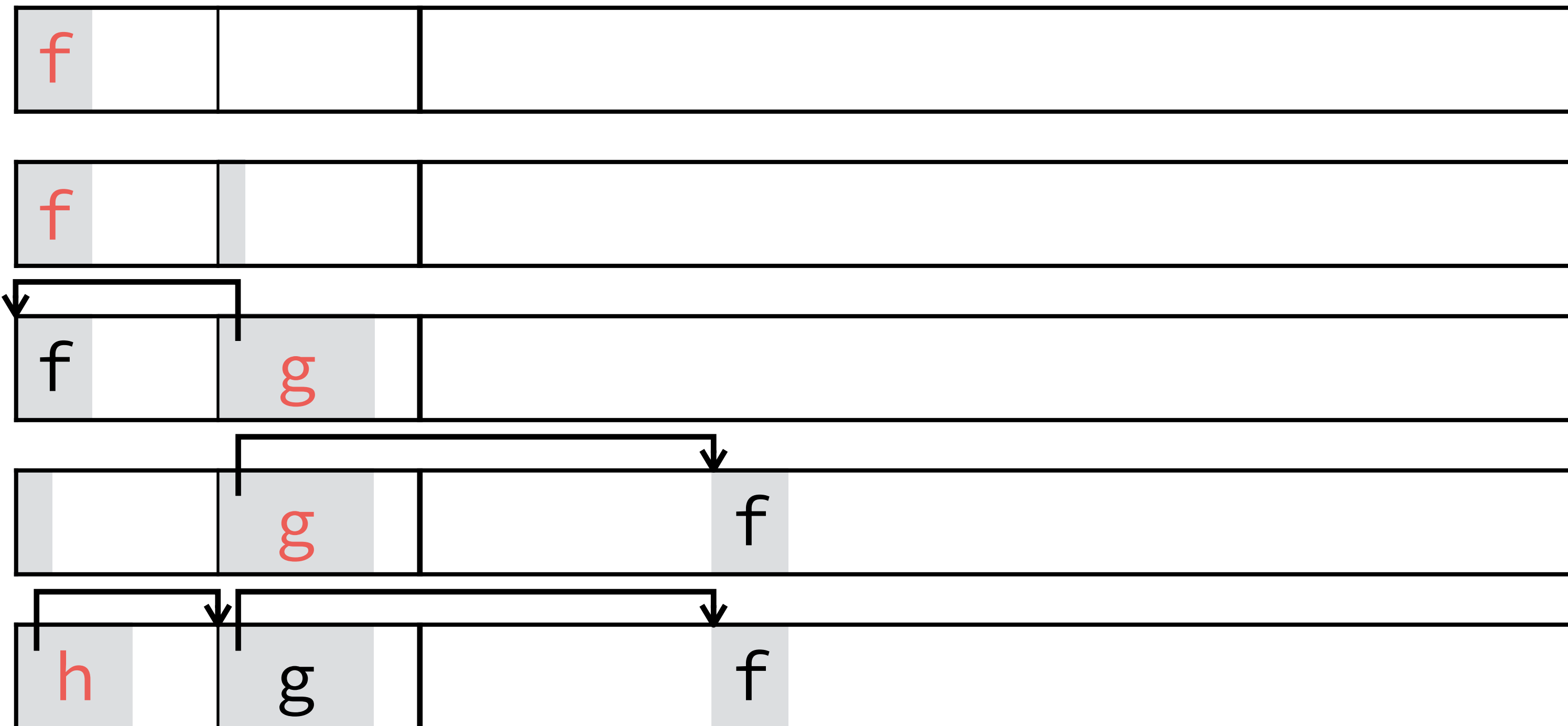


Non-tail calls and returns

top frames

heap

call stack

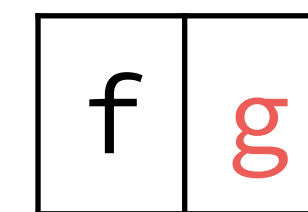
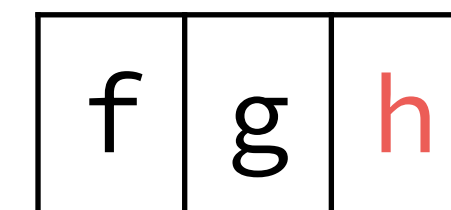
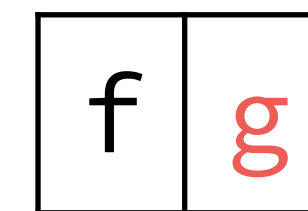
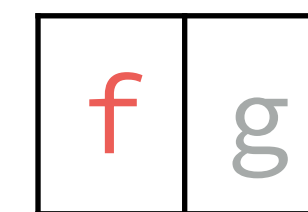
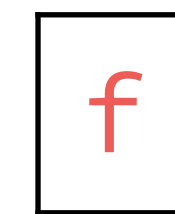
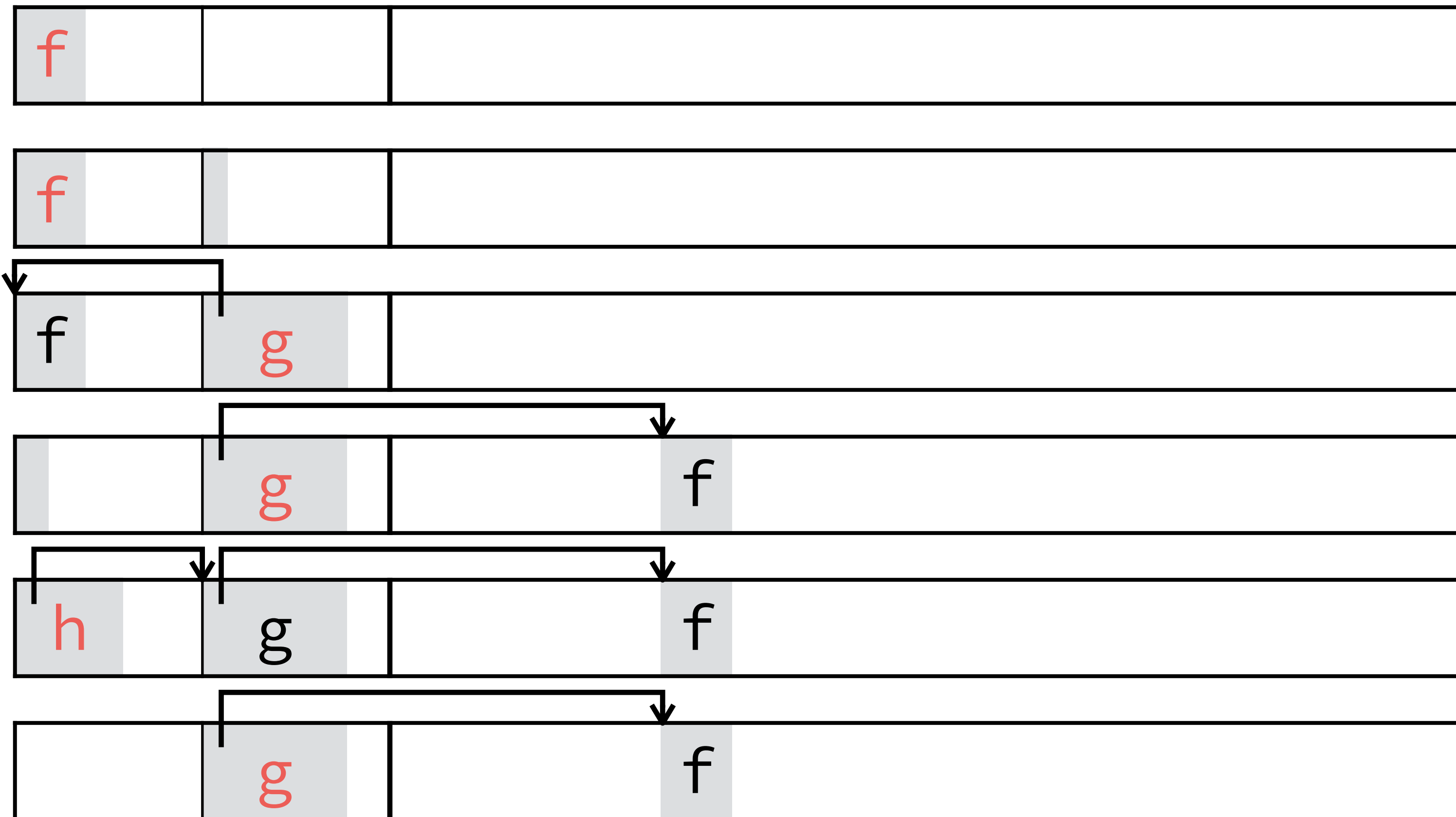


Non-tail calls and returns

top frames

heap

call stack

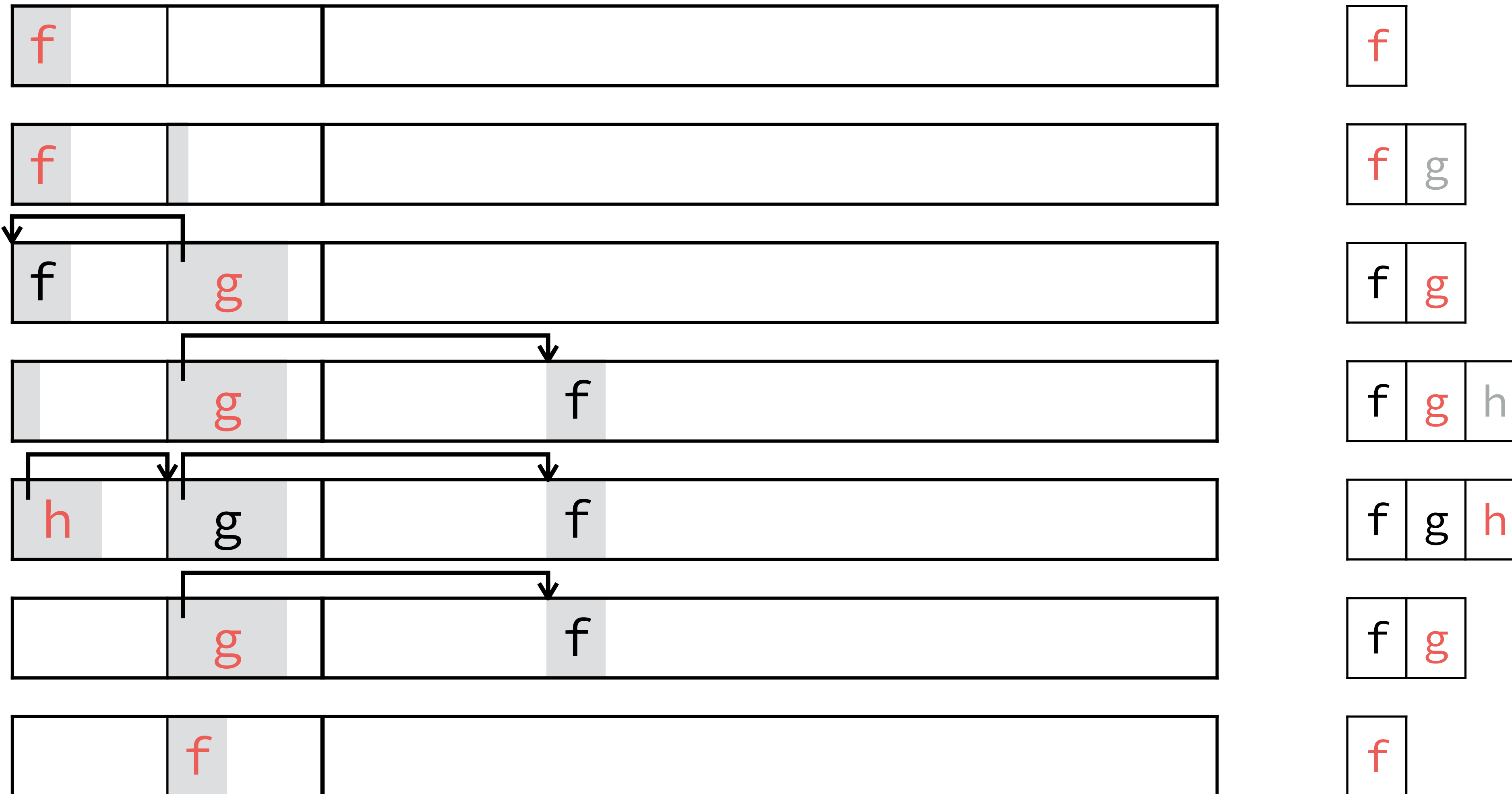


Non-tail calls and returns

top frames

heap

call stack



Tail call

When a function (the caller) wants to tail-call another function (the callee), it:

- stores the callee's arguments in the first register slots of its *own* frame,
- jumps to the callee's first instruction.

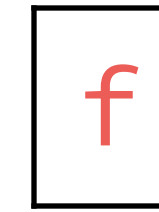
(As an optimization, if the callee directly follows the caller, the jump can be omitted.)

Tail calls

top frames

heap

call stack

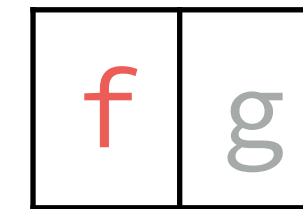
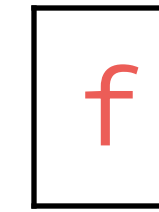


Tail calls

top frames

heap

call stack

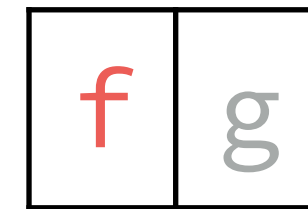
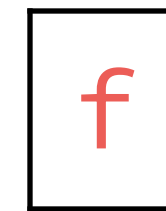
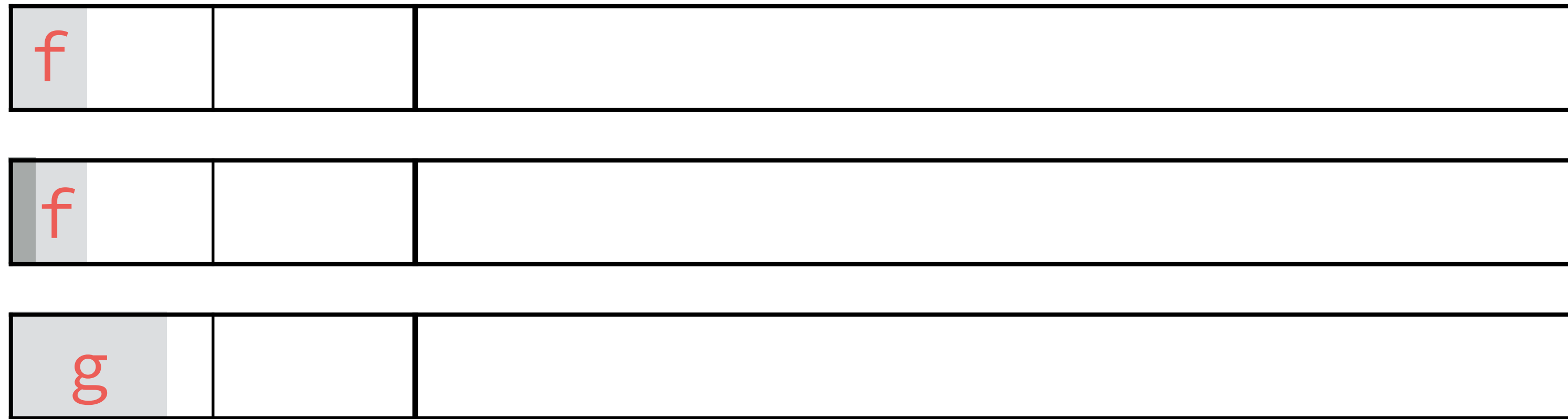


Tail calls

top frames

heap

call stack

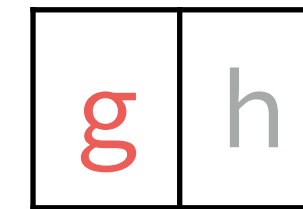
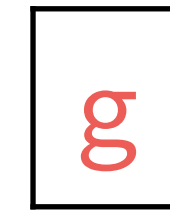
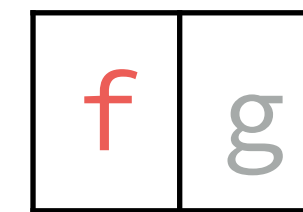
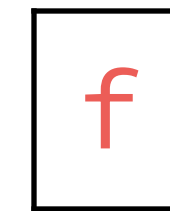
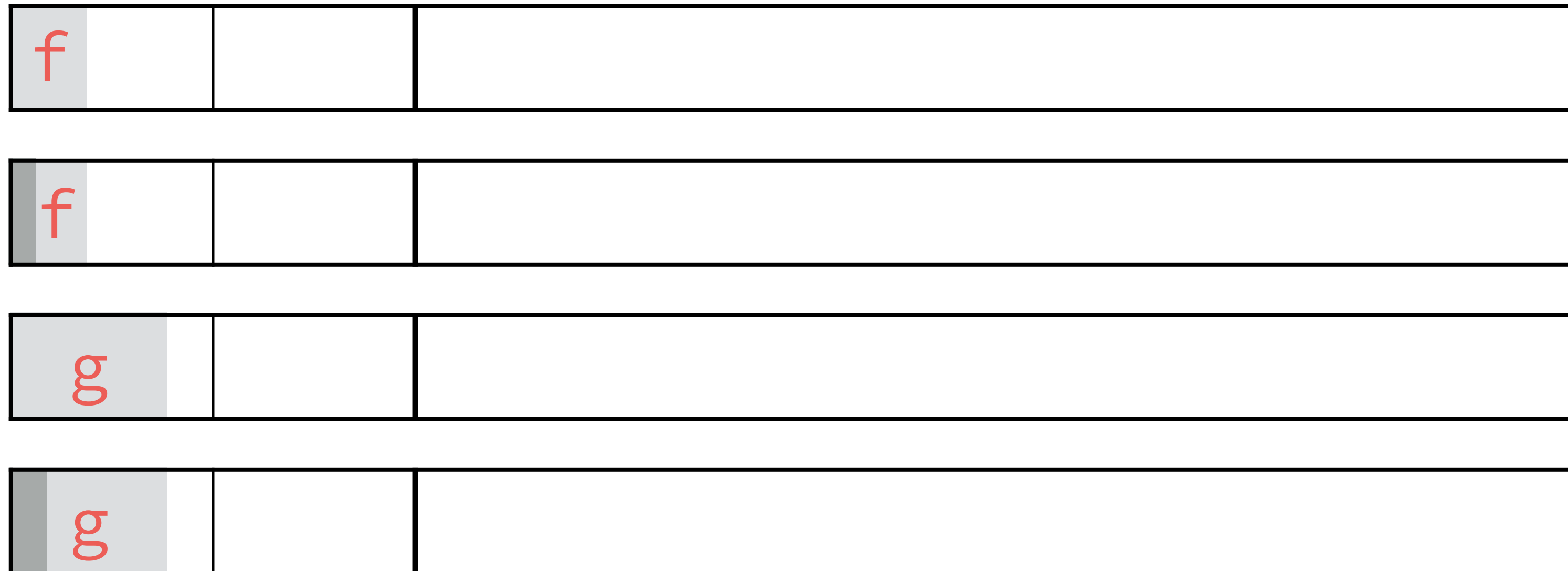


Tail calls

top frames

heap

call stack

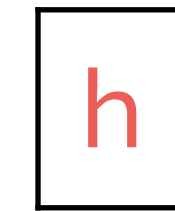
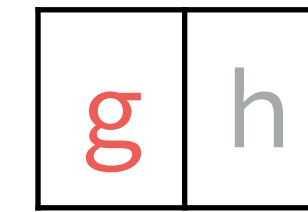
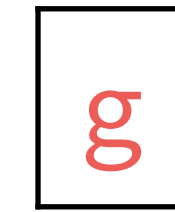
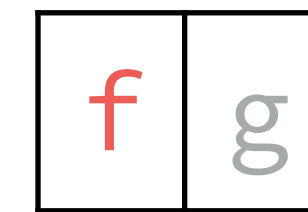
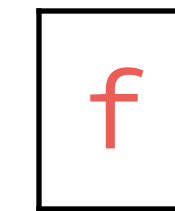
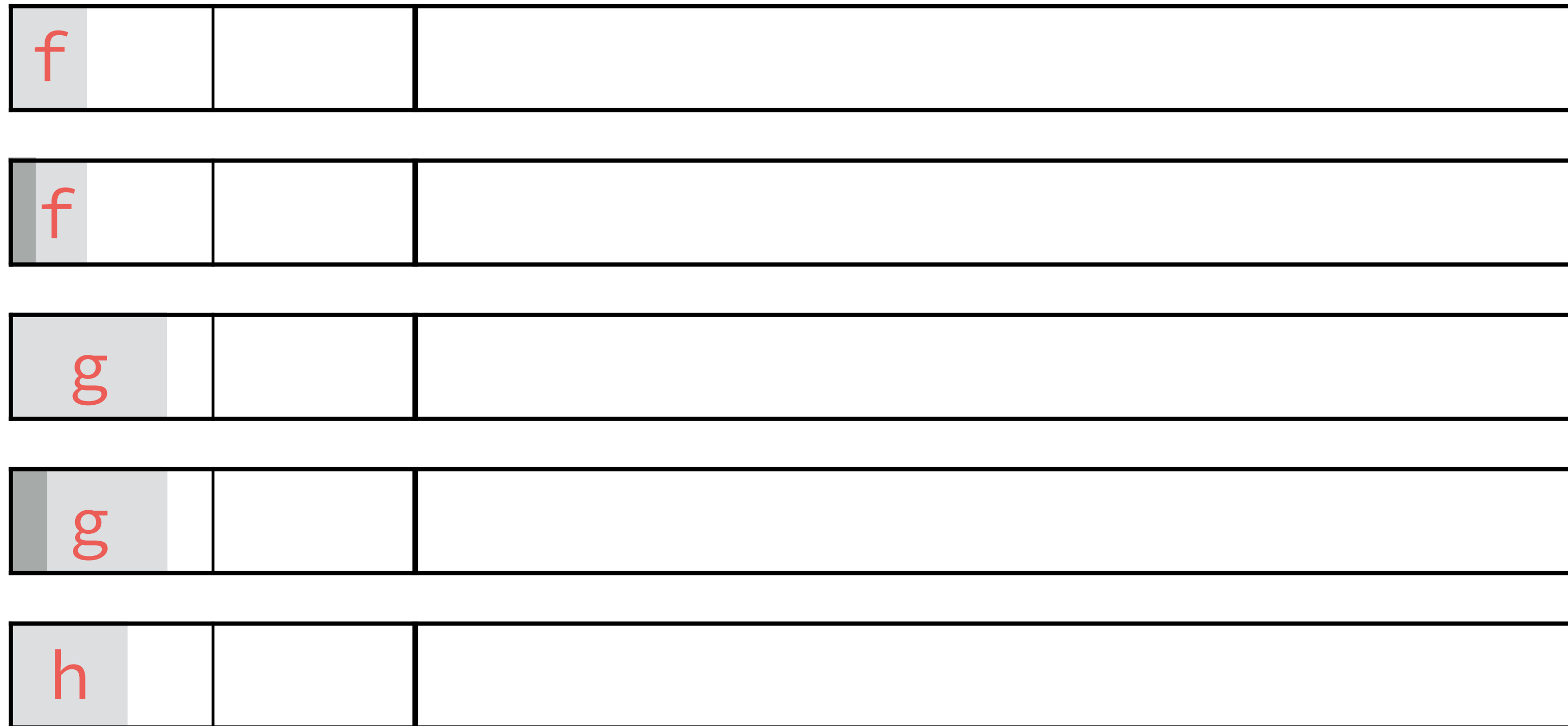


Tail calls

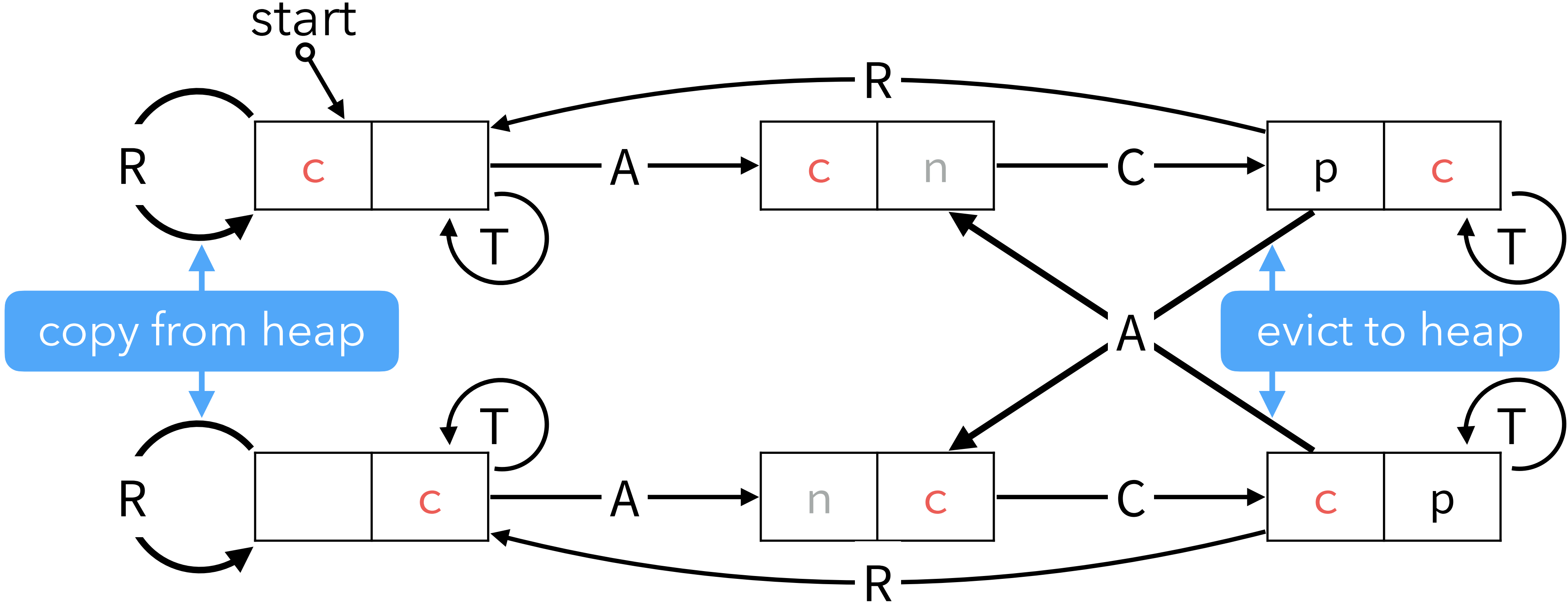
top frames

heap

call stack



Transitions



A - argument push
R - return

C - non-tail call
T - tail call

p, c, n - previous,
current, next frame

Arithmetic instructions (1)

ADD $R_a R_b R_c$ $R_a \leftarrow R_b + R_c$

SUB $R_a R_b R_c$ $R_a \leftarrow R_b - R_c$

MUL $R_a R_b R_c$ $R_a \leftarrow R_b \times R_c$

DIV $R_a R_b R_c$ $R_a \leftarrow R_b / R_c$

MOD $R_a R_b R_c$ $R_a \leftarrow R_b \% R_c$

R_a, R_b, R_c : registers

PC implicitly augmented by 4 by each instruction

Arithmetic instructions (2)

LSL $Ra\ Rb\ Rc$ $Ra \leftarrow Rb \ll Rc$

LSR $Ra\ Rb\ Rc$ $Ra \leftarrow Rb \gg Rc$

AND $Ra\ Rb\ Rc$ $Ra \leftarrow Rb \& Rc$

OR $Ra\ Rb\ Rc$ $Ra \leftarrow Rb | Rc$

XOR $Ra\ Rb\ Rc$ $Ra \leftarrow Rb \wedge Rc$

Ra, Rb, Rc : registers

PC implicitly augmented by 4 by each instruction

Control instructions

JLT $Ra\ Rb\ D^{11}$ if $Ra < Rb$ then $PC \leftarrow PC + 4 \cdot D^{11}$

JLE $Ra\ Rb\ D^{11}$ if $Ra \leq Rb$ then $PC \leftarrow PC + 4 \cdot D^{11}$

JEQ $Ra\ Rb\ D^{11}$ if $Ra = Rb$ then $PC \leftarrow PC + 4 \cdot D^{11}$

JNE $Ra\ Rb\ D^{11}$ if $Ra \neq Rb$ then $PC \leftarrow PC + 4 \cdot D^{11}$

JUMP_I Ra $PC \leftarrow Ra$

JUMP_D D^{27} $PC \leftarrow PC + 4 \cdot D^{27}$

Ra, Rb, Rc : registers,
 D^k : k -bit signed displacement

Call/return instructions

return register

CALL_I $R_a R_b$ $FP'[0] \leftarrow PC + 4, FP'[1] \leftarrow FP, PC \leftarrow R_b, FP \leftarrow FP'$

CALL_D $R_a D^{27}$ *like CALL_NI, except that* $PC \leftarrow PC + 4 \cdot D^{27}$

RET R_a $r \leftarrow R_a, PC \leftarrow FP[0], FP \leftarrow FP[1], R_{RET} \leftarrow r$

+ copy frame
from heap if
necessary

HALT R_a halt execution with the value of R_a

R_a : register,

R_{RET} : return register of matching CALL instruction

D^k : k -bit signed displacement,

r : temporary value

Frame and IO instructions

ARGS $R_a R_b R_c$	append R_a, R_b and R_c to the other top-frame block
FRAME U^8	resize current top-frame block to $2+U^8$
IO 0 R_a	$R_a \leftarrow$ byte read from console, zero-extended to 32 bits
IO 1 R_a	write least-significant byte of R_a to console

+ evict frame to heap if necessary

R_a, R_b, R_c : registers,
 U^k : k -bit unsigned constants
PC implicitly augmented by 4 by each instruction

Register instructions

LDLO $Ra\ S^{19}$ $Ra \leftarrow S^{19}$

LDHI $Ra\ U^{16}$ $Ra \leftarrow (U^{16} \ll 16) \mid (Ra \ \& \ FFFF_{16})$

MOVE $Ra\ Rb$ $Ra \leftarrow Rb$

Ra, Rb : registers,

S^k : k -bit signed constant,

U^k : k -bit unsigned constants

PC implicitly augmented by 4 by each instruction

Block instructions

BALO $R_a R_b T^8$ $R_a \leftarrow$ new block of size R_b and tag T^8

BSIZ $R_a R_b$ $R_a \leftarrow$ size of block R_b

BTAG $R_a R_b$ $R_a \leftarrow$ tag of block R_b

BGET $R_a R_b R_c$ $R_a \leftarrow$ element at index R_c of block R_b

BSET $R_a R_b R_c$ element at index R_c of block $R_b \leftarrow R_a$

R_a, R_b, R_c : registers,

T^8 : 8-bit block tag

PC implicitly augmented by 4 by each instruction

Example

The factorial in (hand-coded) L₃VM assembly:

```
;; R0 contains argument
fact:  FRAME 2
      JNE R0 C0 else
      RET C1
else:  SUB R1 R0 C1
      ARGS R1 C0 C0
      CALL_D R1 fact
      MUL R0 R0 R1
      RET R0
```