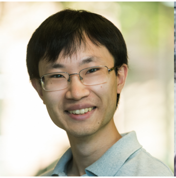
Speculative Optimizations without Fear

Olivier Flückiger, Gabriel Scherer, Ming-Ho Yee, Aviral Goel, Amal Ahmed, Jan Vitek

Northeastern University, Boston, USA

November 3, 2017



1 Context

2 Sourir

3 Formalization

4 Optimizations

Section 1

Context

Our work

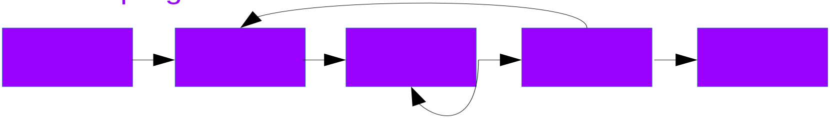
Just-in-time (JIT) compilation is essential to efficient dynamic language implementations.
(Javascript, Lua, R... Java)

There is a blind spot in our formal understanding of JITs: speculation.

We present a language design to study speculative optimizations and prove them correct.

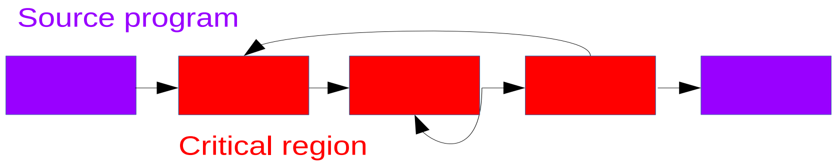
Just-in-time compilation

Source program



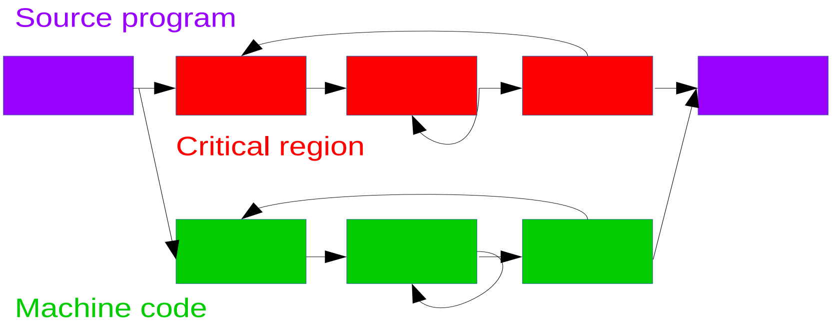
JITs:

Just-in-time compilation



JITs: Profiling

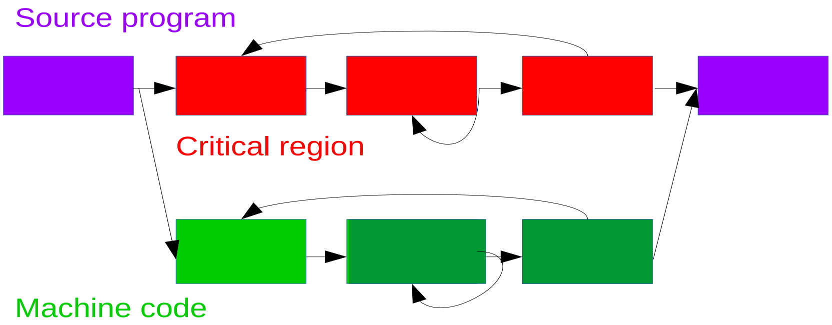
Just-in-time compilation



JITs: Profiling

+ High/Low languages
+ Dynamic code generation/mutation

Just-in-time compilation

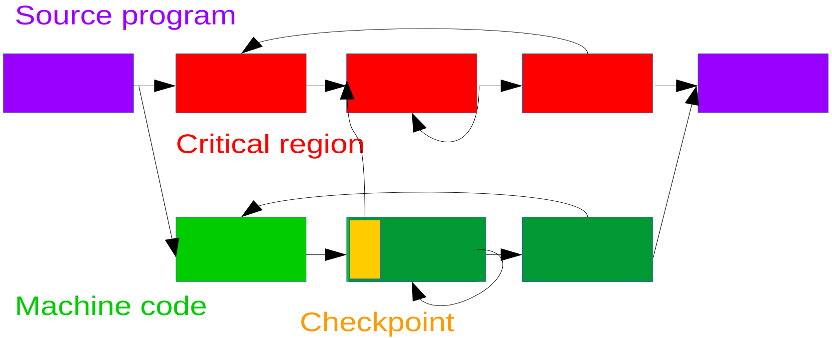


JITs: Profiling

+ High/Low languages
+ Dynamic code generation/mutation

+ Speculation

Just-in-time compilation



JITs: Profiling

+ High/Low languages
+ Dynamic code generation/mutation

+ Speculation and bailout

JITs:

- profiling
- high- and low-level languages (or multi-tiers, etc.)
- dynamic code generation + mutation

- speculation and bailout

JITs:

- profiling
- high- and low-level languages (or multi-tiers, etc.)
- dynamic code generation + mutation
  eliminates interpretation overhead (constant factor)
- speculation and bailout

JITs:

- profiling
- high- and low-level languages (or multi-tiers, etc.)
- dynamic code generation + mutation
  eliminates interpretation overhead (constant factor)
- speculation and bailout
  eliminates **dynamic features** overhead:
  dispatch (OO languages), type checks (Java),
  code loading (Java), redefinable primitives (R...)

JITs:

- profiling
- high- and low-level languages (or multi-tiers, etc.)
- dynamic code generation + mutation
  eliminates interpretation overhead (constant factor)
- speculation and bailout
  eliminates **dynamic features** overhead:
  dispatch (OO languages), type checks (Java),
  code loading (Java), redefinable primitives (R...)

JIT formalization: Myreen [2010]

- Stack language and x86 assembly
- dynamic code generation
- code mutation

mechanized in HOL!

JITs:

- profiling
- high- and low-level languages (or multi-tiers, etc.)
- dynamic code generation + mutation
  eliminates interpretation overhead (constant factor)
- speculation and bailout
  eliminates **dynamic features** overhead:
  dispatch (OO languages), type checks (Java),
  code loading (Java), redefinable primitives (R...)

JIT formalization: Myreen [2010]

- Stack language and x86 assembly
- dynamic code generation
- code mutation

mechanized in HOL!

JITs:

- profiling
- high- and low-level languages (or multi-tiers, etc.)
- dynamic code generation + mutation
  eliminates interpretation overhead (constant factor)
- speculation and bailout
  eliminates **dynamic features** overhead:
  dispatch (OO languages), type checks (Java),
  code loading (Java), redefinable primitives (R...)

JIT formalization: Myreen [2010]

- Stack language and x86 assembly
- dynamic code generation
- code mutation

mechanized in HOL!

What about **speculation**?

JITs:

- profiling
- high- and low-level languages (or multi-tiers, etc.)
- dynamic code generation + mutation
  eliminates interpretation overhead (constant factor)
- speculation and bailout
  eliminates **dynamic features** overhead:
  dispatch (OO languages), type checks (Java),
  code loading (Java), redefinable primitives (R...)

JIT formalization: Myreen [2010]

- Stack language and x86 assembly
- dynamic code generation
- code mutation

mechanized in HOL!

What about **speculation**? This work.

What do we want to know?

Speculation requires keeping **bailout data**.

How should optimizations maintain/transform bailout data?
(inlining is tricky)

Does the presence of checkpoint restrict optimizations?
(hoisting writes or IO is tricky)

When an assumption fails, how much of the other optimizations can keep?
(non-stack-order is tricky)

How should practitioners reason about correctness?

DEEP SPECULATION.

THE OIL-SPECULATOR'S DREAM.

Harper's Weekly cartoon of February 11, 1865. [8]

Sourir

- high- and low-level languages
- dynamic code generation
- speculative optimization and bailout

Sourir

- ~~high- and low-level languages~~ a single bytecode language
- dynamic code generation
- speculative optimization and bailout

Sourir

- ~~high- and low-level languages~~ a single bytecode language
- ~~dynamic code generation~~ one unrolled multi-version program
- speculative optimization and bailout

Sourir

- ~~high- and low-level languages~~ a single bytecode language
- ~~dynamic code generation~~ one unrolled multi-version program
- speculative optimization and bailout a **checkpoint** instruction

Sourir

- ~~high- and low-level languages~~ a single bytecode language
- ~~dynamic code generation~~ one unrolled multi-version program
- speculative optimization and bailout a **checkpoint** instruction

$F_{fun}(c) \rightarrow$
$\quad V_{tough} \rightarrow$
$\quad\quad L_0 : \quad$ **var** $o = 1$
$\quad\quad L_1 : \quad$ **print** $c + o$
$\quad V_{luck} \rightarrow$
$\quad\quad L_0 : \quad$ **assume** $c = 41$ **else** $F_{fun}.V_{tough}.L_1\ [c = c, o = 1]$
$\quad\quad L_1 : \quad$ **print** 42

Contribution

A language design to model speculative optimization: `Sourir`

A kit of correct program transformations and optimizations

A methodology to reason about correct speculative optimizations

Section 2

Sourir

A simple bytecode language

$$
\begin{aligned}
i \;::=\; & \\
\mid\; & \textbf{var}\ x = e \\
\mid\; & \textbf{drop}\, x \\
\mid\; & x \leftarrow e \\
\mid\; & \textbf{array}\, x[e] \\
\mid\; & \textbf{array}\, x = [e^*] \\
\mid\; & x[e_1] \leftarrow e_2 \\
\mid\; & \textbf{branch}\, e\ L_1\ L_2 \\
\mid\; & \textbf{goto}\, L \\
\mid\; & \textbf{print}\, e \\
\mid\; & \textbf{read}\, x \\
\mid\; & \textbf{call}\ x = e(e^*) \\
\mid\; & \textbf{return}\, e \\
\mid\; & \textbf{assume}\ e^*\ \textbf{else}\ \xi\ \tilde{\xi}^* \\
\mid\; & \textbf{stop}
\end{aligned}
$$

$$
\begin{aligned}
e \;::=\; & \\
\mid\; & se \\
\mid\; & x[se] \\
\mid\; & \textbf{length}(se) \\
\mid\; & primop\,(se^*)
\end{aligned}
$$

$$
\begin{aligned}
se \;::=\; & \\
\mid\; & lit \\
\mid\; & F \\
\mid\; & x
\end{aligned}
$$

$$
\begin{aligned}
lit \;::=\; & \\
\mid\; & \ldots, -1, 0, 1, \ldots \\
\mid\; & \textbf{nil} \mid \textbf{true} \mid \textbf{false}
\end{aligned}
$$

Versions

$P ::= (F(x^*) \rightarrow D_F)^*$ **program**: a list of named functions
$D_F ::= (V \rightarrow I)^*$ **function definition**: list of versioned instruction streams
$I ::= (L : i)^*$ **instruction stream** with labeled instructions

$F_{fun}(c) \rightarrow$
  $V_{tough} \rightarrow$
    L_0 : **var** $o = 1$
    L_1 : **print** $c + o$
  $V_{luck} \rightarrow$
    L_0 : **assume** $c = 41$ **else** $F_{fun}.V_{tough}.L_1\ [c = c, o = 1]$
    L_1 : **print** 42

Checkpoints

Checkpoint: **guards** + **bailout data**.

$$\textbf{assume } c = 41 \textbf{ else } F_{fun}.V_{tough}.L_1 \; [c = c, o = 1]$$

Guards: just a list of expressions returning booleans.

Bailout data:

- where to go: $F_f.V_w.L_l$
- in what state: $[x_1 = e_1, .., x_n = e_n]$
- (plus more: see inlining)

Checkpoints

Checkpoint: **guards** + **bailout data**.

$$\textbf{assume } c = 41 \textbf{ else } F_{fun}.V_{tough}.L_1 \; [c = c, o = 1]$$

Guards: just a list of expressions returning booleans.

Bailout data:

- where to go: $F_f.V_w.L_l$
- in what state: $[x_1 = e_1, .., x_n = e_n]$
- (plus more: see inlining)

Not a branch. (inlining)
Checkpoints simplify optimizations...

Checkpoints

Checkpoint: **guards** + **bailout data**.

$$\textbf{assume } c = 41 \textbf{ else } F_{fun}.V_{tough}.L_1 \; [c = c, o = 1]$$

Guards: just a list of expressions returning booleans.

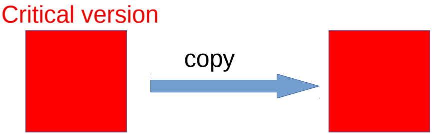
Bailout data:

- where to go: $F_f.V_w.L_l$
- in what state: $[x_1 = e_1, .., x_n = e_n]$
- (plus more: see inlining)

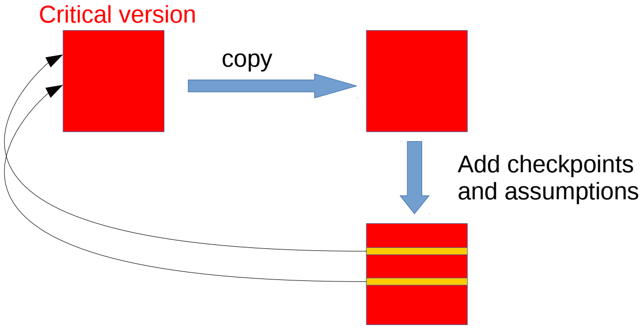
Not a branch. (inlining)
Checkpoints simplify optimizations...and correctness proofs!

Speculative optimization pipeline

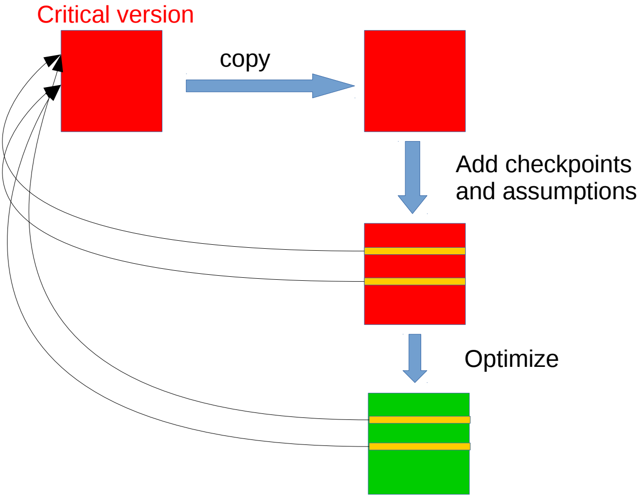
Speculative optimization pipeline



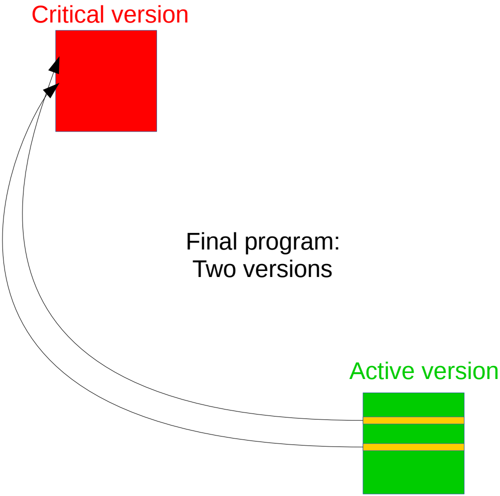
Speculative optimization pipeline



Speculative optimization pipeline



Speculative optimization pipeline



Final program:
Two versions

Speculative optimization pipeline

Section 3

Formalization

Execution: Operational semantics

Configurations:

$$C ::= \langle P\ I\ L\ K^*\ M\ E \rangle$$

Actions:

$$A ::= \ \textbf{read}\ \mathit{lit} \mid \textbf{print}\ \mathit{lit} \qquad A_\tau := \ A \mid \tau \qquad T ::= \ A^*.$$

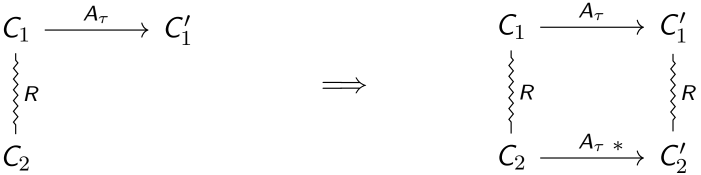
Reduction:

$$C_1 \xrightarrow{A_\tau}{}^* \ C_2 \qquad C_1 \xrightarrow{T}{}^* \ C_2$$

Equivalence: (weak) bisimulation

Relation R between the configurations over P_1 and P_2.

R is a weak **simulation** if:



R is a weak **bisimulation** if R and R^{-1} are simulations.

Bailout invariants

Version invariant: All versions of a function are equivalent.
(Necessary to replace the active version)

Bailout invariant: Bailing out **more** than necessary is correct.
(Necessary to add new assumptions)

Section 4

Optimizations

Branch pruning – from the kit

$$V_{base} \rightarrow$$

$$\begin{array}{|ll} L_1 : & \textbf{branch}\ tag = INT\ \ L_{int}\ \ L_{nonint} \\ L_{int} : & \ldots \\ L_{nonint} : & \ldots \end{array}$$

Branch pruning – from the kit

$$V_{base} \rightarrow \left| \begin{array}{ll} L_1 : & \textbf{branch}\ tag = INT\ \ L_{int}\ \ L_{nonint} \\ L_{int} : & \dots \\ L_{nonint} : & \dots \end{array} \right.$$

$$V_{opt} \rightarrow \left| \begin{array}{ll} L_0 : & \textbf{assume}\ tag = INT\ \textbf{else}\ F.V_{base}.L_1\ \delta \\ L_1 : & \textbf{branch}\ tag = INT\ \ L_{int}\ \ L_{nonint} \\ L_{int} : & \dots \\ L_{nonint} : & \dots \end{array} \right.$$

Checkpoint + guard inserted

Bailout invariant!

Branch pruning – from the kit

$$V_{base} \rightarrow$$

$$\left|\begin{array}{ll} L_1 : & \textbf{branch}\ tag = INT\ \ L_{int}\ \ L_{nonint} \\ L_{int} : & \ldots \\ L_{nonint} : & \ldots \end{array}\right.$$

$$V_{opt} \rightarrow \left|\begin{array}{ll} L_0 : & \textbf{assume}\ tag = INT\ \textbf{else}\ F.V_{base}.L_1\ \delta \\ L_1 : & \textbf{branch}\ \textbf{true}\ \ L_{int}\ \ L_{nonint} \\ L_{int} : & \ldots \\ L_{nonint} : & \ldots \end{array}\right.$$

constant folding

Branch pruning – from the kit

$$V_{base} \rightarrow$$

$$\left|\begin{array}{ll} L_1 : & \textbf{branch}\ tag = INT\ \ L_{int}\ \ L_{nonint} \\ L_{int} : & \ldots \\ L_{nonint} : & \ldots \end{array}\right.$$

$$V_{opt} \rightarrow \quad \left|\begin{array}{ll} L_0 : & \textbf{assume}\ tag = INT\ \textbf{else}\ F.V_{base}.L_1\ \delta \\ L_{int} : & \ldots \end{array}\right.$$

unreachable code elimination

Inlining

$F_{main}() \rightarrow$
$\quad V_{inlined} \rightarrow$

L_0 :	**array** $vec = [1, 2, 3, 4]$
L_2 :	**var** $size =$ **nil**
L_3 :	**var** $obj = vec$
L_{cp_1} :	**assume** $obj \neq$ **nil else** $\ldots$
L_5 :	**var** $len =$ **length**(obj)
L_6 :	$size \leftarrow len * 4$
L_7 :	**drop** len
L_8 :	**drop** obj
L_9 :	**goto** L_{ret}
L_{ret} :	**print** $size$

$\quad V_{base} \rightarrow \quad \ldots$

$F_{main}() \rightarrow$
$\quad V_{base} \rightarrow$

L_0 :	**array** $vec = [1, 2, 3, 4]$
L_2 :	**call** $size = F_{size}(vec)$
L_{ret} :	**print** $size$

$F_{size}(obj) \rightarrow$
$\quad V_{opt} \rightarrow$

L_{cp_1} :	**assume** $obj \neq$ **nil else** ξ
L_{vec} :	**var** $len =$ **length**(obj)
L_3 :	**return** $len * 4$

$\quad V_{base} \rightarrow \quad \ldots$

Inlining

$F_{main}(\,) \rightarrow$

$\quad V_{inlined} \rightarrow$

L_0 :	**array** $vec = [1, 2, 3, 4]$
L_2 :	**var** $size =$ **nil**
L_3 :	**var** $obj = vec$
L_{cp_1} :	**assume** $obj \neq$ **nil else** $\ldots$
L_5 :	**var** $len =$ **length**(obj)
L_6 :	$size \leftarrow len * 4$
L_7 :	**drop** len
L_8 :	**drop** obj
L_9 :	**goto** L_{ret}
L_{ret} :	**print** $size$

$\quad V_{base} \rightarrow \ \ldots$

$F_{main}(\,) \rightarrow$

$\quad V_{base} \rightarrow$

L_0 :	**array** $vec = [1, 2, 3, 4]$
L_2 :	**call** $size = F_{size}(vec)$
L_{ret} :	**print** $size$

$F_{size}(obj) \rightarrow$

$\quad V_{opt} \rightarrow$

L_{cp_1} :	**assume** $obj \neq$ **nil else** ξ
L_{vec} :	**var** $len =$ **length**(obj)
L_3 :	**return** $len * 4$

$\quad V_{base} \rightarrow \ \ldots$

assume $obj \neq$ **nil else** $\xi \ \langle F_{main}.V_{base}.L_{ret}\ size\ [vec = vec]\rangle$

Conclusion

All you need for speculation: versions + checkpoints.

Future work: bidirectional transformations.

Thanks!
Questions?

Magnus O. Myreen. Verified just-in-time compiler on x86. In **Principles of Programming Languages (POPL)**, 2010.