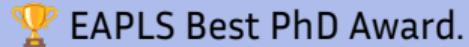


Aurèle Barrière

Doctorat, IRISA 
avec Sandrine Blazy et David Pichardie

*Vérification formelle de
compilation à la volée (JIT)*
2019-2022



PostDoc, EPFL 
avec Clément Pit-Claudel
*Vers des moteurs de regex modernes
linéaires et vérifiés*
2023-2025

Stages : SNU  (2017), Federico II  (2017), Princeton  (2018).

Intégration : Équipe CAMUS

Comment faire confiance à l'exécution d'un programme sur le web?

Comment faire confiance à l'exécution d'un programme sur le web?

Un besoin de garanties

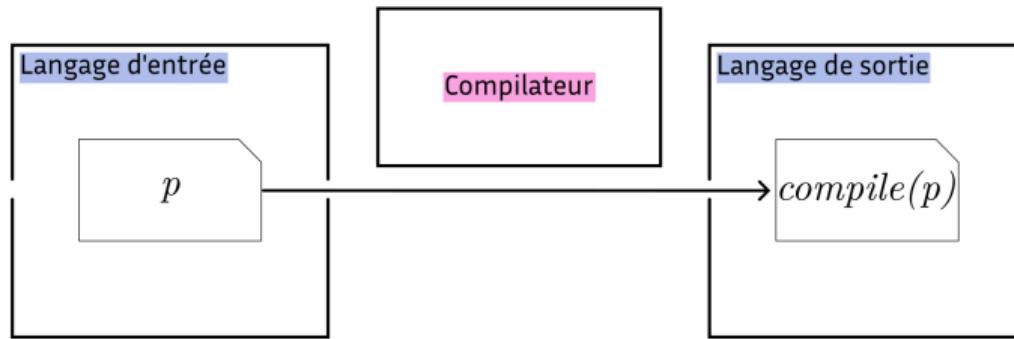
- Les navigateurs sont des environnements d'exécution, pour JavaScript et WebAssembly.
- Nouvelles techniques de compilation (par exemple, compilation à la volée/JIT).
- Leurs bugs de compilation sont dangereux! Google Chrome et Firefox en 2025 : [\[CVE-2025-0291\]](#),
[\[CVE-2025-0434\]](#), [\[CVE-2025-0445\]](#), [\[CVE-2025-0611\]](#), [\[CVE-2025-0612\]](#), [\[CVE-2025-0995\]](#),
[\[CVE-2025-0998\]](#), [\[CVE-2025-0999\]](#), [\[CVE-2025-1011\]](#), [\[CVE-2025-1914\]](#), [\[CVE-2025-1933\]](#).

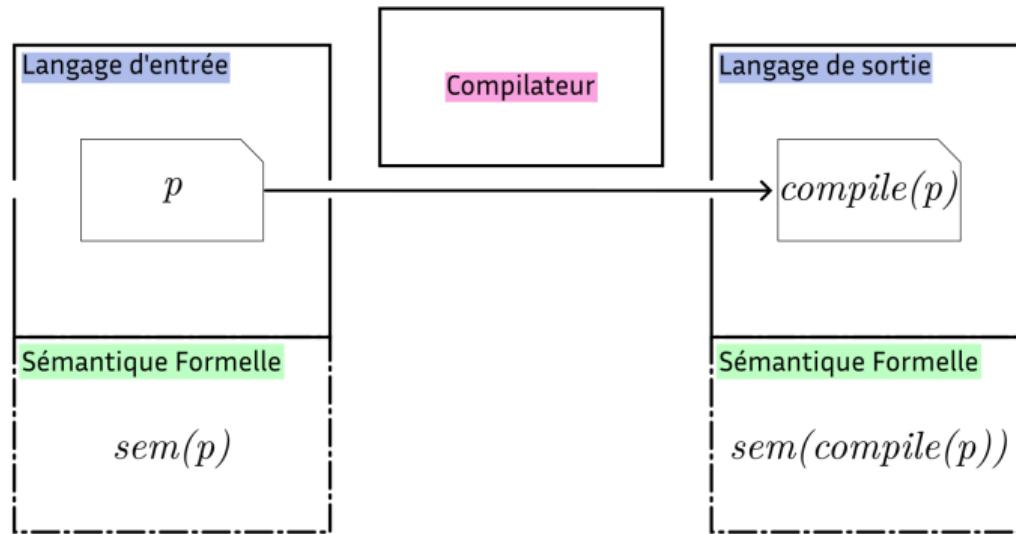
Comment faire confiance à l'exécution d'un programme sur le web?

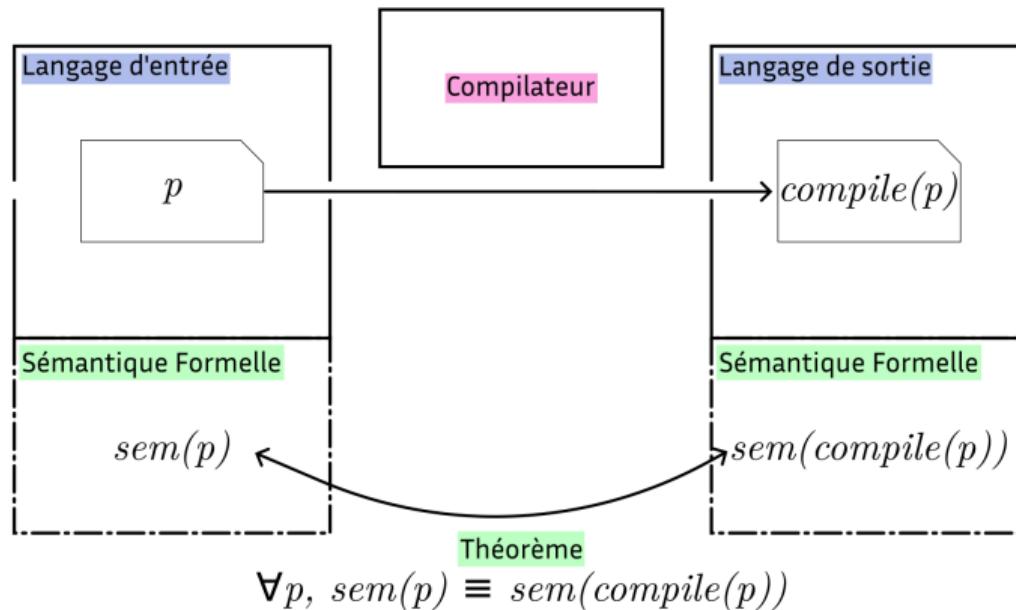
Un besoin de garanties

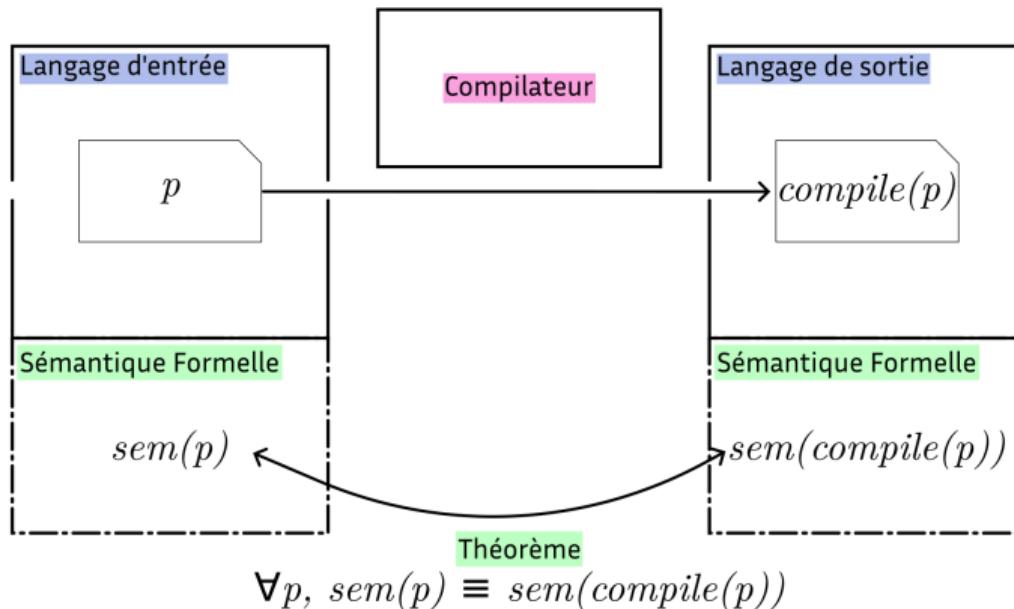
- Les navigateurs sont des environnements d'exécution, pour JavaScript et WebAssembly.
- Nouvelles techniques de compilation (par exemple, compilation à la volée/JIT).
- Leurs bugs de compilation sont dangereux! Google Chrome et Firefox en 2025 : [\[CVE-2025-0291\]](#),
[\[CVE-2025-0434\]](#), [\[CVE-2025-0445\]](#), [\[CVE-2025-0611\]](#), [\[CVE-2025-0612\]](#), [\[CVE-2025-0995\]](#),
[\[CVE-2025-0998\]](#), [\[CVE-2025-0999\]](#), [\[CVE-2025-1011\]](#), [\[CVE-2025-1914\]](#), [\[CVE-2025-1933\]](#).

Comment implémenter ces environnements d'exécution sans bugs?



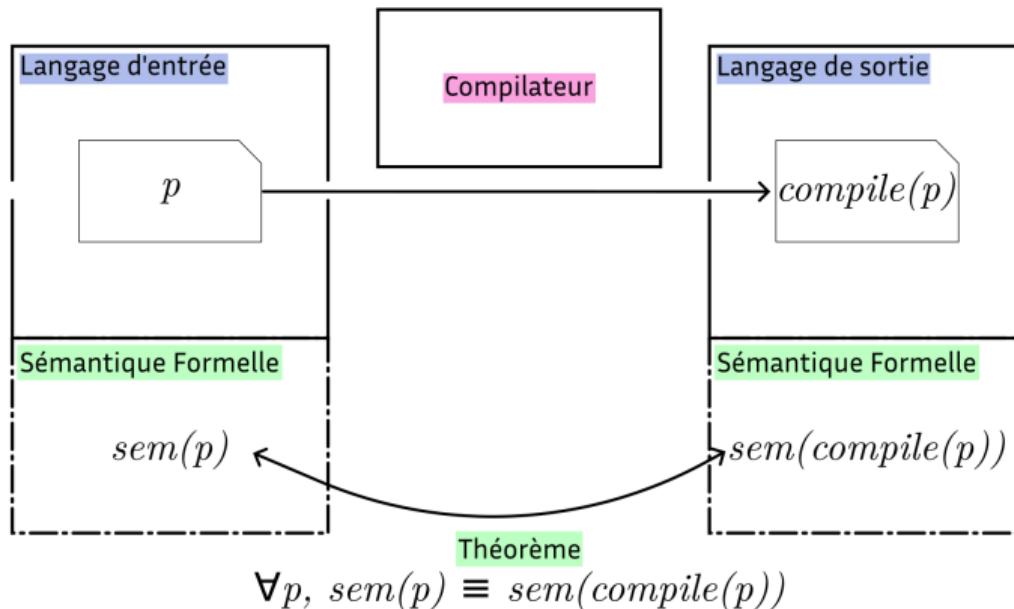






Compilateurs formellement vérifiés dans un assistant de preuve

CompCert (Coq/Rocq) [Leroy, POPL'2006], CakeML (HOL) [Kumar et al. POPL'2014]...



Compilateurs formellement vérifiés dans un assistant de preuve

CompCert (Coq/Rocq) [Leroy, POPL'2006], CakeML (HOL) [Kumar et al. POPL'2014]...

Un problème

Les techniques de compilation et d'exécution utilisées sur le web ont largement dévié de la théorie.

Concevoir à la fois la théorie et des implémentations vérifiées pour un web de confiance.

Concevoir à la fois la théorie et des implémentations vérifiées pour un web de confiance.

Bénéfices

- Obtenir des implémentations de confiance, vérifiées avec un assistant de preuve (Coq/Rocq).
- Comprendre les techniques modernes.
- Concevoir de nouvelles techniques.

Concevoir à la fois la théorie et des implémentations vérifiées pour un web de confiance.

Bénéfices

- Obtenir des implémentations de confiance, vérifiées avec un assistant de preuve (Coq/Rocq).
- Comprendre les techniques modernes.
- Concevoir de nouvelles techniques.

Deux cas de compilation non traditionnelle du web

Doctorat : Vérification formelle de compilation à la volée (JIT)

PostDoc : Étude formelle des regex JavaScript

JIT (Just-in-Time) = entremêler **exécution et compilation** du programme.

Publications : [ACMBooks'25], [POPL'23], [POPL'21], [CoqPL'20].

Prix de thèse : 🏆 EAPLS Best PhD Dissertation Award.

Collaboration : Olivier Flückiger & Jan Vitek (Northeastern 🇺🇸), créateurs du JIT Rir pour le langage R.

JIT (Just-in-Time) = entremêler **exécution et compilation** du programme.



Publications : [ACMBooks'25], [POPL'23], [POPL'21], [CoqPL'20].

Prix de thèse : 🏆 EAPLS Best PhD Dissertation Award.

Collaboration : Olivier Flückiger & Jan Vitek (Northeastern 🇺🇸), créateurs du JIT Rir pour le langage R.

JIT (Just-in-Time) = entremêler **exécution et compilation** du programme.

CVE-2019-11707, CVE-2019-11708: Multiple Zero-Day Vulnerabilities in Mozilla Firefox Exploited in the Wild



Satnam Narang

June 18, 2019 | 3 Min Read



Security researchers discover two zero-day vulnerabilities in Mozilla Firefox used in targeted attacks.



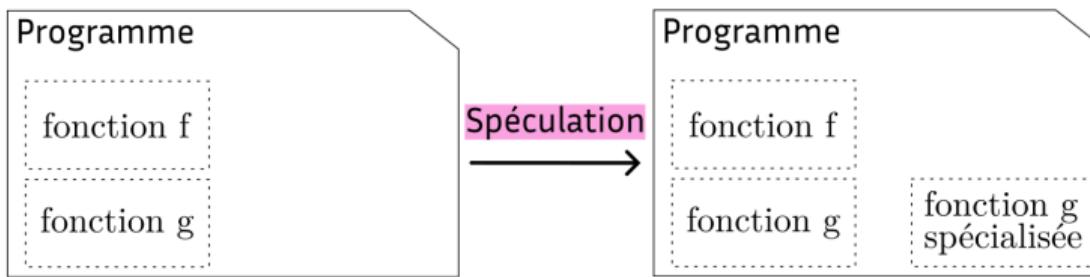
Les dangers du JIT

Comment écrire un compilateur JIT correct?

Publications : [ACMBooks'25], [POPL'23], [POPL'21], [CoqPL'20].

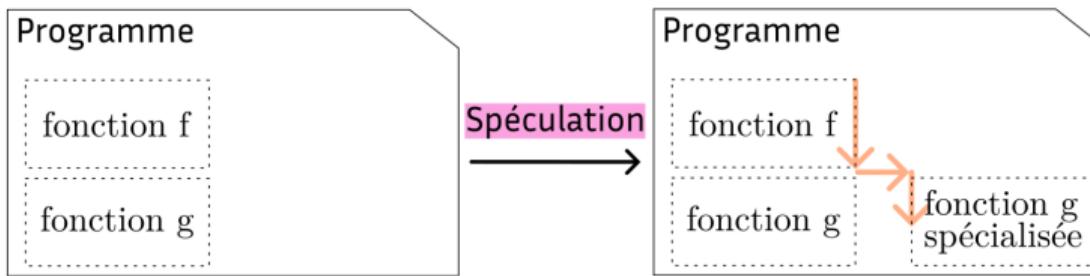
Prix de thèse : 🏆 EAPLS Best PhD Dissertation Award.

Collaboration : Olivier Flückiger & Jan Vitek (Northeastern 🇺🇸), créateurs du JIT Rir pour le langage R.



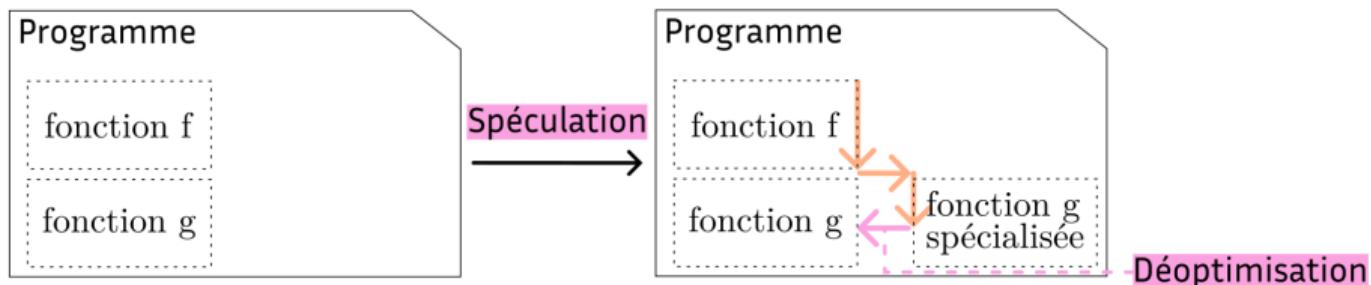
Spéculation

Compiler des versions spécialisées de fonctions.



Spéculation

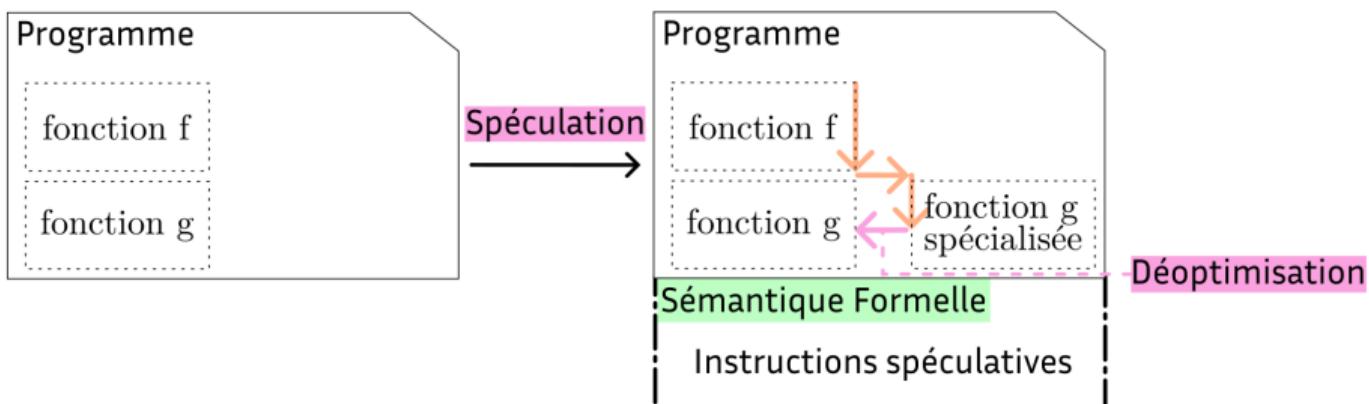
Compiler des versions spécialisées de fonctions.



Spéculation

Compiler des versions spécialisées de fonctions.

Déoptimisation : sauter dynamiquement de la fonction spécialisée et compilée vers la version originale.



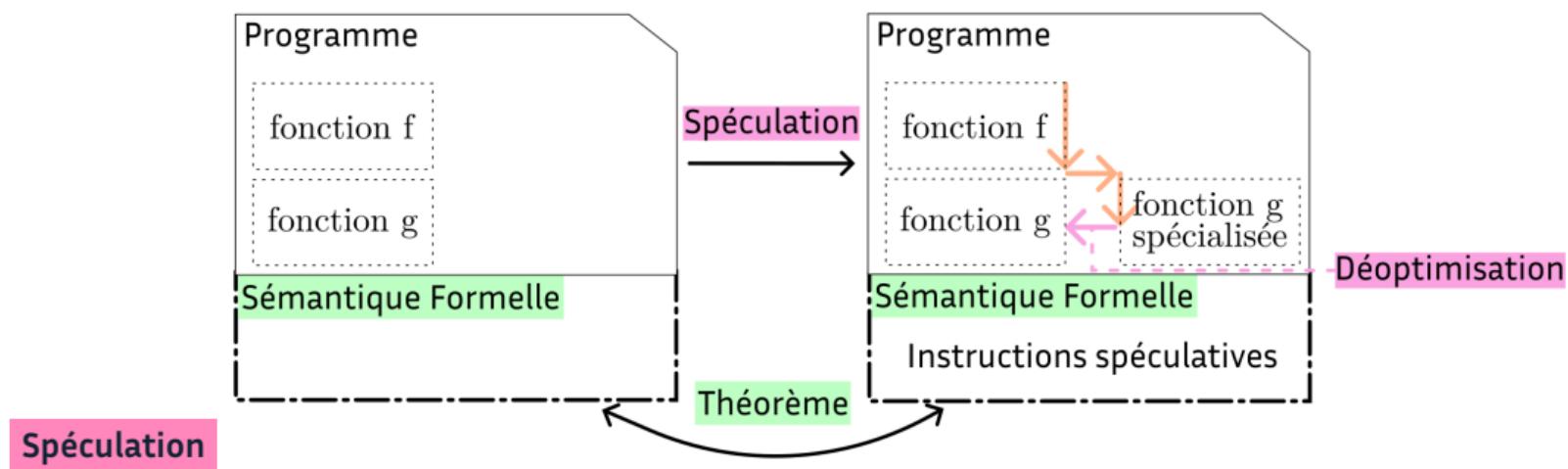
Spéculation

Compiler des versions spécialisées de fonctions.

Déoptimisation : sauter dynamiquement de la fonction spécialisée et compilée vers la version originale.

Contributions

Sémantique formelle pour des instructions spéculatives (difficulté : non déterminisme).



Compiler des versions spécialisées de fonctions.

Déoptimisation : sauter dynamiquement de la fonction spécialisée et compilée vers la version originale.

Contributions

Sémantique formelle pour des instructions spéculatives (difficulté : non déterminisme).

Vérification de leur insertion, manipulation et compilation.

Une méthode de référence pour spéculer dans un JIT.

Mon doctorat

Des prototypes de JITs vérifiés et exécutables.

Artéfacts :

+30K lignes de Coq/Rocq

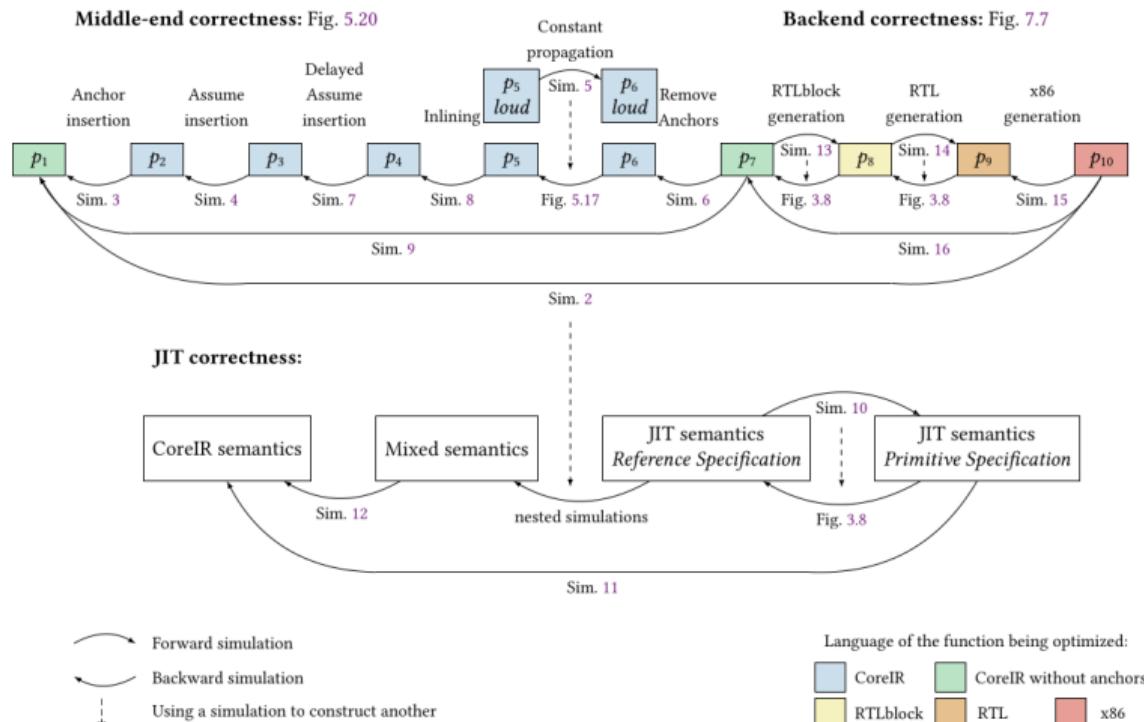


Figure 8.1 – Composing all our simulations for an effectful JIT with speculation and native code generation

$r ::=$	a	Caractère
	$r_1 r_2$	Séquence
	$r_1 r_2$	Disjonction
	r^*	Étoile

$r ::= a$	Caractère
$r_1 r_2$	Séquence
$r_1 r_2$	Disjonction
r^*	Étoile
<hr/>	
[a - z]	Classe de caractères
\$, ^	Ancre
(r)	Groupe de capture
(?= r)	Lookahead
(?<= r)	Lookbehind
\1	Backreference

Complexité : linéaire , inconnue , NP-dur .

$r ::=$	a	Caractère
	$r_1 r_2$	Séquence
	$r_1 r_2$	Disjonction
	r^*	Étoile
<hr/>		
	$[a - z]$	Classe de caractères
	$\$, ^$	Ancre
	(r)	Groupe de capture
	$(?= r)$	Lookahead
	$(?<= r)$	Lookbehind
	$\backslash 1$	Backreference

Problème : complexité exponentielle

Vulnérabilité ReDoS : 12% des serveurs JS vulnérables.
`"a".repeat(100).match(/(a*)*b/)`: 10^{14} ans.

Complexité : linéaire, inconnue, NP-dur.

$r ::=$	a	Caractère
	$r_1 r_2$	Séquence
	$r_1 r_2$	Disjonction
	r^*	Étoile
<hr/>		
	$[a - z]$	Classe de caractères
	$$, ^$	Ancre
	(r)	Groupe de capture
	$(?= r)$	Lookahead
	$(?<= r)$	Lookbehind
	$\backslash 1$	Backreference

Problème : complexité exponentielle

Vulnérabilité ReDoS : 12% des serveurs JS vulnérables.
`"a".repeat(100).match(/(a*)*b/)`: 10^{14} ans.

 Solution dans V8 (Google Chrome/Node.JS) :
un moteur linéaire pour les fonctionnalités linéaires.

Complexité : linéaire, inconnue, NP-dur.

$r ::= a$	Caractère
$r_1 r_2$	Séquence
$r_1 r_2$	Disjonction
r^*	Étoile
<hr/>	
$[a - z]$	Classe de caractères
$$, ^$	Ancre
(r)	Groupe de capture
$(?= r)$	Lookahead
$(?<= r)$	Lookbehind
$\backslash 1$	Backreference

Complexité : linéaire , inconnue , NP-dur .

Problème : complexité exponentielle

Vulnérabilité ReDoS : 12% des serveurs JS vulnérables.
`"a".repeat(100).match(/(a*)*b/)`: 10^{14} ans.

 Solution dans V8 (Google Chrome/Node.JS) :
un moteur linéaire pour les fonctionnalités linéaires.

Des problèmes algorithmiques et sémantiques

J'ai montré que :
Les algorithmes linéaires étaient faux ou non linéaires.
Les modèles sémantiques étaient incomplets ou faux.

Les spécificités sémantiques de JavaScript

- Les groupes de captures ont une sémantique unique (réinitialisation à chaque itération).
- **Nouveau :** L'étoile a une sémantique différente! “`ab`”.`match(/(a?b??)*/)`

Les spécificités sémantiques de JavaScript

- Les groupes de captures ont une sémantique unique (réinitialisation à chaque itération).
- **Nouveau :** L'étoile a une sémantique différente! “`ab`”.`match(/(a?b??)*/)`

Fonctionnalité	État de l'art linéaire (V8)	Mes nouveaux algorithmes
Quantificateurs nullables (*,+)	incorrect	$O(r \times s)$

$|r|$: taille de la regex

$|s|$: taille de la chaîne de caractères

Les spécificités sémantiques de JavaScript

- Les groupes de captures ont une sémantique unique (réinitialisation à chaque itération).
- **Nouveau :** L'étoile a une sémantique différente ! “`ab`”.`match(/(a?b??)*/)`

Fonctionnalité	État de l'art linéaire (V8)	Mes nouveaux algorithmes
Quantificateurs nullables (*,+)	incorrect	$O(r \times s)$
Groupes de capture quantifiés	$O(r ^2 \times s)$	$O(r \times s)$
Plus non nullable	$O(2^{ r } \times s)$	$O(r \times s)$
Plus nullable greedy	$O(2^{ r } \times s)$	$O(r \times s)$

$|r|$: taille de la regex

$|s|$: taille de la chaîne de caractères

Les spécificités sémantiques de JavaScript

- Les groupes de captures ont une sémantique unique (réinitialisation à chaque itération).
- **Nouveau :** L'étoile a une sémantique différente ! “`ab`”.`match(/(a?b??)*/)`

Fonctionnalité	État de l'art linéaire (V8)	Mes nouveaux algorithmes
Quantificateurs nullables (*,+)	incorrect	$O(r \times s)$
Groupes de capture quantifiés	$O(r ^2 \times s)$	$O(r \times s)$
Plus non nullable	$O(2^{ r } \times s)$	$O(r \times s)$
Plus nullable greedy	$O(2^{ r } \times s)$	$O(r \times s)$
Lookaheads et Lookbehinds	non supporté	$O(r \times s)$

Le premier algorithme linéaire pour Lookaheads et Lookbehinds !

Grâce à la sémantique des groupes de capture JavaScript.

Des restrictions applicables à d'autres langages.

$|r|$: taille de la regex

$|s|$: taille de la chaîne de caractères

Nouveaux algorithmes

Mes algorithmes sont intégrés dans V8 (2500 lignes de C++) :



Nouvelle sémantique

Thèse de master encadrée : une sémantique en Coq/Rocq pour les regex JavaScript.

🏅 compétition étudiante de PLDI.

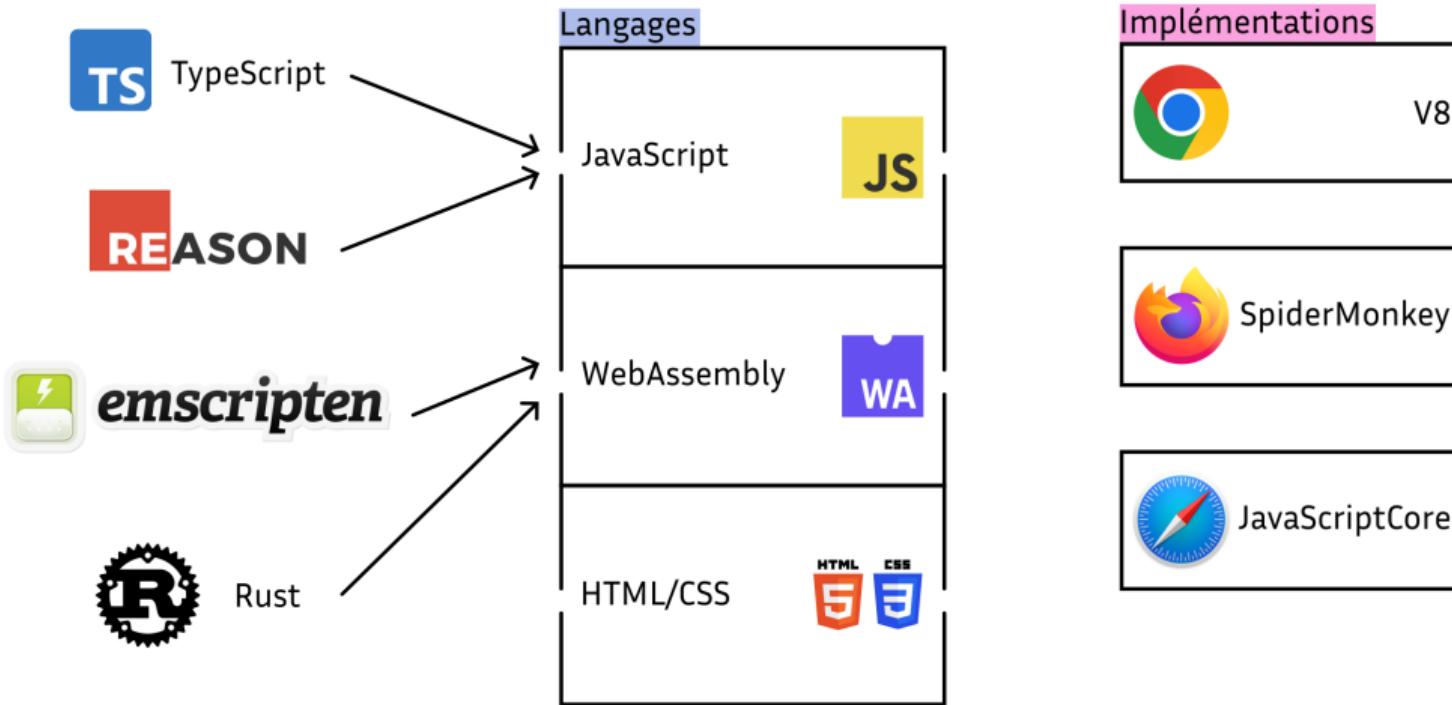
On peut enfin vérifier formellement des moteurs de regex JavaScript !

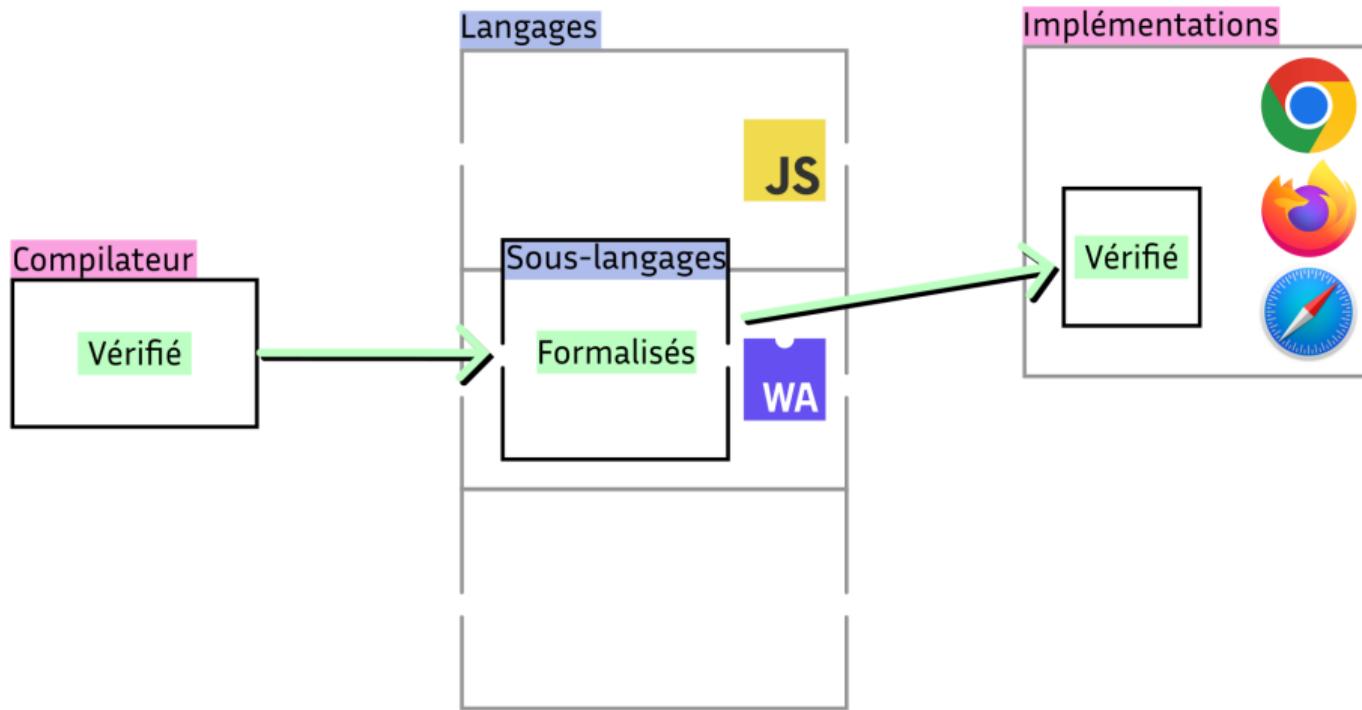
Publications : [PLDI'24], [ICFP'24]
Encadrement de 12 projets d'étudiants.

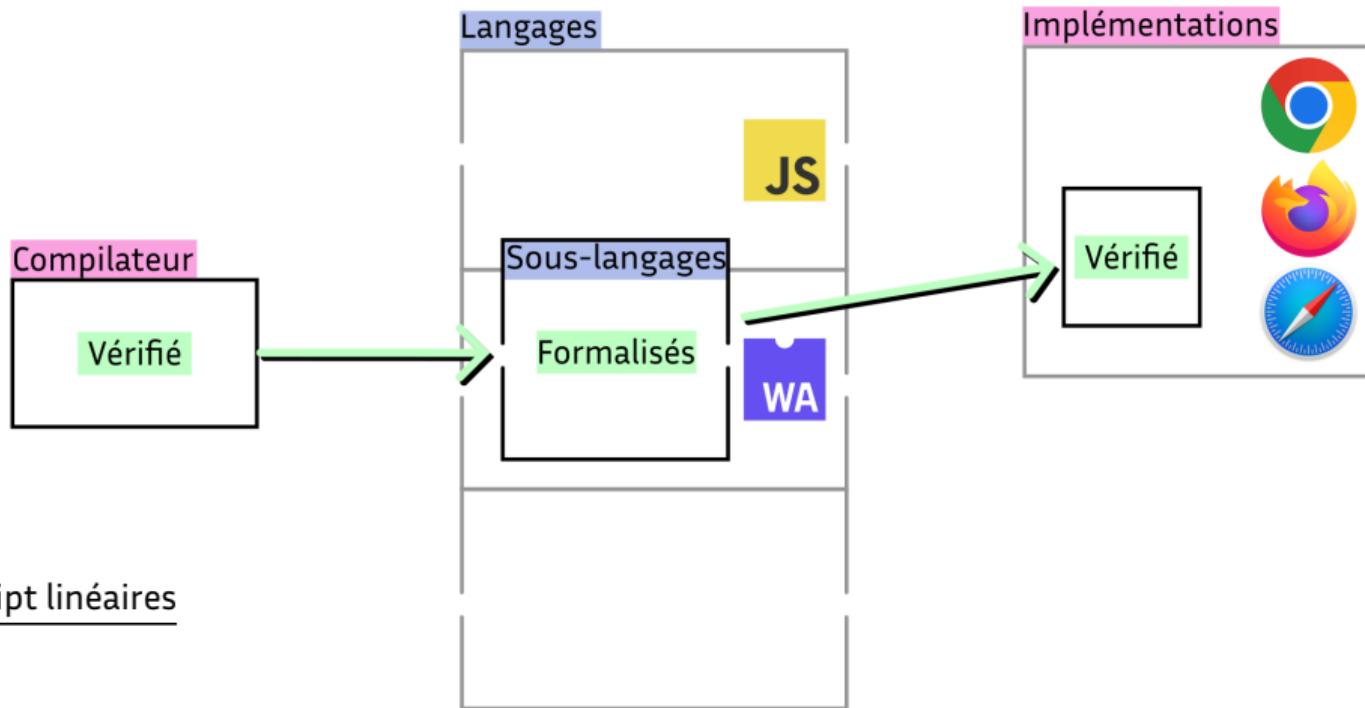
Financements :
Open Research Data Contribute Grant (32 000€)
Swiss National Science Foundation Project (497 000€)

Programme de Recherche

Vers une plateforme web de confiance : vérification formelle de compilateurs et d'environnements d'exécution



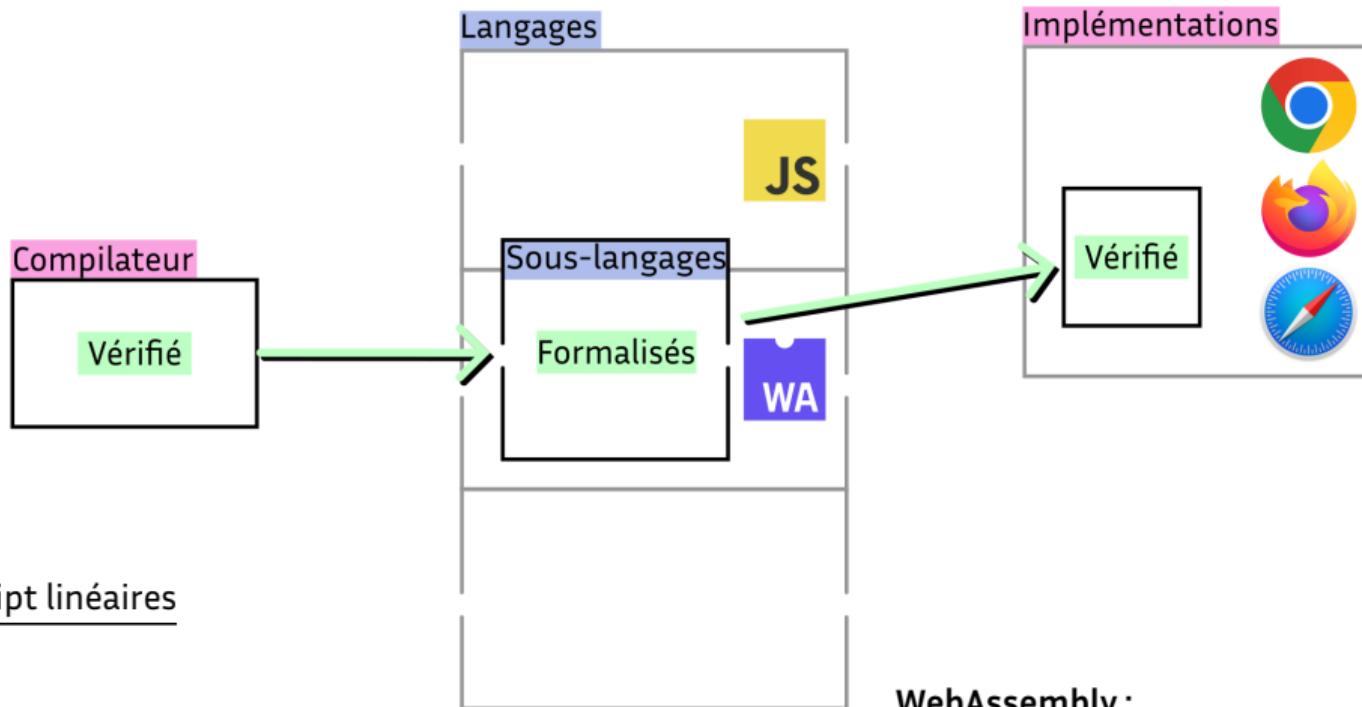




Deux axes :

Regex JavaScript linéaires

Un sous-ensemble de WebAssembly



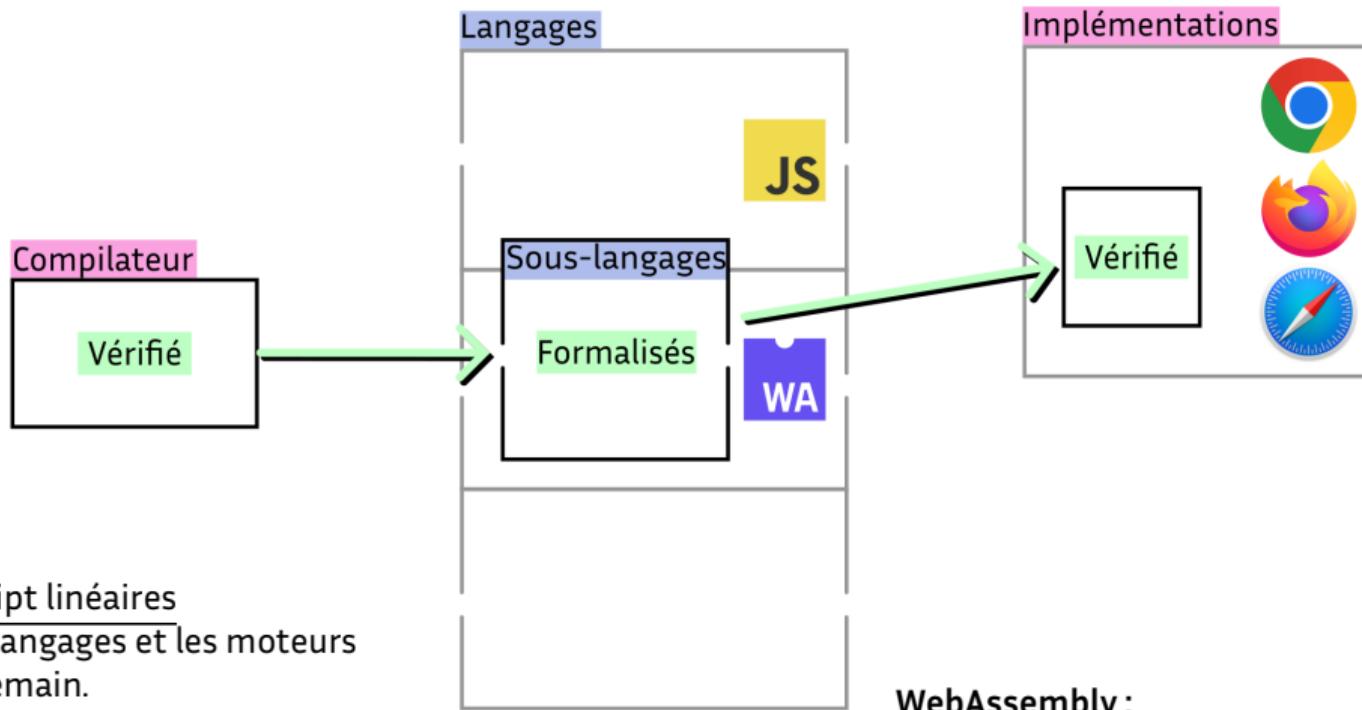
Deux axes :

Regex JavaScript linéaires

Un sous-ensemble de WebAssembly

WebAssembly :

- Bytecode bas niveau.
- Sémantique isolant chaque module.
- Avec une sémantique formelle.



Deux axes :

Regex JavaScript linéaires

Concevoir les langages et les moteurs de regex de demain.

Un sous-ensemble de WebAssembly

Compilation formellement vérifiée pour des programmes compartimentés.

WebAssembly :

- Bytecode bas niveau.
- Sémantique isolant chaque module.
- Avec une sémantique formelle.

Regex JavaScript

Sous-ensemble linéaire

JS

Regex JavaScript

Sous-ensemble linéaire

Lookaheads

Lookbehinds

Groupes de capture

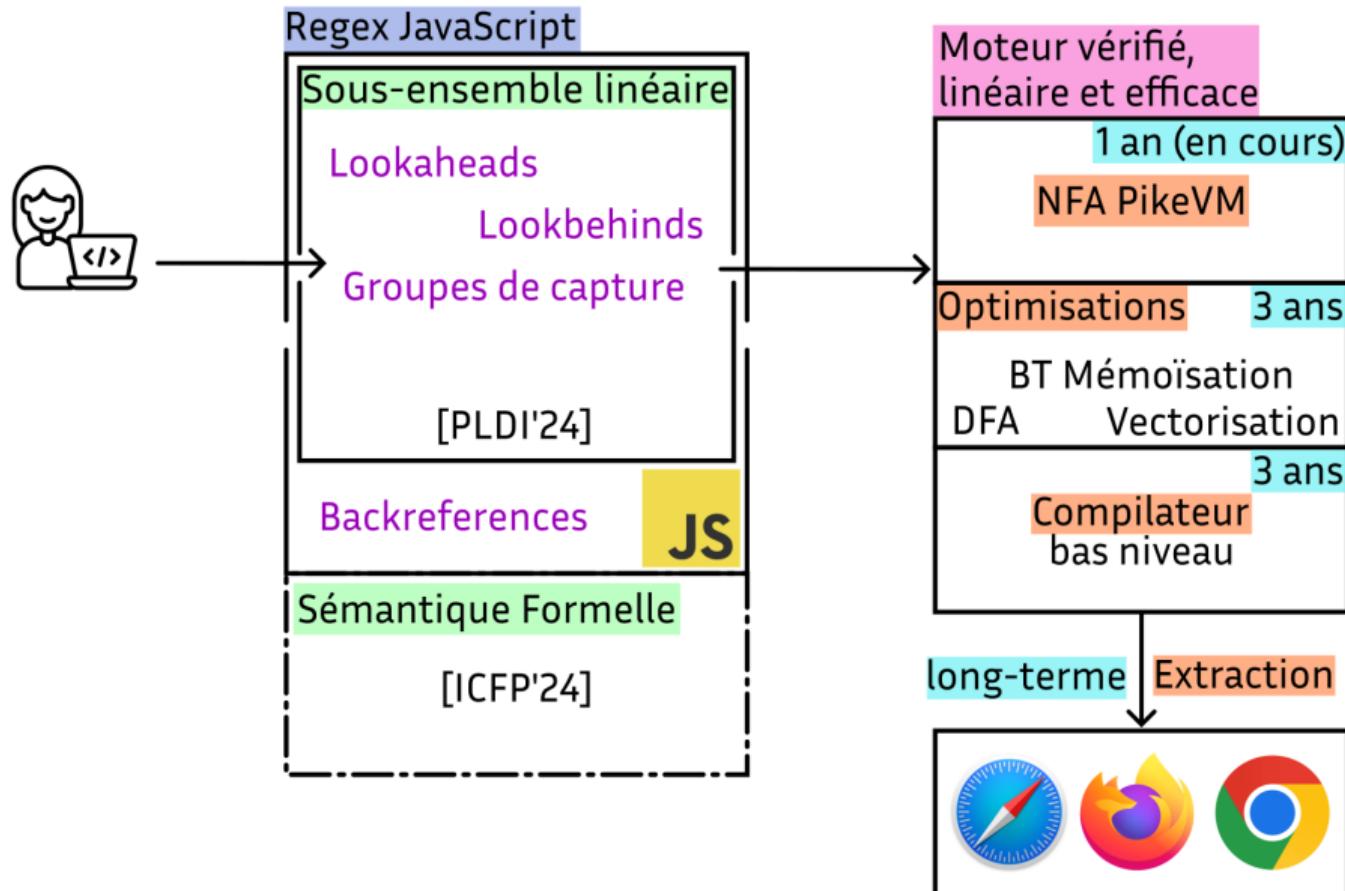
[PLDI'24]

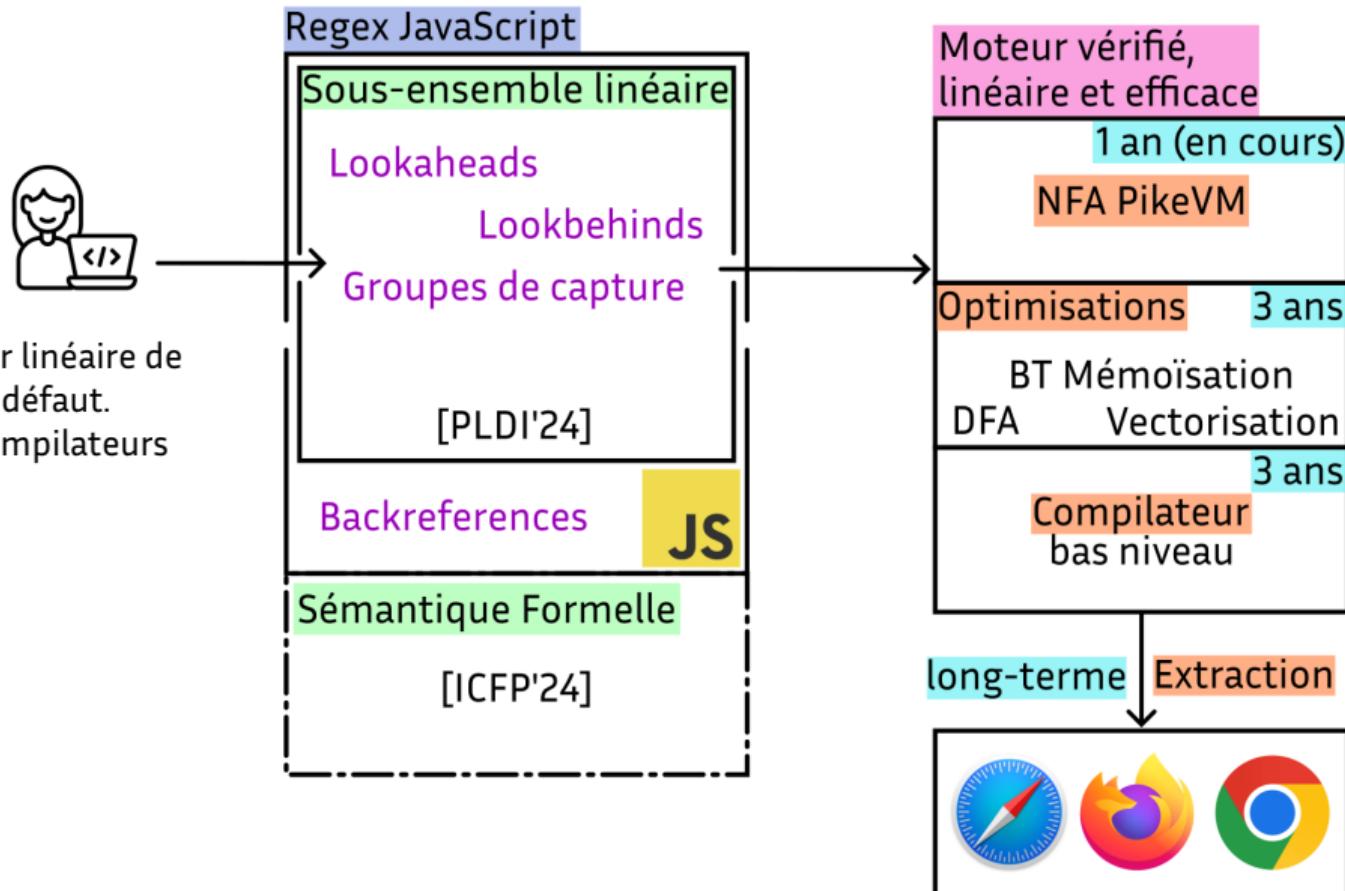
Backreferences

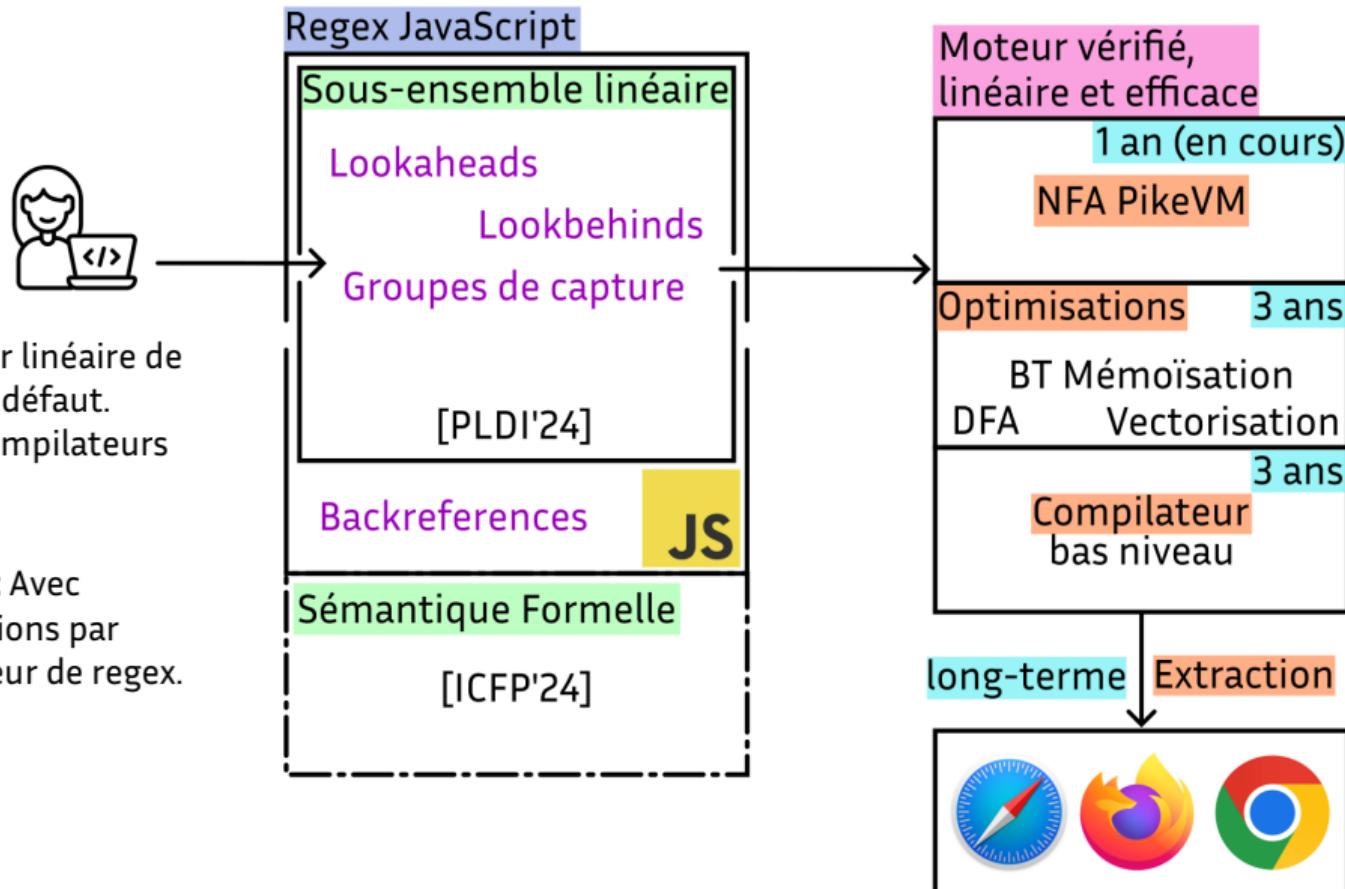
JS

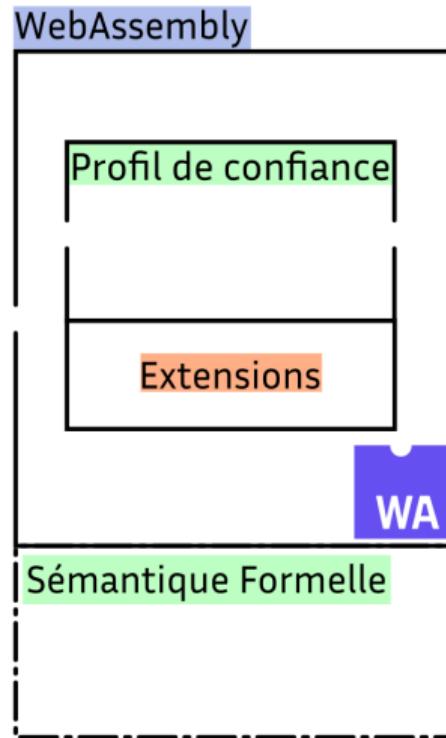
Sémantique Formelle

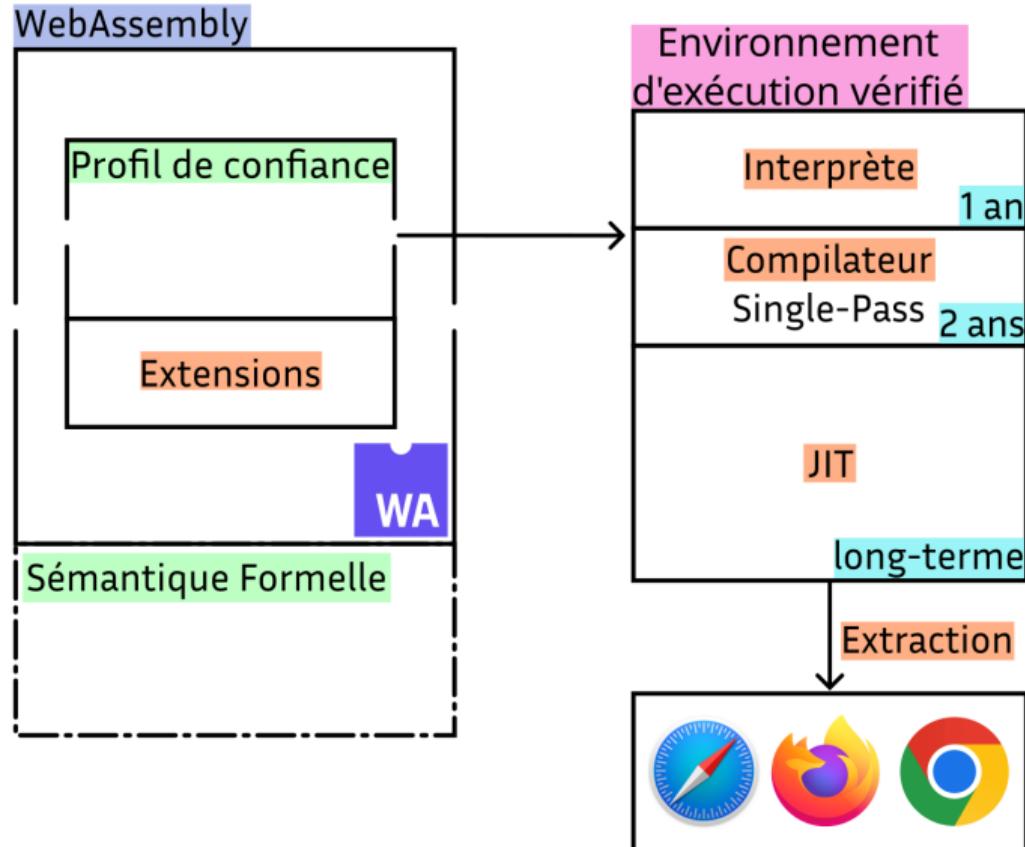
[ICFP'24]

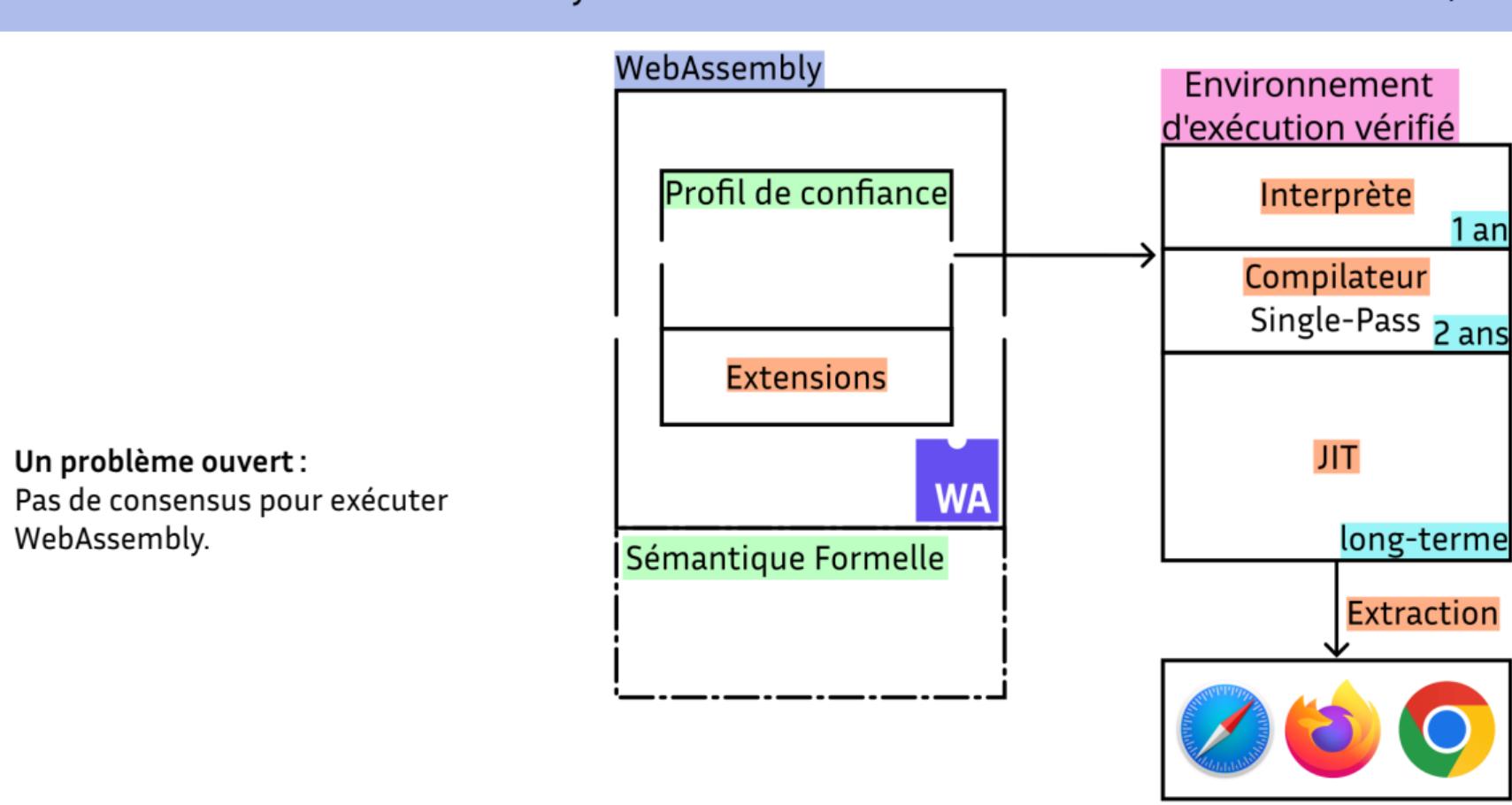






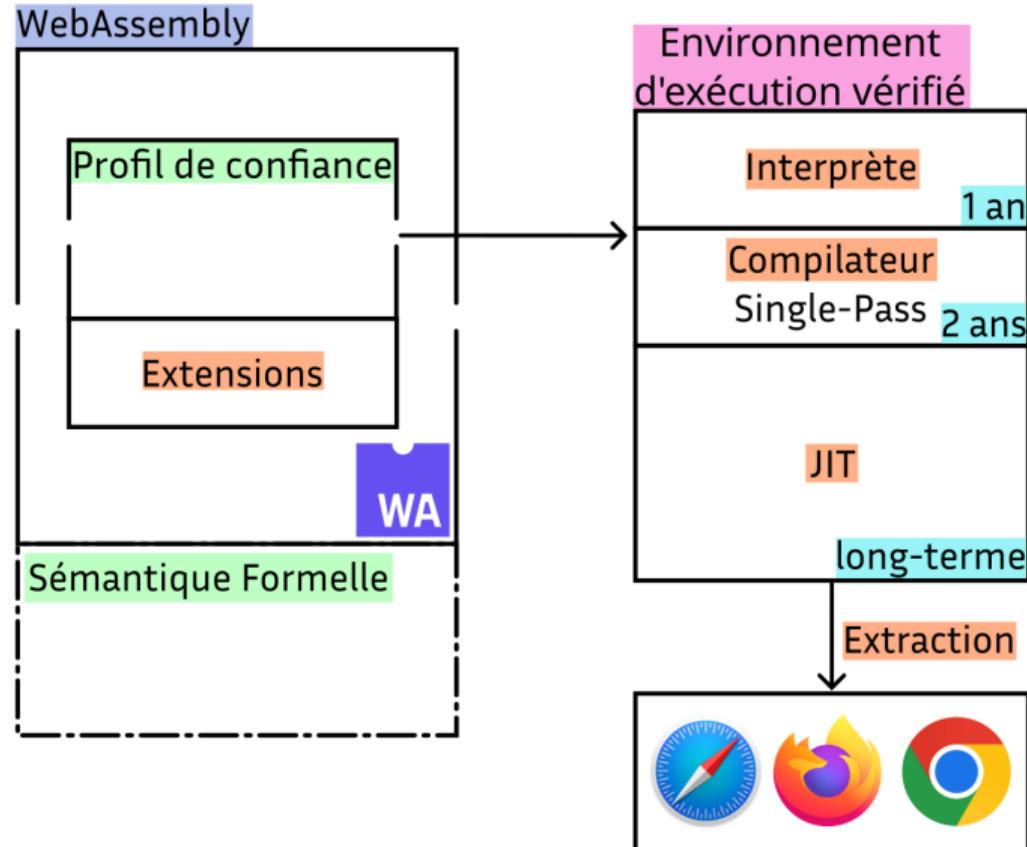


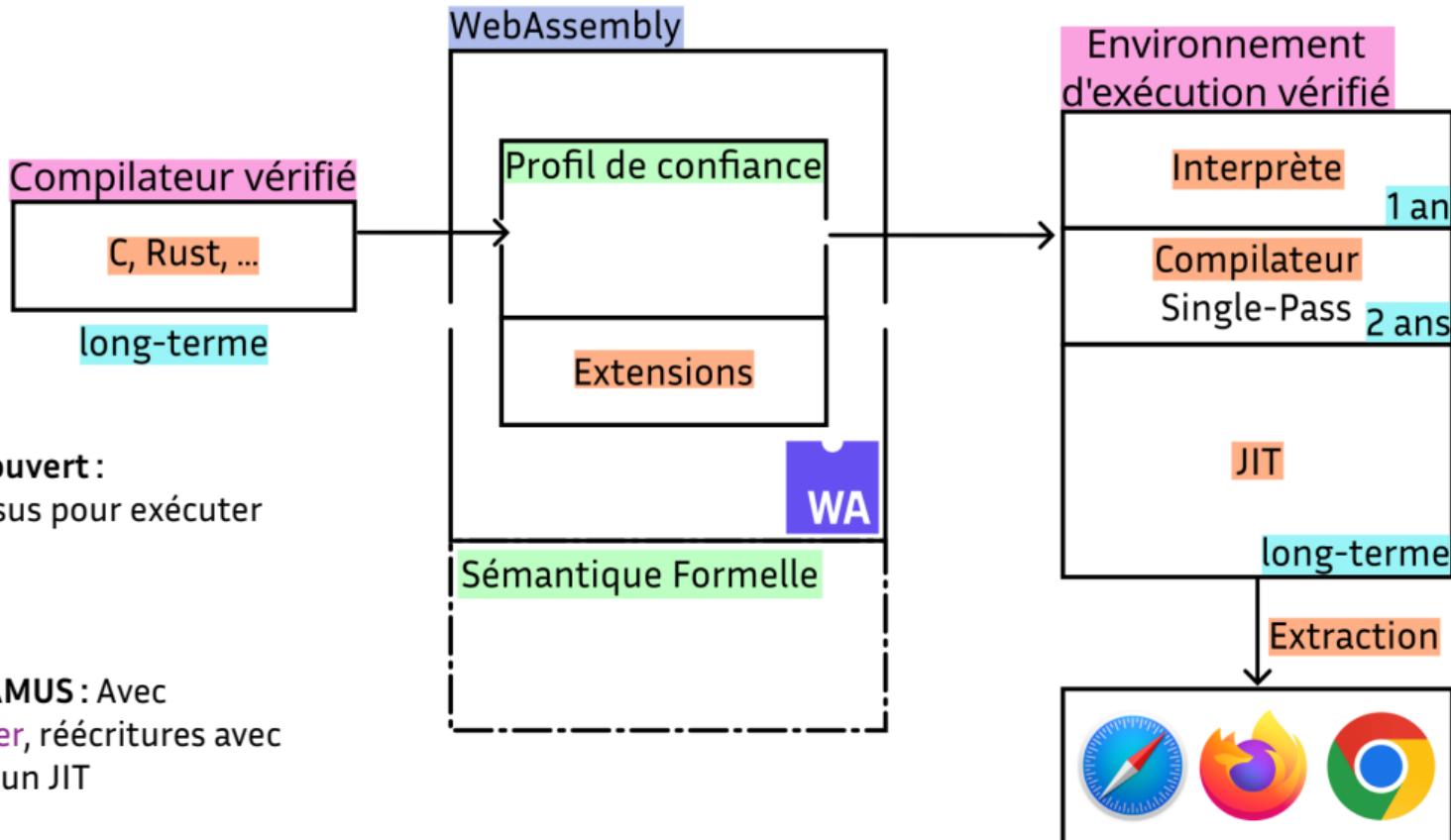




Un problème ouvert :
Pas de consensus pour exécuter
WebAssembly.

Intégration CAMUS : Avec
[Thomas Koehler](#), réécritures avec
e-graphs pour un JIT
WebAssembly.





L'équipe CAMUS : de nombreux thèmes communs

Optimisations vérifiées

Arthur Charguéraud et Thomas Koehler

Utiliser OptiTrust pour optimiser un moteur de regex vérifié.

Sémantiques de langages

Arthur Charguéraud et Jens Gustedt

Formaliser les sémantiques des langages exécutés.

Vectorisation

Bérenger Bramas et Thomas Koehler

Optimiser un moteur de regex.

Compilation avec spécialisation

Philippe Clauss

Des similarités avec la spéculation des JITs.

Preuves de complexité asymptotique

Arthur Charguéraud

Prouver la linarité d'un moteur de regex.

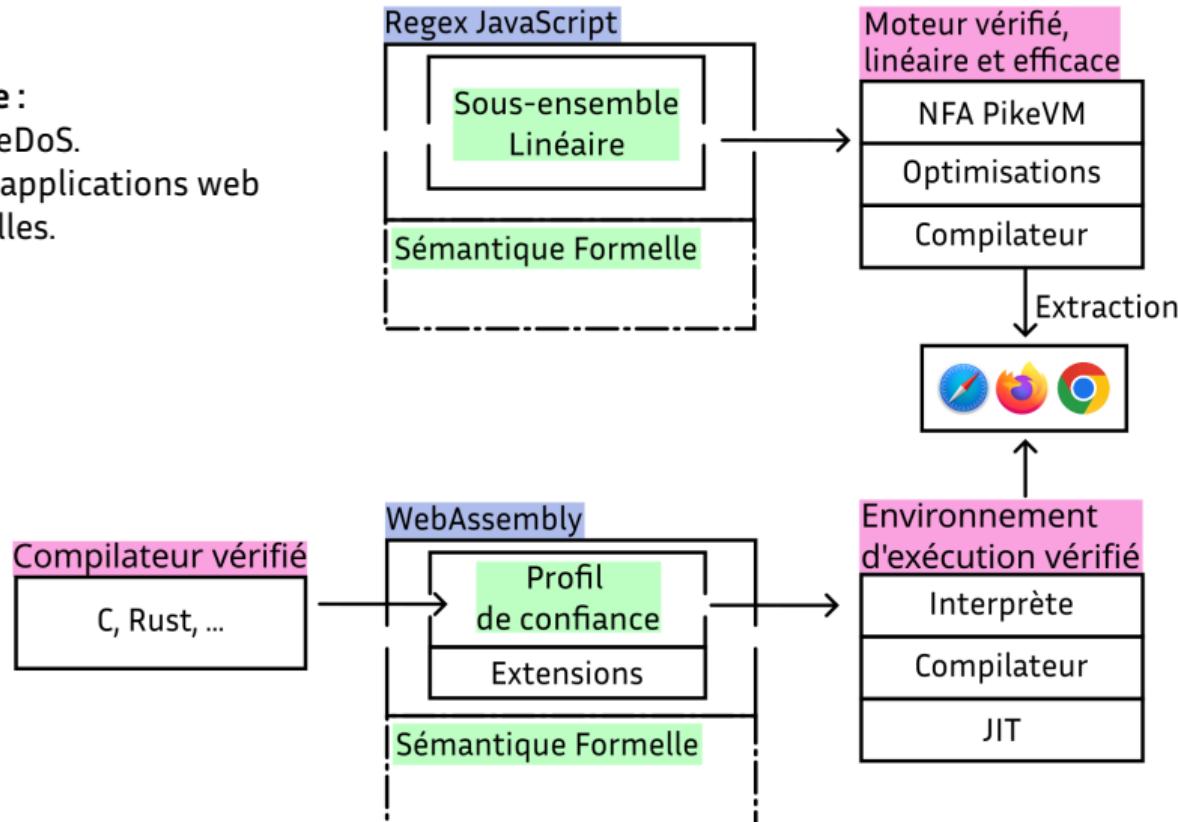
Nouvelle expertise apportée

JITs, Regex, Compilation formellement vérifiée.

Programme de recherche :

Vaincre la vulnérabilité ReDoS.

Déployer et exécuter des applications web avec des garanties formelles.



Programme de recherche :

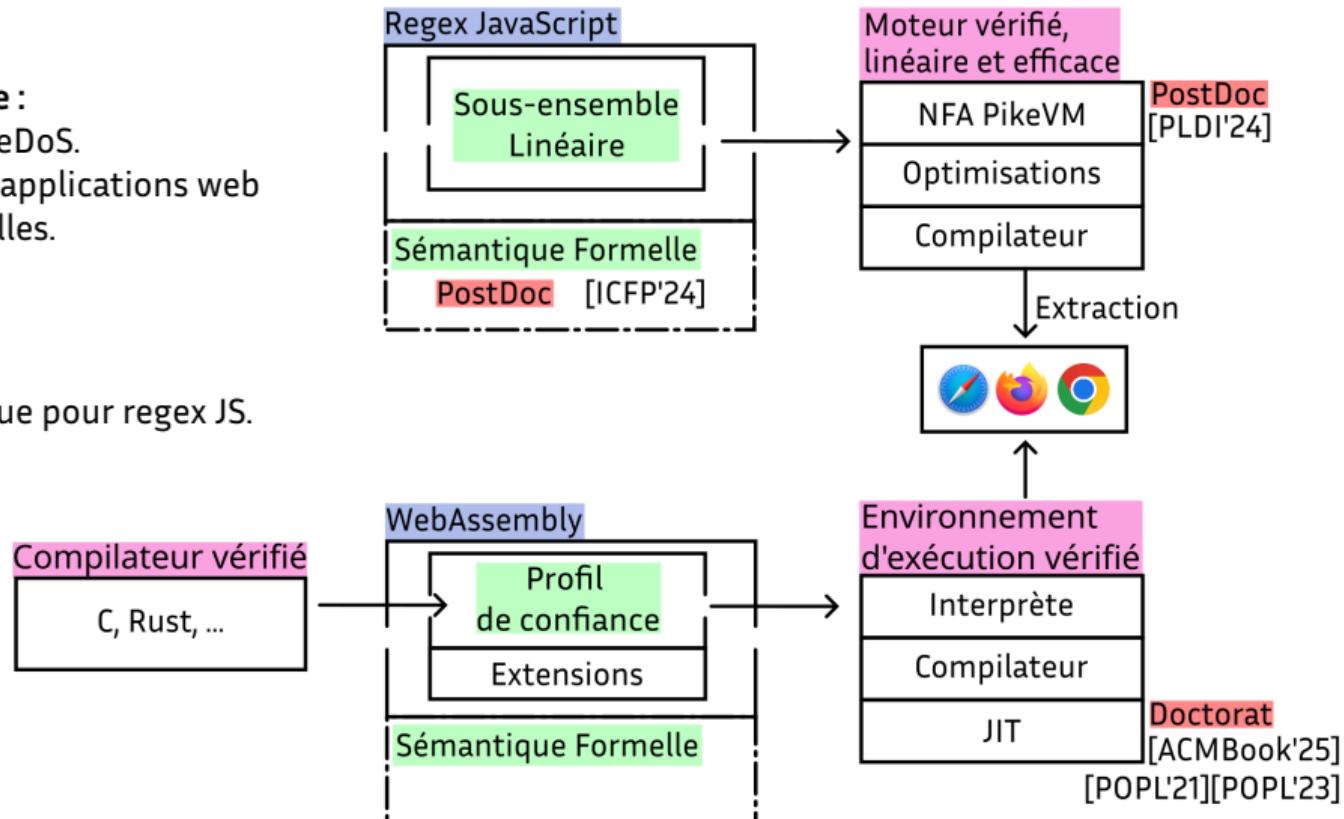
Vaincre la vulnérabilité ReDoS.

Déployer et exécuter des applications web avec des garanties formelles.

Travaux précédents :

JITs vérifiés.

Algorithmes et sémantique pour regex JS.



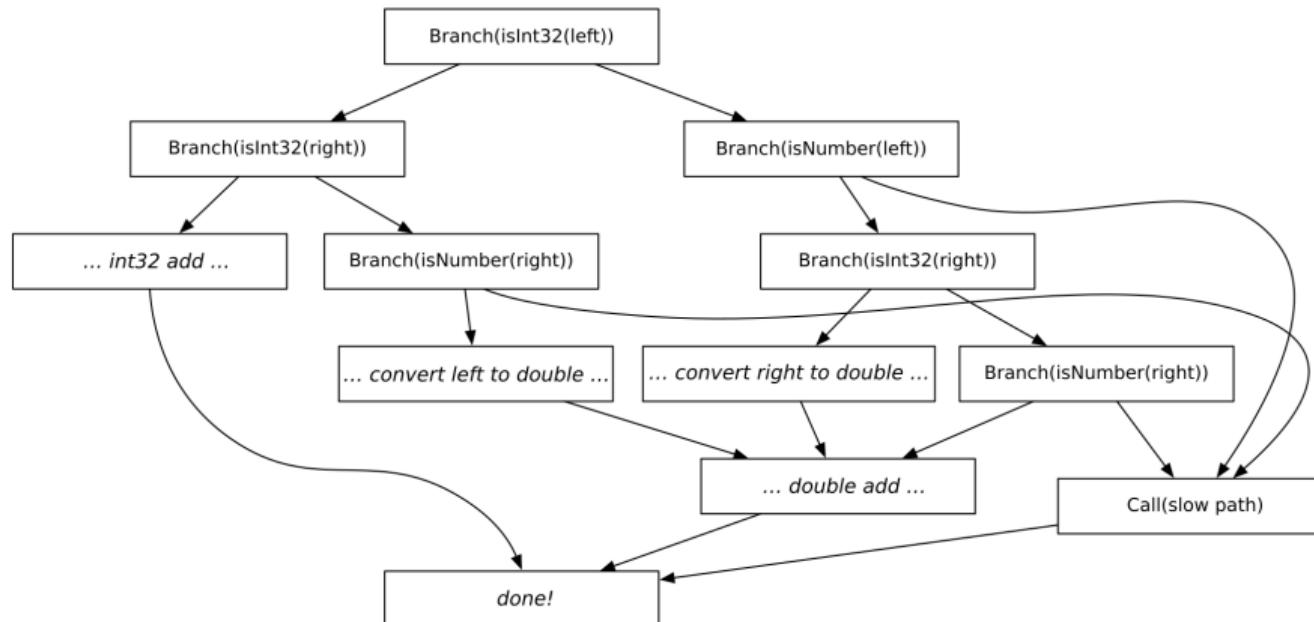
Transparents supplémentaires

- 1/14 Quelles garanties pour le Web?
- 2/14 Mon domaine : compilation formellement vérifiée
- 3/14 Ma méthodologie
- 4/14 Doctorat
- 5/14 Spéculation & Déoptimisation
- 6/14 JITs vérifiés
- 7/14 PostDoc
- 8/14 Nouveaux algorithmes linéaires
- 9/14 Avancements algorithmiques et sémantiques
- 10/14 Vers une plateforme web de confiance
- 11/14 Un moteur vérifié, linéaire, efficace, intégrable
- 12/14 Un sous-ensemble WebAssembly de confiance
- 13/14 Intégration
- 14/14 Vérification formelle pour un Web de confiance

Transparents supplémentaires :
Spéculer dans un langage dynamique
Insérer des instructions spéculatives
Définition Simulations Imbriquées
Simulations imbriquées, exécution
Simulations imbriquées, optimisation
Regex modernes avec priorité
L'étoile JavaScript est unique
Simulation de NFA et étoile
Dupliquer le graphe pour l'étoile JavaScript
Lookarounds et groupes de capture
3 étapes pour les lookarounds
Sémantique formelle pour les regex JavaScript
Une mécanisation de confiance

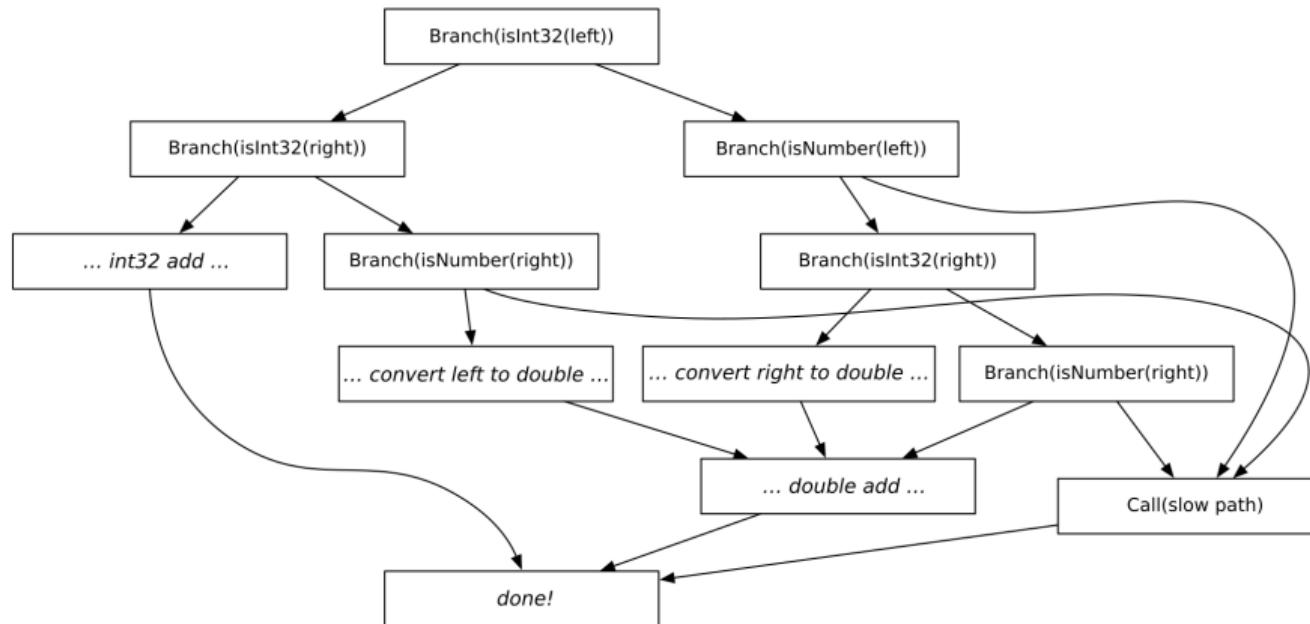
Pour exécuter `left + right`, un exemple issu de JavaScriptCore.

Pour exécuter `left + right`, un exemple issu de JavaScriptCore.



Difficile à compiler.
Restreint des optimisations.

Pour exécuter `left + right`, un exemple issu de JavaScriptCore.



Difficile à compiler.
Restreint des optimisations.

Spéculation de type

En spéculant sur les types de `left` et `right`, le graphe est réduit à un seul nœud. Un seul test si les variables sont utilisées plusieurs fois.

```
Function F (x, y):
```

```
Version Base:
```

```
l1: a ← 1
```

```
b ← y + x
```

```
Return (a + b)
```

```
Function F (x, y):  
Version Base:  
l1: a ← 1
```

b ← y + x
Return (a + b)

JIT : il est possible qu'on spécle ici plus tard

Function F (x, y):
Version Base:
l1: a ← 1

Version Opt:
l1: a ← 1
Anchor F.l1 [x ← x, y ← y]

b ← y + x
b ← y + x

Return (a + b)

Return (a + b)

Timeline :

- Créer une version Opt, avec des points de synchronisation, Anchor

JIT : il est possible qu'on spécle ici plus tard

Function F (x, y):
Version Base:
l1: a ← 1

b ← y + x

Return (a + b)

Version Opt:
l1: a ← 1
Anchor F.l1 [x ← x, y ← y]

b ← y + x

Return (a + b)

Timeline :

- Créer une version Opt, avec des points de synchronisation, Anchor

JIT : spéculons que x est égal à 9

Function F (x, y):

Version Base:

l1: a ← 1

b ← y + x

Return (a + b)

Version Opt:

l1: a ← 1

Anchor F.l1 [x ← x, y ← y]

Assume (x=9) F.l1 [x ← x, y ← y]

b ← y + x

Return (a + b)

Timeline :

- Créer une version Opt, avec des points de synchronisation, Anchor
- Insérer spéculation, Assume

JIT : spéculons que x est égal à 9

Function F (x, y):

Version Base:

l1: a ← 1

b ← y + x

Return (a + b)

Version Opt:

l1: a ← 1

Anchor F.l1 [x ← x, y ← y]

Assume (x=9) F.l1 [x ← x, y ← y]

b ← y + 9

Return (1 + b)

Timeline :

- Créer une version Opt, avec des points de synchronisation, Anchor
- Insérer spéculation, Assume
- Propagation de constantes

Function F (x, y):

Version Base:

l1: a ← 1

b ← y + x

Return (a + b)

Version Opt:

l1: a ← 1

Anchor F.l1 [x ← x, y ← y]

Assume (x=9) F.l1 [x ← x, y ← y]

Assume (y=7) F.l1 [x ← x, y ← y]

b ← y + 9

Return (1 + b)

Timeline :

- Créer une version Opt, avec des points de synchronisation, Anchor
- Insérer spéculation, Assume
- Propagation de constantes
- Insérer spéculation

JIT : spéculons que y est égal à 7

Function F (x, y):

Version Base:

l1: a ← 1

b ← y + x

Return (a + b)

Version Opt:

l1: a ← 1

Anchor F.l1 [x ← x, y ← y]

Assume (x=9) F.l1 [x ← x, y ← y]

Assume (y=7) F.l1 [x ← x, y ← y]

b ← 16

Return (17)

Timeline :

- Créer une version Opt, avec des points de synchronisation, Anchor
- Insérer spéculation, Assume
- Propagation de constantes
- Insérer spéculation
- Propagation de constantes

Function F (x, y):

Version Base:

l1: a ← 1

b ← y + x

Return (a + b)

Version Opt:

Anchor F.l1 [x ← x, y ← y]
Assume (x=9) F.l1 [x ← x, y ← y]
Assume (y=7) F.l1 [x ← x, y ← y]

Return (17)

Timeline :

- Créer une version Opt, avec des points de synchronisation, Anchor
- Insérer spéculation, Assume
- Propagation de constantes
- Insérer spéculation
- Propagation de constantes
- Élimination de code mort

Function F (x, y):

Version Base:

l1: a ← 1

b ← y + x

Return (a + b)

Version Opt:

Anchor F.l1 [x ← x, y ← y]
Assume (x=9) F.l1 [x ← x, y ← y]
Assume (y=7) F.l1 [x ← x, y ← y]

Return (17)

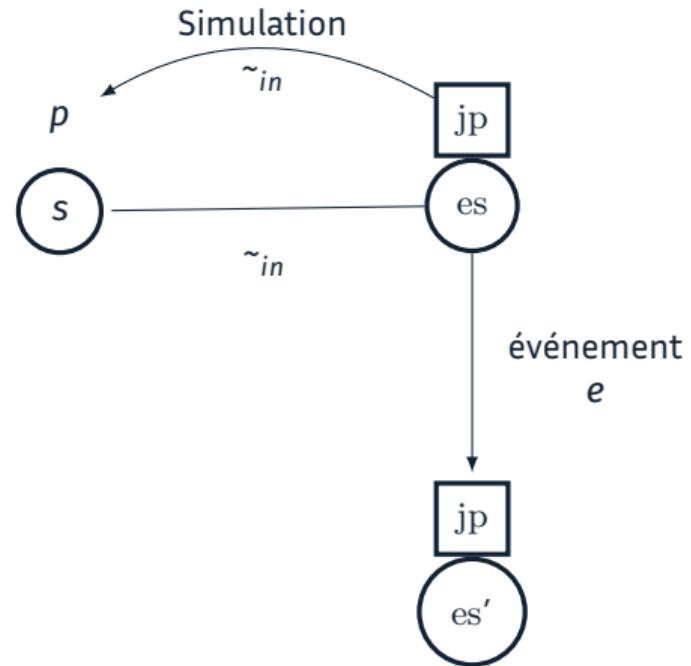
Timeline :

- Créer une version Opt, avec des points de synchronisation, Anchor
- Insérer spéculation, Assume
- Propagation de constantes
- Insérer spéculation
- Propagation de constantes
- Élimination de code mort

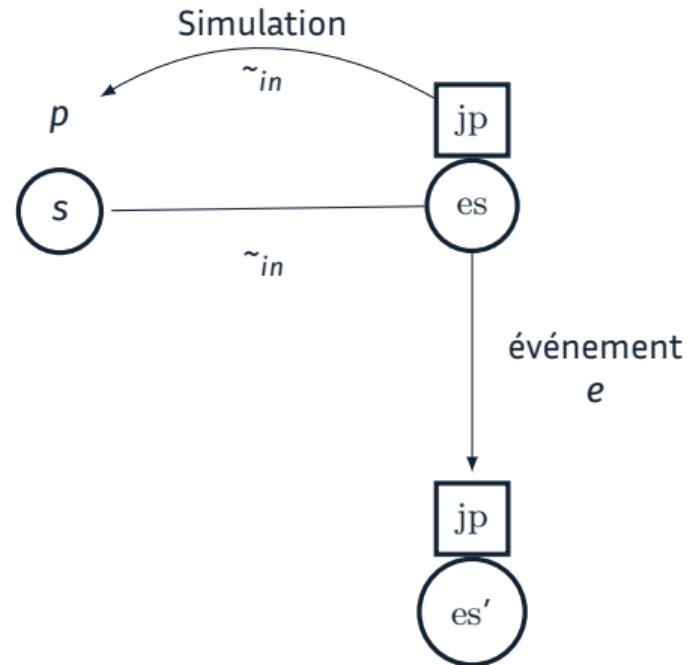
Spéculation vérifiée

- Définir la sémantique formelle des instructions spéculatives
- Prouver des simulations pour chaque étape (et plus encore : inlining,...).

Exécution (par ex. interprète) :
état d'exécution es mis à jour.

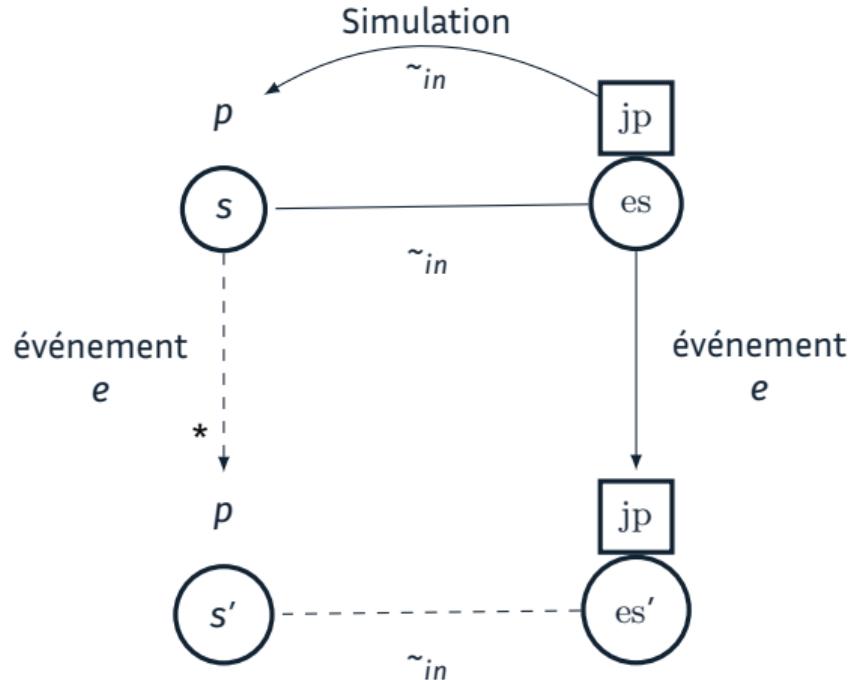


Exécution (par ex. interprète) :
état d'exécution es mis à jour.



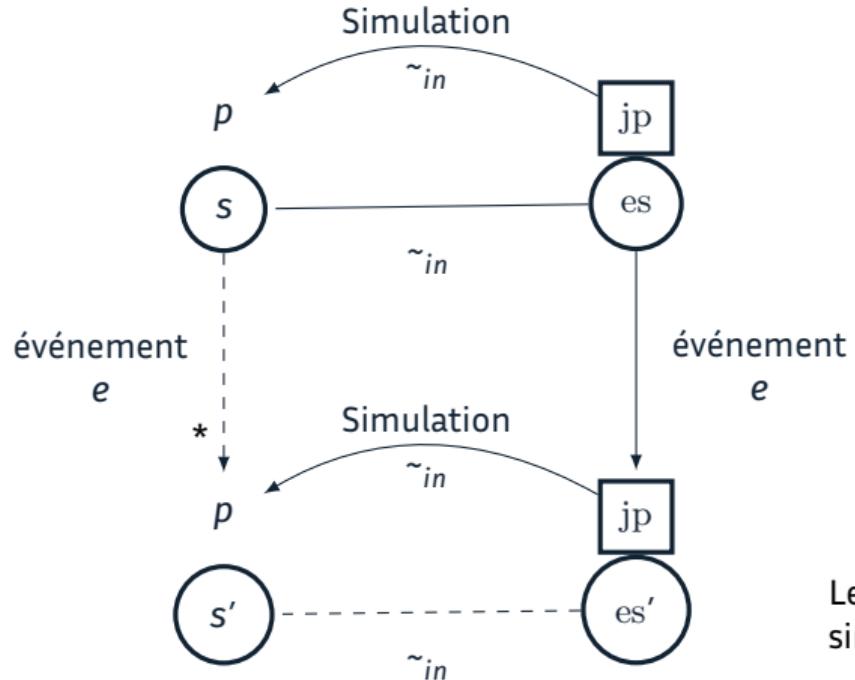
Exécution (par ex. interprète) :
état d'exécution es mis à jour.

On utilise la simulation interne
pour trouver une exécution
équivalente du source.



Exécution (par ex. interprète) :
état d'exécution es mis à jour.

On utilise la simulation interne
pour trouver une exécution
équivalente du source.



Le programme jp est toujours
simulé avec p.

(1) Initialisation dynamique

$$\forall s_y, \text{ si } s_y \text{ est un état de synchronisation, alors } s_y \sim_{int} s_y$$
(2) Préservation de progrès

$$\forall s_1 s'_1 t s_2, s_1 \sim_{int} s_2 \wedge s_1 \xrightarrow[p_1]{t} s'_1 \implies \exists t' s'_2, s_2 \xrightarrow[p]{t'} s'_2$$

(3) Diagramme interne

$$\forall s_1 s_2 s'_2 t, s_1 \sim_{int} s_2 \wedge s_2 \xrightarrow[p]{t} s'_2 \implies$$

$$(\exists s'_1, s_1 \xrightarrow[p_1]{t} s'_1 \wedge s'_1 \sim_{int} s'_2) \vee$$

$$(s_1 \sim_{int} s'_2 \wedge m_{int}(s'_2) < m_{int}(s_2) \wedge t = \emptyset)$$

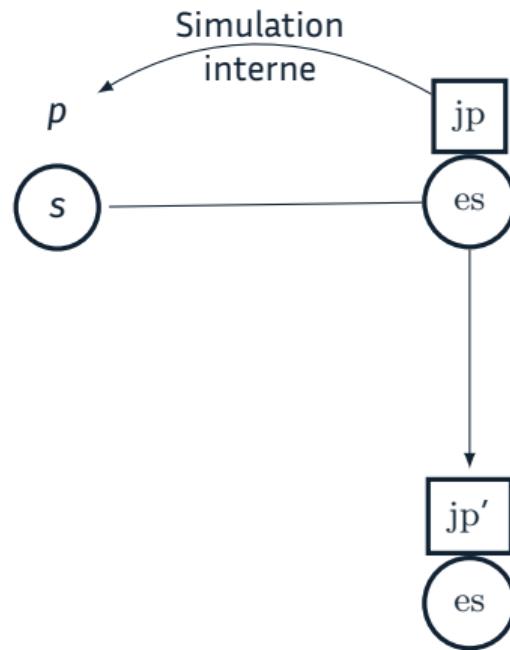
Simulation Interne \sim_{int} m_{int} p_1 p

$s \sim_{int} e$

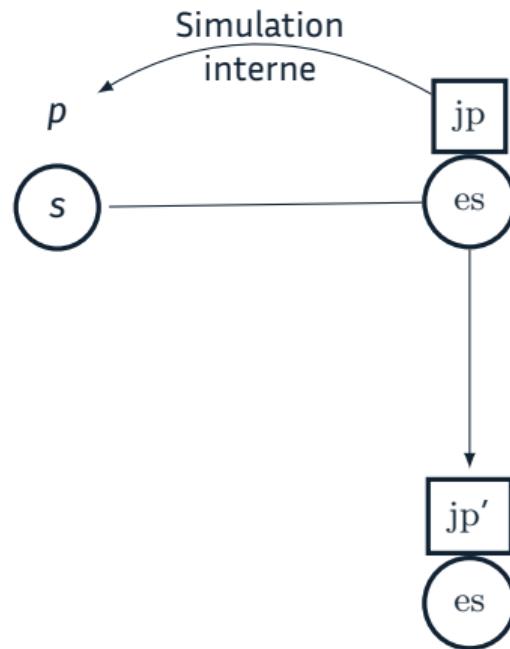
Simulation Interne \sim_{int} m_{int} p_1 p

$s \sim_{ext} (e, p, n, ps)$

Optimisation : programme jp
mis à jour.



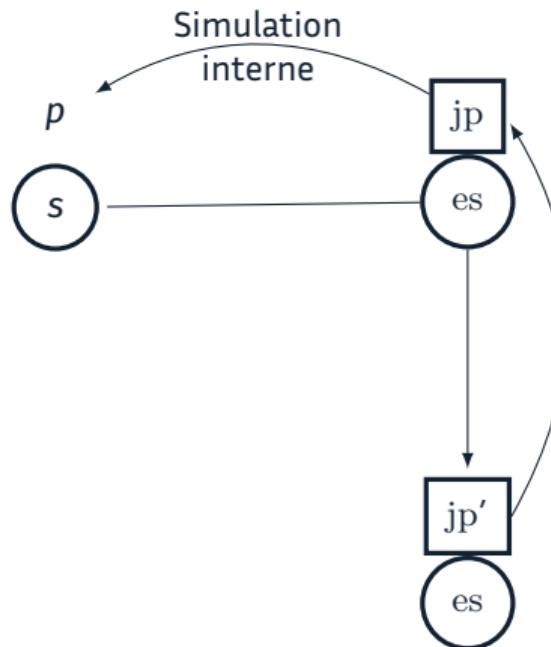
Optimisation : programme jp
mis à jour.



Il faut prouver que jp' est
toujours simulé avec p .

Optimisation : programme jp mis à jour.

On prouve l'optimisation dynamique correcte avec une simulation.

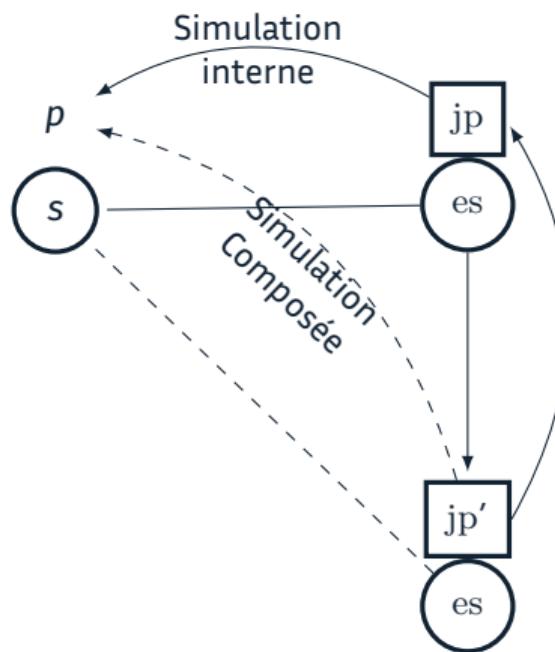


Il faut prouver que jp' est toujours simulé avec p .

Theorem optimizer_correct:
forall jp, jp',
optimizer jp = jp' →
backward_simulation jp jp'.

Optimisation : programme jp mis à jour.

On prouve l'optimisation dynamique correcte avec une simulation.



Il faut prouver que jp' est toujours simulé avec p .

Theorem `optimizer_correct`:
 $\forall jp\ jp',$
`optimizer` $jp = jp' \rightarrow$
`backward_simulation` $jp\ jp'$.

On compose cette simulation avec la simulation interne existante.

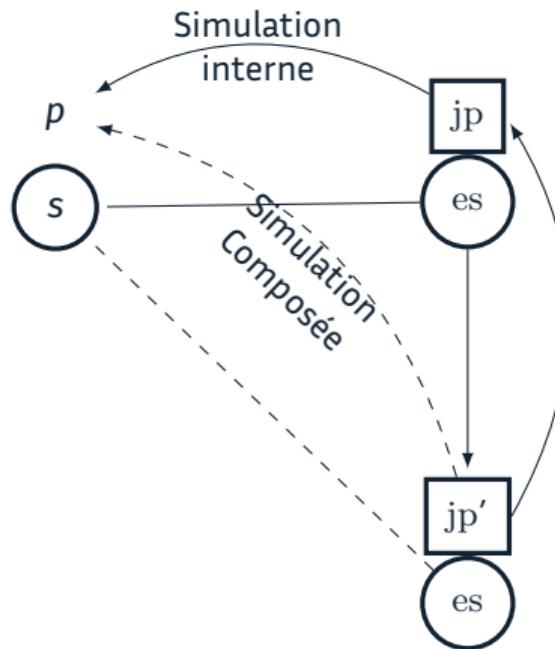
Optimisation : programme jp mis à jour.

On prouve l'optimisation dynamique correcte avec une simulation.

Conclusion

Avec cette technique, on prouve une optimisation dynamique avec une simulation, **comme dans le cas statique!**

Il faut prouver que jp' est toujours simulé avec p .



Theorem optimizer_correct:
 $\forall jp \, jp', \text{optimizer } jp = jp' \rightarrow \text{backward_simulation } jp \, jp'$

On compose cette simulation avec la simulation interne existante.

Matching ambigu

Dans les langages modernes (JavaScript, Python, PCRE, RE2, Rust), il faut retourner plus qu'un booléen.

- Match de $d \mid b$ sur "abcd" = "b".
- Match de $a \mid ab$ sur "abc" = "a".
- Match de ab^* sur "abbccc" = "abb".

Quel match choisir ?

Matching ambigu

Dans les langages modernes (JavaScript, Python, PCRE, RE2, Rust), il faut retourner plus qu'un booléen.

- Match de $d \mid b$ sur "abcd" = "b".
- Match de $a \mid ab$ sur "abc" = "a".
- Match de ab^* sur "abbccc" = "abb".

Quel match choisir ?

Règles de priorité

- Le match qui commence le plus tôt a la priorité.
- Dans $r_1 \mid r_2$, r_1 a la priorité (*non commutatif!*).
- Pour les quantificateurs *greedy* (*, +), priorité au nombre maximum d'itérations.

Matching ambigu

Dans les langages modernes (JavaScript, Python, PCRE, RE2, Rust), il faut retourner plus qu'un booléen.

- Match de $d \mid b$ sur "abcd" = "b".
- Match de $a \mid ab$ sur "abc" = "a".
- Match de ab^* sur "abbccc" = "abb".

Quel match choisir ?

Règles de priorité

- Le match qui commence le plus tôt a la priorité.
- Dans $r_1 \mid r_2$, r_1 a la priorité (*non commutatif!*).
- Pour les quantificateurs *greedy* (*, +), priorité au nombre maximum d'itérations.

Que faire pour ϵ^* ? La sémantique des quantificateurs doit éviter les répétitions infinies .

Deux manières d'éviter les boucles infinies

La majorité des langages : boucles d'ε interdites.

JavaScript : Les itérations ne peuvent pas matcher la chaîne vide..

Deux manières d'éviter les boucles infinies

La majorité des langages : boucles d' ϵ interdites.

JavaScript : Les itérations ne peuvent pas matcher la chaîne vide..

(($a \mid \epsilon$) ($\epsilon \mid b$))^{*} on "ab"

Deux manières d'éviter les boucles infinies

La majorité des langages : boucles d' ϵ interdites.

JavaScript : Les itérations ne peuvent pas matcher la chaîne vide..

($(a \mid \epsilon)$ $(\epsilon \mid b)$) * on "ab"
 \xrightarrow{a}

Deux manières d'éviter les boucles infinies

La majorité des langages : boucles d' ϵ interdites.

JavaScript : Les itérations ne peuvent pas matcher la chaîne vide..

($(a \mid \epsilon)$ $(\epsilon \mid b)$) * on "ab"
 \xrightarrow{a} $\xrightarrow{\epsilon}$

Deux manières d'éviter les boucles infinies

La majorité des langages : boucles d' ϵ interdites.

JavaScript : Les itérations ne peuvent pas matcher la chaîne vide..

($(a \mid \epsilon)$ $(\epsilon \mid b)$) * on "ab"

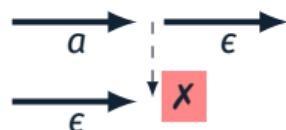


Deux manières d'éviter les boucles infinies

La majorité des langages : boucles d' ϵ interdites.

JavaScript : Les itérations ne peuvent pas matcher la chaîne vide..

($(a \mid \epsilon)$ $(\epsilon \mid b)$) * on "ab"

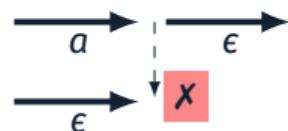


Deux manières d'éviter les boucles infinies

La majorité des langages : boucles d' ϵ interdites.

JavaScript : Les itérations ne peuvent pas matcher la chaîne vide..

($(a \mid \epsilon)$ $(\epsilon \mid b)$)* on "ab"



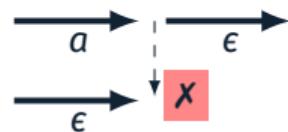
Résultat : 1 itération, matchant "a".

Deux manières d'éviter les boucles infinies

La majorité des langages : boucles d' ϵ interdites.

JavaScript : Les itérations ne peuvent pas matcher la chaîne vide..

($(a \mid \epsilon) \ (\epsilon \mid b)$)* on "ab"



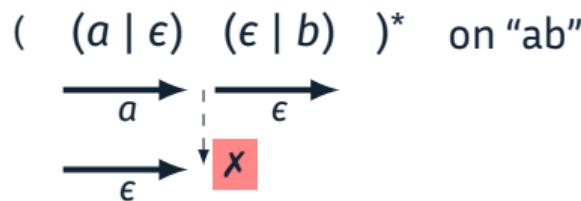
Résultat : 1 itération, matchant "a".

($(a \mid \epsilon) \ (\epsilon \mid b)$)* on "ab"

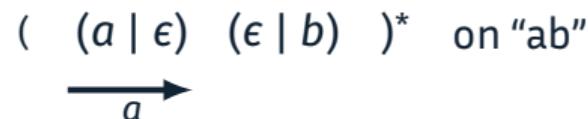
Deux manières d'éviter les boucles infinies

La majorité des langages : boucles d' ϵ interdites.

JavaScript : Les itérations ne peuvent pas matcher la chaîne vide..



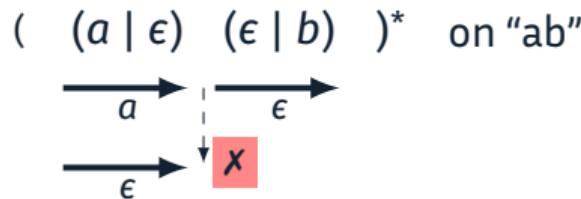
Résultat : 1 itération, matchant "a".



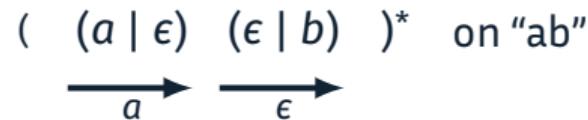
Deux manières d'éviter les boucles infinies

La majorité des langages : boucles d' ϵ interdites.

JavaScript : Les itérations ne peuvent pas matcher la chaîne vide..



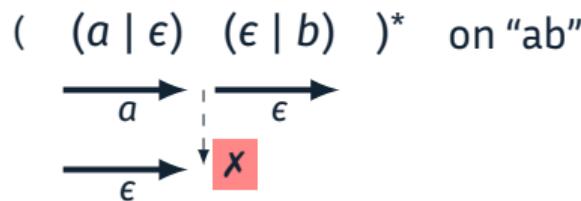
Résultat : 1 itération, matchant "a".



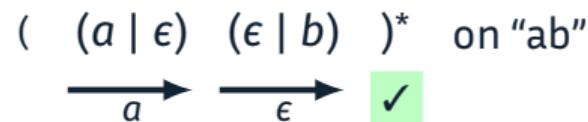
Deux manières d'éviter les boucles infinies

La majorité des langages : boucles d' ϵ interdites.

JavaScript : Les itérations ne peuvent pas matcher la chaîne vide..



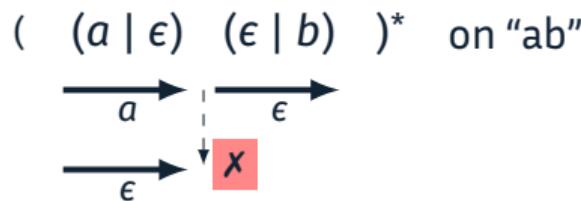
Résultat : 1 itération, matchant "a".



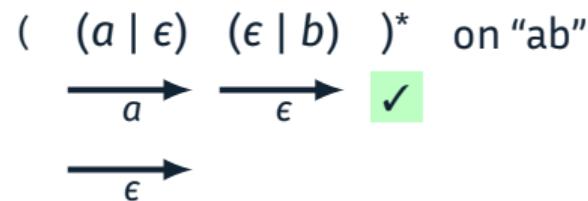
Deux manières d'éviter les boucles infinies

La majorité des langages : boucles d' ϵ interdites.

JavaScript : Les itérations ne peuvent pas matcher la chaîne vide..



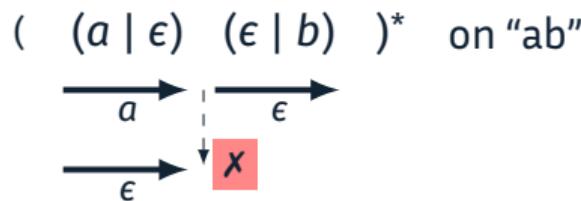
Résultat : 1 itération, matchant "a".



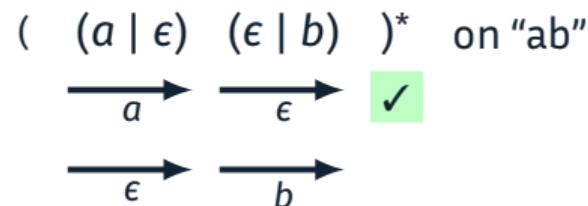
Deux manières d'éviter les boucles infinies

La majorité des langages : boucles d' ϵ interdites.

JavaScript : Les itérations ne peuvent pas matcher la chaîne vide..



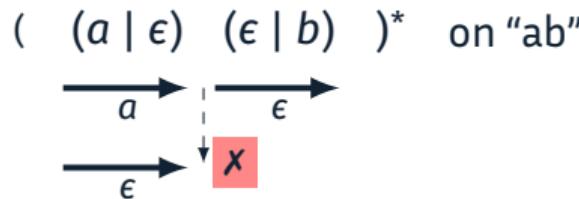
Résultat : 1 itération, matchant "a".



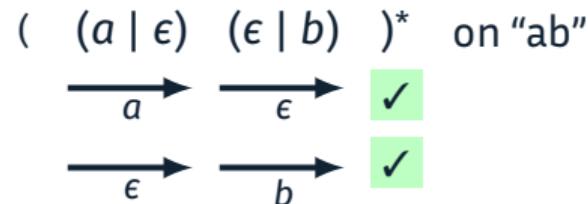
Deux manières d'éviter les boucles infinies

La majorité des langages : boucles d' ϵ interdites.

JavaScript : Les itérations ne peuvent pas matcher la chaîne vide..



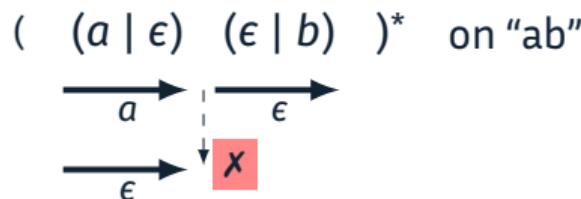
Résultat : 1 itération, matchant "a".



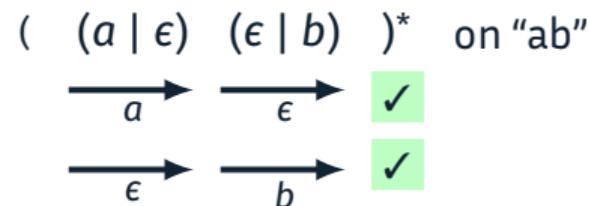
Deux manières d'éviter les boucles infinies

La majorité des langages : boucles d' ϵ interdites.

JavaScript : Les itérations ne peuvent pas matcher la chaîne vide..



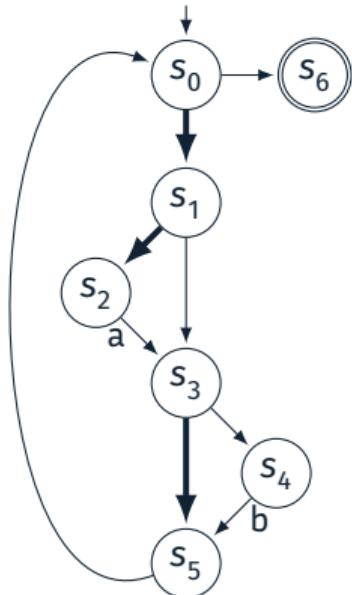
Résultat : 1 itération, matchant "a".



Résultat : 2 itérations, matchant "ab".

Simulation de NFA sur "ab"

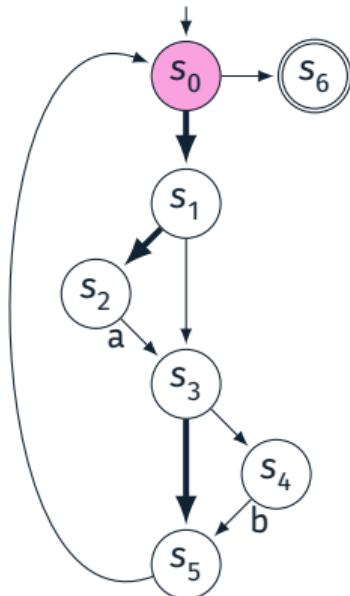
- Construire le NFA.
- Parcours en largeur pour trouver le chemin de plus haute priorité.
- Ne jamais visiter deux fois la même configuration (état du NFA + position de la chaîne).



NFA de $((a \mid \epsilon) (\epsilon \mid b))^*$
avec arêtes de priorité en **gras**.

Simulation de NFA sur "ab"

- Construire le NFA.
- Parcours en largeur pour trouver le chemin de plus haute priorité.
- Ne jamais visiter deux fois la même configuration (état du NFA + position de la chaîne).

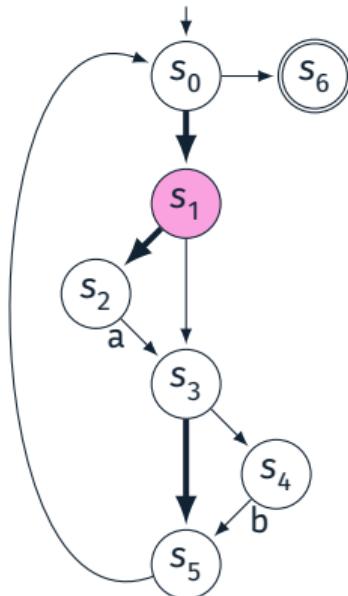


On ne trouve pas le chemin de plus haute priorité pour la sémantique JavaScript!

NFA de $((a \mid \epsilon) (\epsilon \mid b))^*$
avec arêtes de priorité en **gras**.

Simulation de NFA sur "ab"

- Construire le NFA.
- Parcours en largeur pour trouver le chemin de plus haute priorité.
- Ne jamais visiter deux fois la même configuration (état du NFA + position de la chaîne).

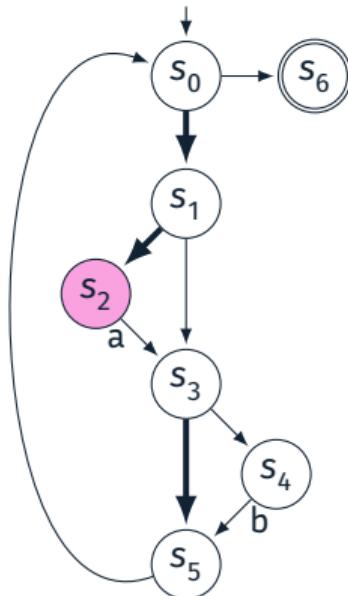


On ne trouve pas le chemin de plus haute priorité pour la sémantique JavaScript!

NFA de $((a \mid \epsilon) (\epsilon \mid b))^*$
avec arêtes de priorité en **gras**.

Simulation de NFA sur "ab"

- Construire le NFA.
- Parcours en largeur pour trouver le chemin de plus haute priorité.
- Ne jamais visiter deux fois la même configuration (état du NFA + position de la chaîne).

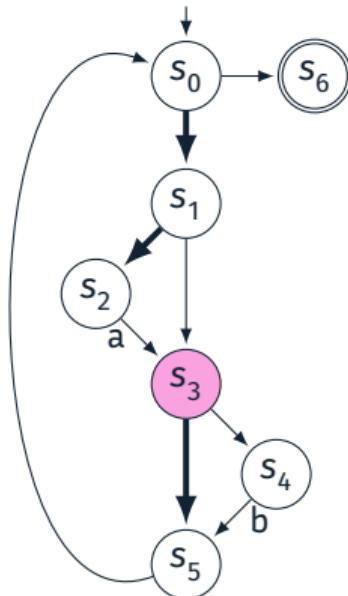


On ne trouve pas le chemin de plus haute priorité pour la sémantique JavaScript!

NFA de $((a \mid \epsilon) (\epsilon \mid b))^*$
avec arêtes de priorité en **gras**.

Simulation de NFA sur "ab"

- Construire le NFA.
- Parcours en largeur pour trouver le chemin de plus haute priorité.
- Ne jamais visiter deux fois la même configuration (état du NFA + position de la chaîne).

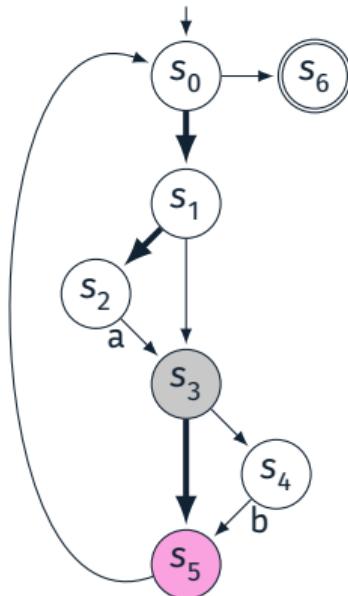


On ne trouve pas le chemin de plus haute priorité pour la sémantique JavaScript!

NFA de $((a \mid \epsilon) (\epsilon \mid b))^*$
avec arêtes de priorité en **gras**.

Simulation de NFA sur "ab"

- Construire le NFA.
- Parcours en largeur pour trouver le chemin de plus haute priorité.
- Ne jamais visiter deux fois la même configuration (état du NFA + position de la chaîne).

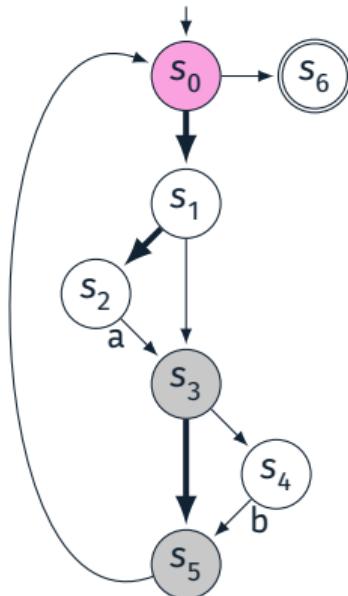


On ne trouve pas le chemin de plus haute priorité pour la sémantique JavaScript!

NFA de $((a \mid \epsilon) (\epsilon \mid b))^*$
avec arêtes de priorité en **gras**.

Simulation de NFA sur "ab"

- Construire le NFA.
- Parcours en largeur pour trouver le chemin de plus haute priorité.
- Ne jamais visiter deux fois la même configuration (état du NFA + position de la chaîne).

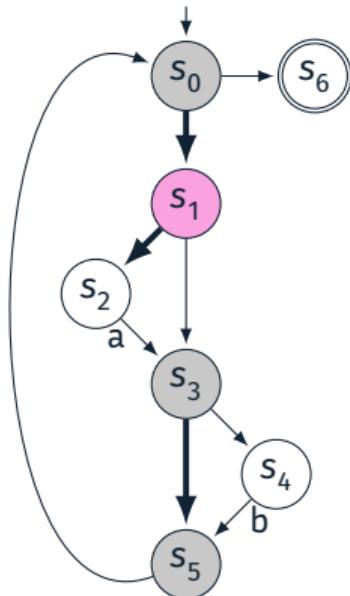


On ne trouve pas le chemin de plus haute priorité pour la sémantique JavaScript!

NFA de $((a \mid \epsilon) (\epsilon \mid b))^*$
avec arêtes de priorité en **gras**.

Simulation de NFA sur "ab"

- Construire le NFA.
- Parcours en largeur pour trouver le chemin de plus haute priorité.
- Ne jamais visiter deux fois la même configuration (état du NFA + position de la chaîne).

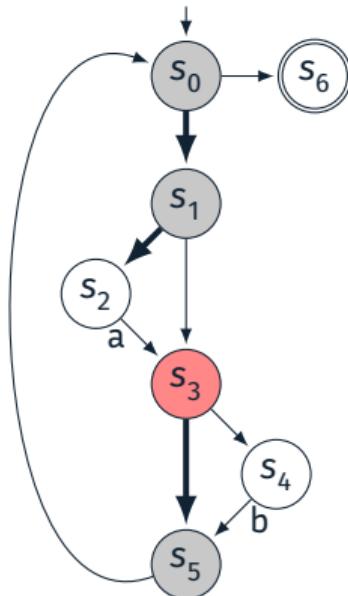


On ne trouve pas le chemin de plus haute priorité pour la sémantique JavaScript!

NFA de $((a \mid \epsilon) (\epsilon \mid b))^*$
avec arêtes de priorité en **gras**.

Simulation de NFA sur "ab"

- Construire le NFA.
- Parcours en largeur pour trouver le chemin de plus haute priorité.
- Ne jamais visiter deux fois la même configuration (état du NFA + position de la chaîne).

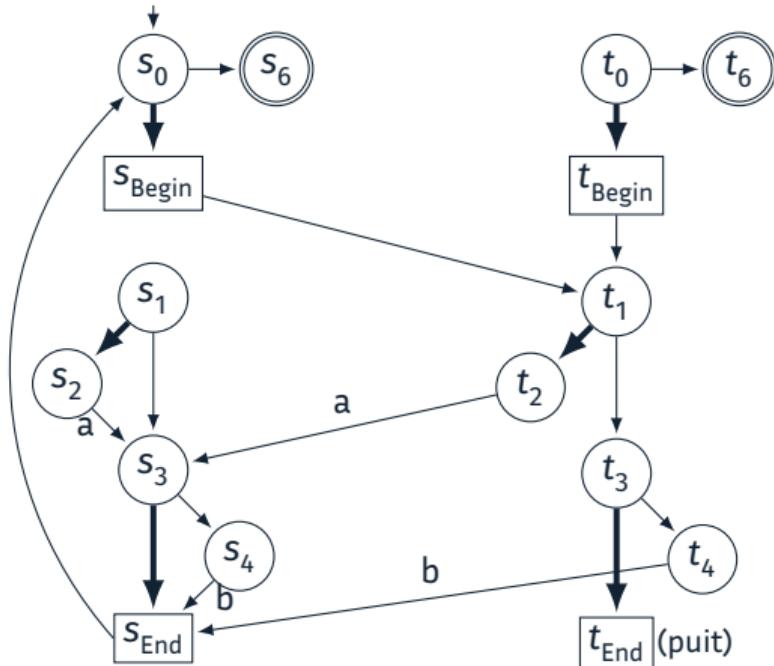


On ne trouve pas le chemin de plus haute priorité pour la sémantique JavaScript!

NFA de $((a \mid \epsilon) (\epsilon \mid b))^*$
avec arêtes de priorité en **gras**.

Nouvelle construction de NFA

- 2 copies du NFA.
- Nouveaux nœuds Begin et End.
- Begin pointe à droite.
- Lire un caractère pointe à gauche.

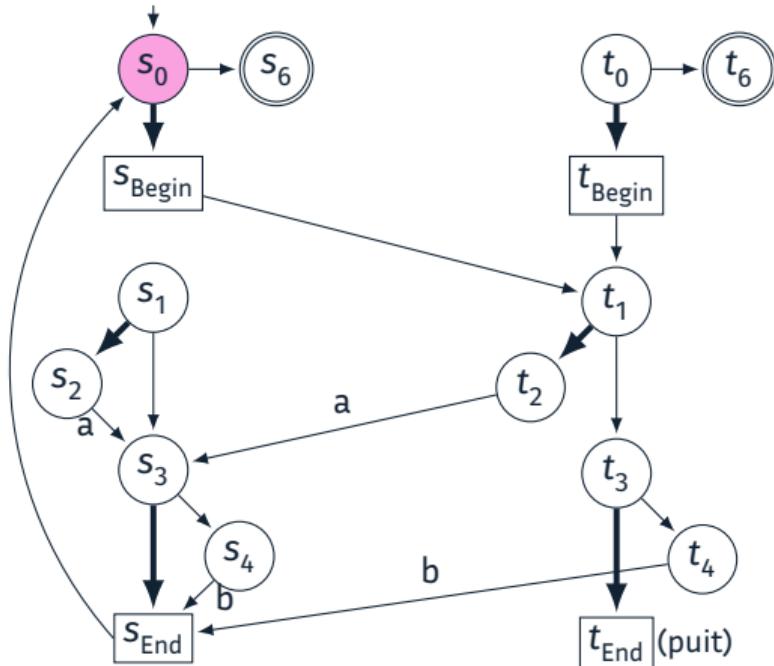


On peut sortir de l'étoile.

On ne peut pas sortir de l'étoile.

Nouvelle construction de NFA

- 2 copies du NFA.
- Nouveaux nœuds Begin et End.
- Begin pointe à droite.
- Lire un caractère pointe à gauche.

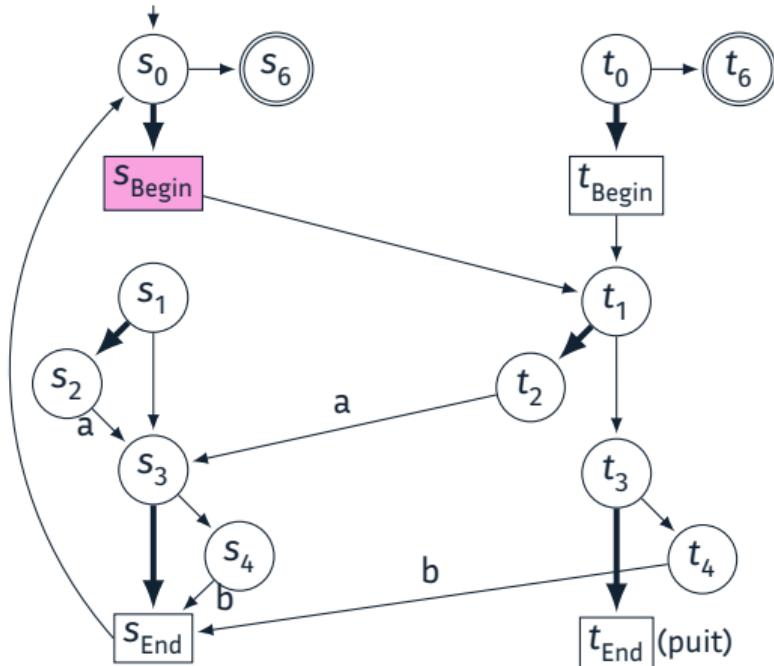


On peut sortir de l'étoile.

On ne peut pas sortir de l'étoile.

Nouvelle construction de NFA

- 2 copies du NFA.
- Nouveaux nœuds Begin et End.
- Begin pointe à droite.
- Lire un caractère pointe à gauche.

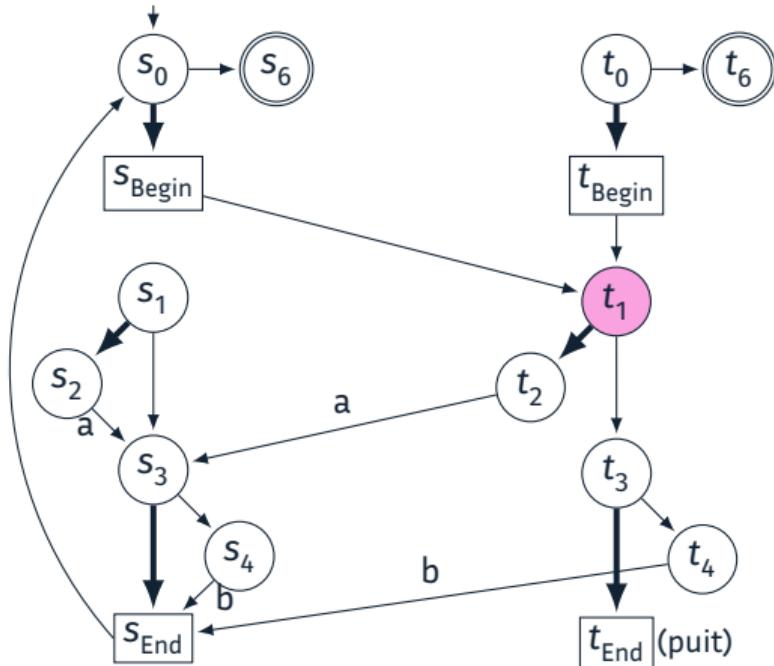


On peut sortir de l'étoile.

On ne peut pas sortir de l'étoile.

Nouvelle construction de NFA

- 2 copies du NFA.
- Nouveaux nœuds Begin et End.
- Begin pointe à droite.
- Lire un caractère pointe à gauche.

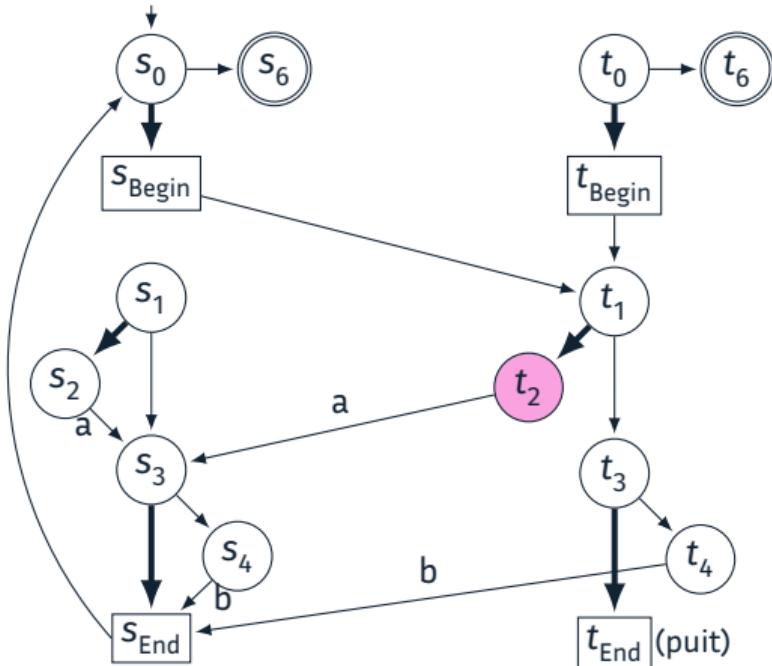


On peut sortir de l'étoile.

On ne peut pas sortir de l'étoile.

Nouvelle construction de NFA

- 2 copies du NFA.
- Nouveaux nœuds Begin et End.
- Begin pointe à droite.
- Lire un caractère pointe à gauche.

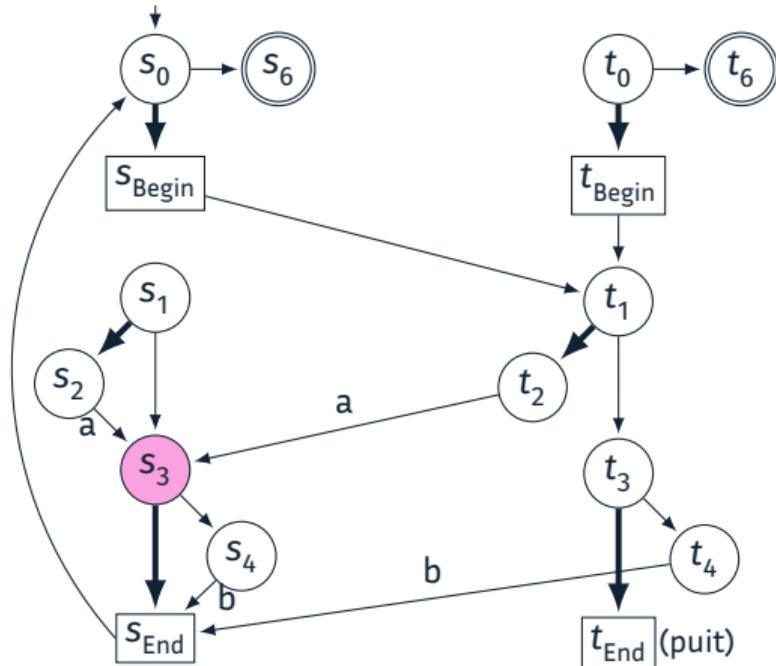


On peut sortir de l'étoile.

On ne peut pas sortir de l'étoile.

Nouvelle construction de NFA

- 2 copies du NFA.
- Nouveaux nœuds Begin et End.
- Begin pointe à droite.
- Lire un caractère pointe à gauche.

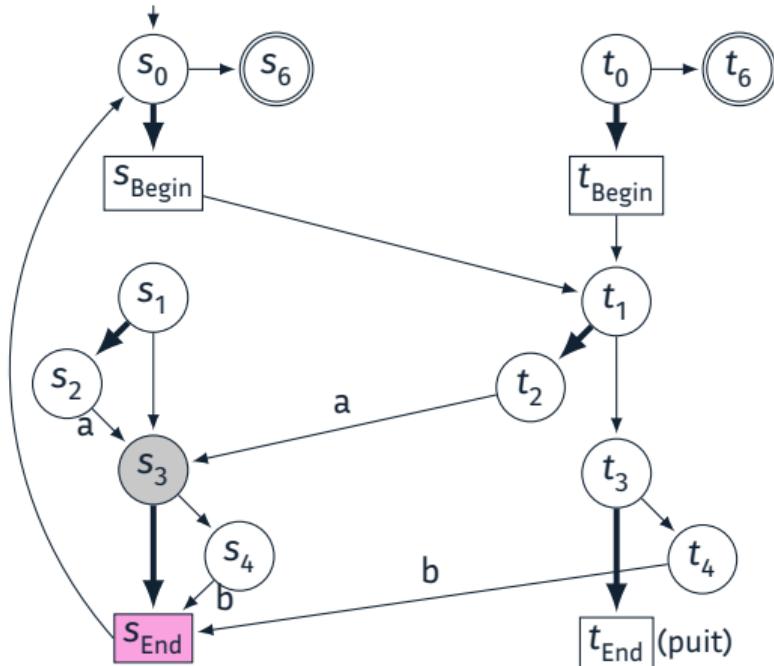


On peut sortir de l'étoile.

On ne peut pas sortir de l'étoile.

Nouvelle construction de NFA

- 2 copies du NFA.
- Nouveaux nœuds Begin et End.
- Begin pointe à droite.
- Lire un caractère pointe à gauche.

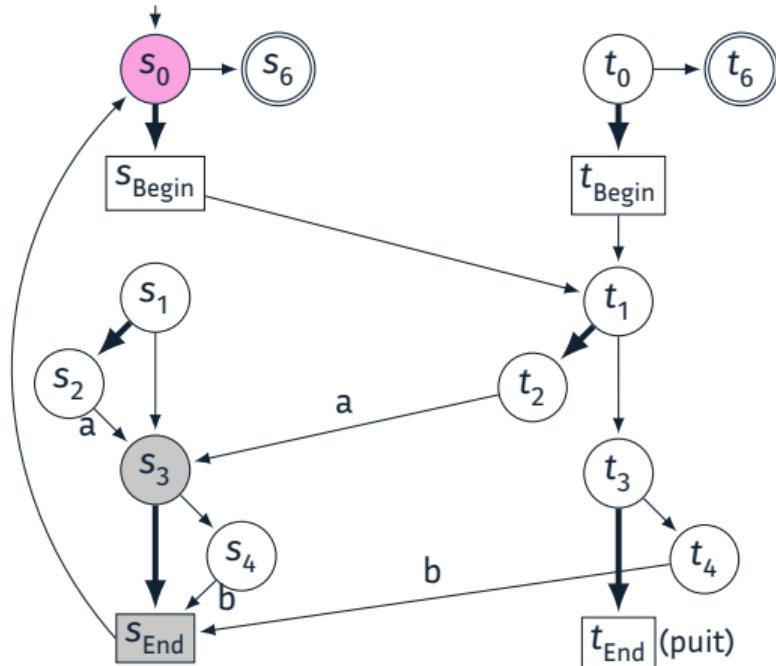


On peut sortir de l'étoile.

On ne peut pas sortir de l'étoile.

Nouvelle construction de NFA

- 2 copies du NFA.
- Nouveaux nœuds Begin et End.
- Begin pointe à droite.
- Lire un caractère pointe à gauche.

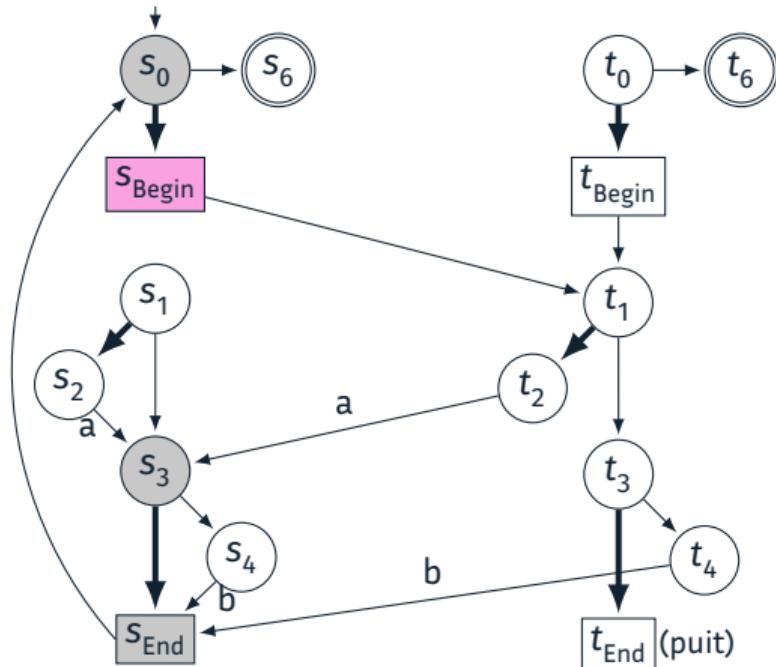


On peut sortir de l'étoile.

On ne peut pas sortir de l'étoile.

Nouvelle construction de NFA

- 2 copies du NFA.
- Nouveaux nœuds Begin et End.
- Begin pointe à droite.
- Lire un caractère pointe à gauche.

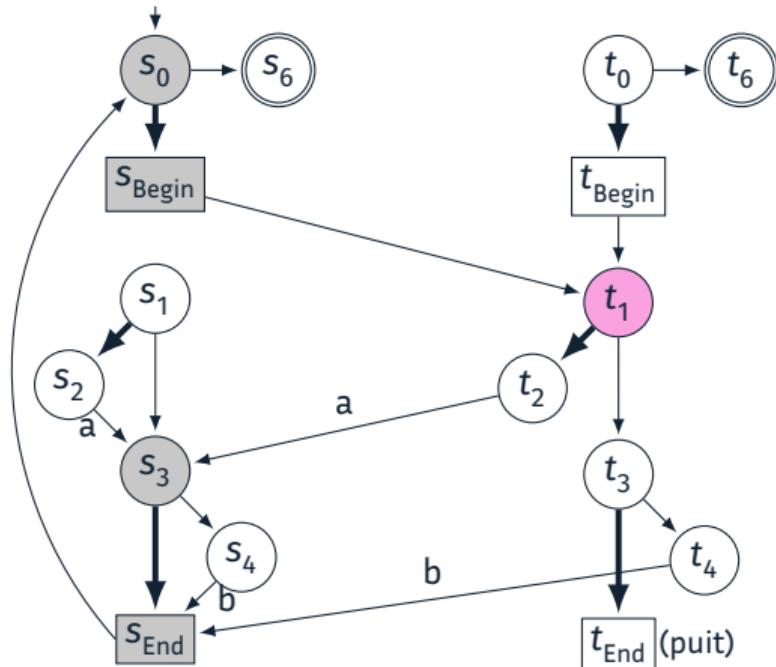


On peut sortir de l'étoile.

On ne peut pas sortir de l'étoile.

Nouvelle construction de NFA

- 2 copies du NFA.
- Nouveaux nœuds Begin et End.
- Begin pointe à droite.
- Lire un caractère pointe à gauche.

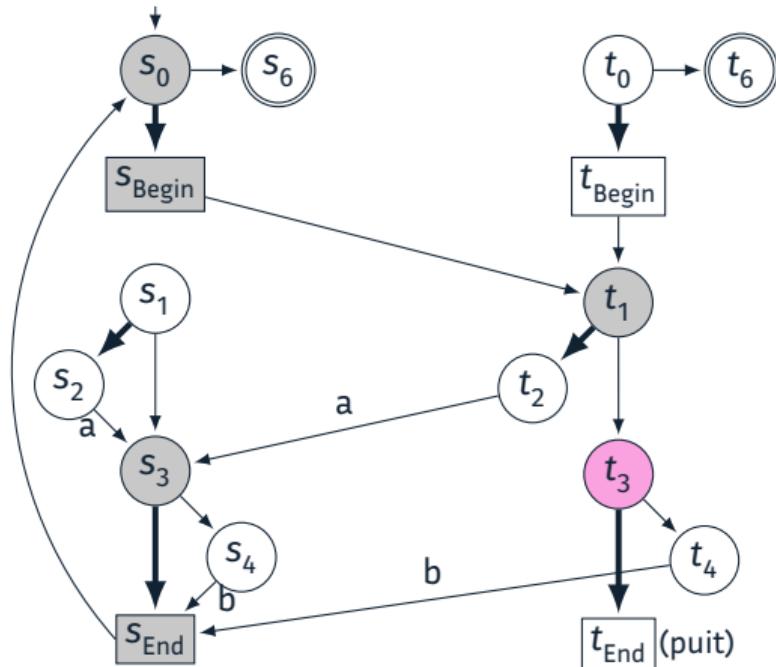


On peut sortir de l'étoile.

On ne peut pas sortir de l'étoile.

Nouvelle construction de NFA

- 2 copies du NFA.
- Nouveaux nœuds Begin et End.
- Begin pointe à droite.
- Lire un caractère pointe à gauche.

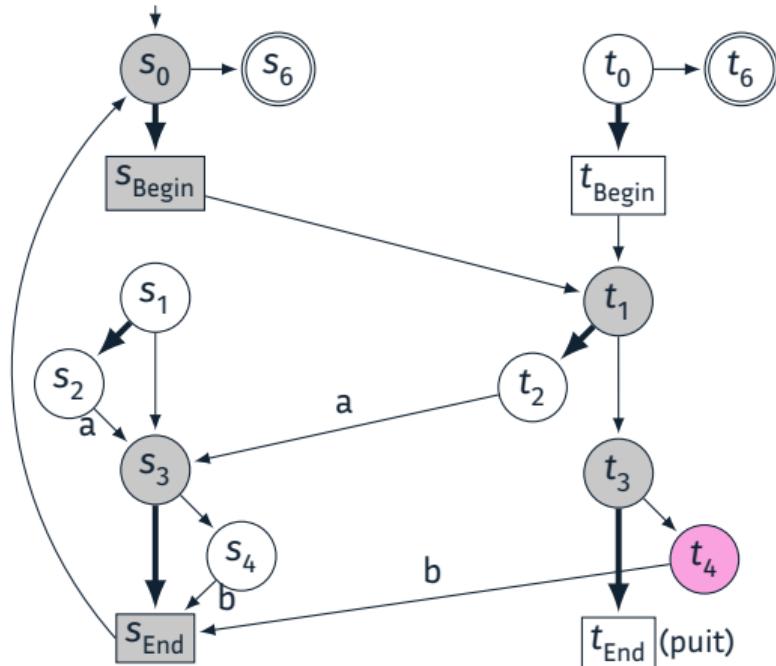


On peut sortir de l'étoile.

On ne peut pas sortir de l'étoile.

Nouvelle construction de NFA

- 2 copies du NFA.
- Nouveaux nœuds Begin et End.
- Begin pointe à droite.
- Lire un caractère pointe à gauche.

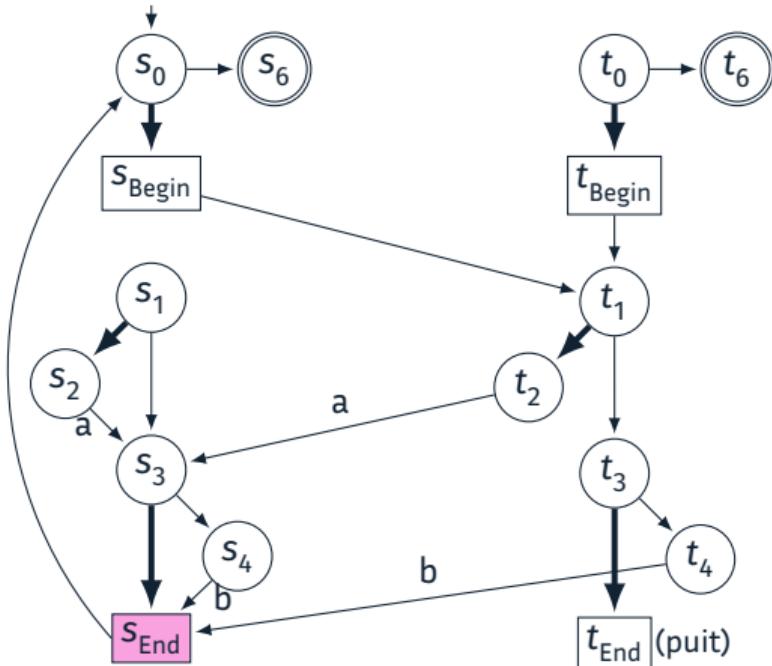


On peut sortir de l'étoile.

On ne peut pas sortir de l'étoile.

Nouvelle construction de NFA

- 2 copies du NFA.
- Nouveaux nœuds Begin et End.
- Begin pointe à droite.
- Lire un caractère pointe à gauche.

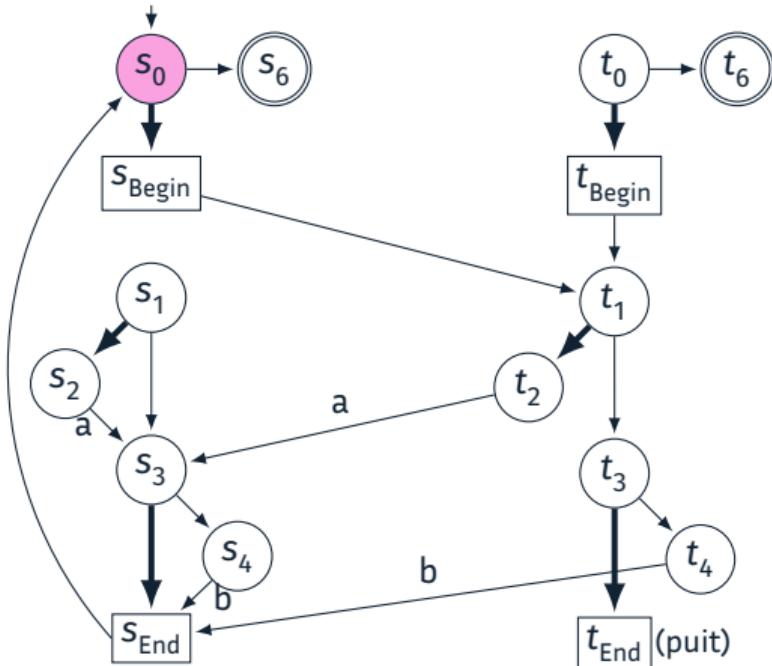


On peut sortir de l'étoile.

On ne peut pas sortir de l'étoile.

Nouvelle construction de NFA

- 2 copies du NFA.
- Nouveaux nœuds Begin et End.
- Begin pointe à droite.
- Lire un caractère pointe à gauche.

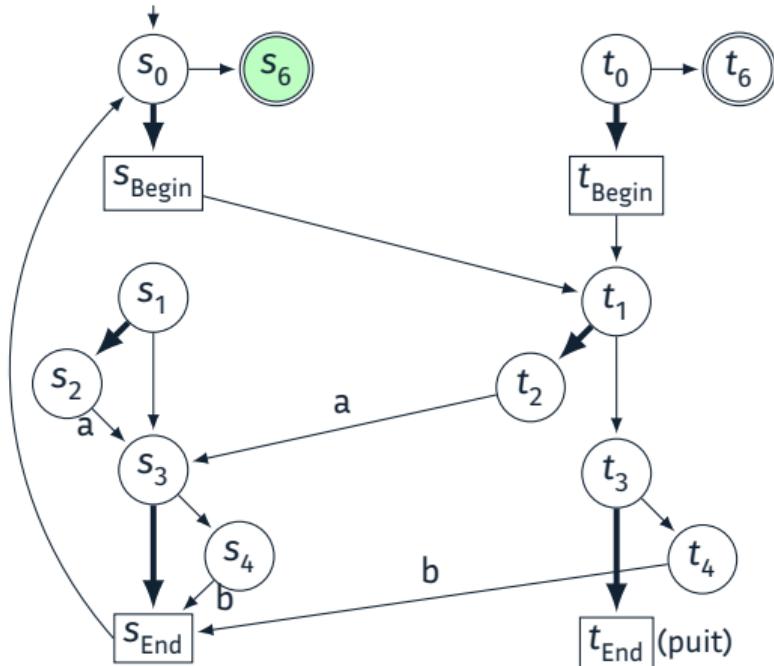


On peut sortir de l'étoile.

On ne peut pas sortir de l'étoile.

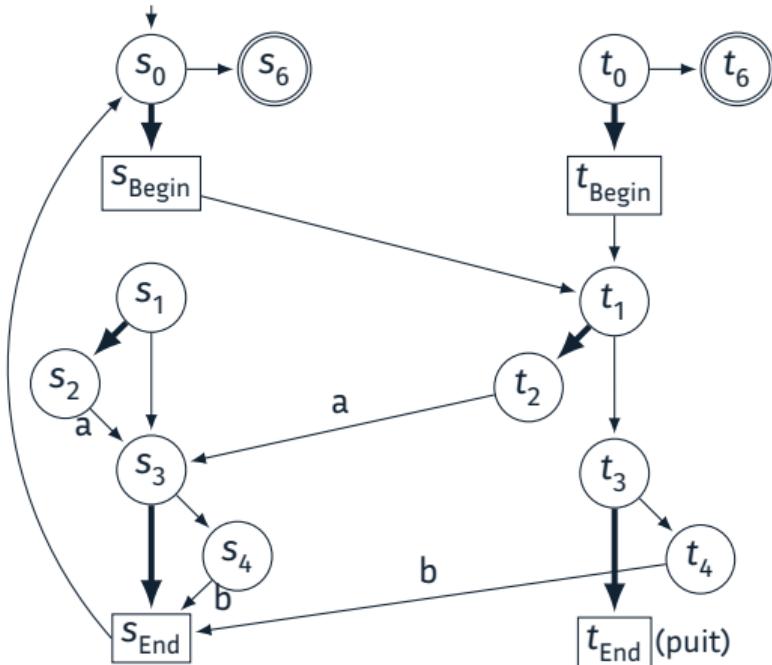
Nouvelle construction de NFA

- 2 copies du NFA.
- Nouveaux nœuds Begin et End.
- Begin pointe à droite.
- Lire un caractère pointe à gauche.



On peut sortir de l'étoile.

On ne peut pas sortir de l'étoile.



Nouvelle construction de NFA

- 2 copies du NFA.
- Nouveaux nœuds Begin et End.
- Begin pointe à droite.
- Lire un caractère pointe à gauche.

Algorithme linéaire pour l'étoile JavaScript

- Suit la sémantique JavaScript.
- Linéaire (deux copies seulement, même pour les étoiles imbriquées).
- Implémenté dans V8.

On peut sortir de l'étoile.

On ne peut pas sortir de l'étoile.

Groupes de capture

Retourner la sous-chaîne matchée en dernier par la sous-regex entre parenthèses.

Matcher $a(b \mid c)d$ sur "acde" = ("acd", "c").

Groupes de capture

Retourner la sous-chaîne matchée en dernier par la sous-regex entre parenthèses.

Matcher $a(b \mid c)d$ sur "acde" = ("acd", "c").

En JavaScript seulement : réinitialisation des captures

À chaque itération de l'étoile, réinitialiser les valeurs des groupes dans l'étoile.

Matcher $((a \mid b)^*$ sur "ab" = ("ab", "b", **undefined**).

Groupes de capture

Retourner la sous-chaîne matchée en dernier par la sous-regex entre parenthèses.

Matcher $a(b \mid c)d$ sur "acde" = ("acd", "c").

En JavaScript seulement : réinitialisation des captures

À chaque itération de l'étoile, réinitialiser les valeurs des groupes dans l'étoile.

Matcher $((a \mid b)^*$ sur "ab" = ("ab", "b", **undefined**).

Lookarounds

Une condition dans une regex. $a(?=b)$ matche les "a", seulement si ils sont suivis d'un "b".

$[0-9]^+(?=^\circ C)$ matche "12" dans "12°C", mais rien dans "12 mars".

Idée clé 1 - Précalcul

En temps linéaire, on peut précalculer chaque position où un lookaround est vrai.

En *inversant* la regex.

Idée clé 1 - Précalcul

En temps linéaire, on peut précalculer chaque position où un lookaround est vrai.

En *inversant* la regex.

Un algorithme en 3 étapes

- Précalculer une table.

Idée clé 1 - Précalcul

En temps linéaire, on peut précalculer chaque position où un lookahead est vrai.

En *inversant* la regex.

Un algorithme en 3 étapes

- Précalculer une table.

Exemple : $(a \ (?=(a | b)))^*$ sur "aaac".

	a	a	a	c
(a b)	✓	✓	✓	✗
				✗

Idée clé 1 - Précalcul

En temps linéaire, on peut précalculer chaque position où un lookahead est vrai.

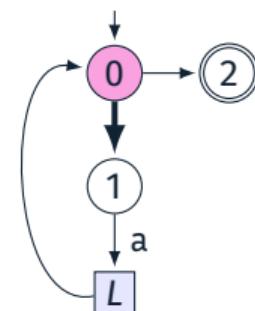
En *inversant* la regex.

Un algorithme en 3 étapes

- Précalculer une table.
- Matcher l'expression principale.

Exemple : $(a \ (?=(a | b)))^*$ sur "aaac".

	a	a	a	c	
(a b)	✓	✓	✓	x	x



Idée clé 1 - Précalcul

En temps linéaire, on peut précalculer chaque position où un lookahead est vrai.

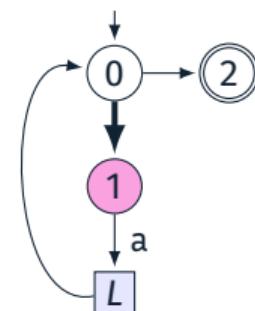
En *inversant* la regex.

Un algorithme en 3 étapes

- Précalculer une table.
- Matcher l'expression principale.

Exemple : $(a \ (?=(a | b)))^*$ sur "aaac".

	a	a	a	c	
(a b)	✓	✓	✓	x	x



Idée clé 1 - Précalcul

En temps linéaire, on peut précalculer chaque position où un lookahead est vrai.

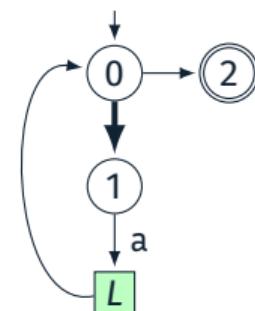
En *inversant* la regex.

Un algorithme en 3 étapes

- Précalculer une table.
- Matcher l'expression principale.

Exemple : $(a \ (?=(a | b)))^*$ sur "aaac".

	a	a	a	c
(a b)	✓	✓	✓	x



Idée clé 1 - Précalcul

En temps linéaire, on peut précalculer chaque position où un lookahead est vrai.

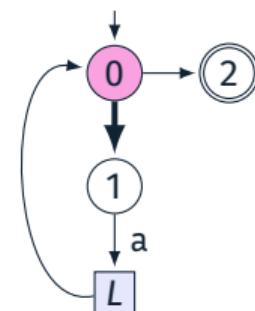
En *inversant* la regex.

Un algorithme en 3 étapes

- Précalculer une table.
- Matcher l'expression principale.

Exemple : $(a \ (?=(a | b)))^*$ sur "aaac".

	a	a	a	c	
(a b)	✓	✓	✓	x	x



Idée clé 1 - Précalcul

En temps linéaire, on peut précalculer chaque position où un lookahead est vrai.

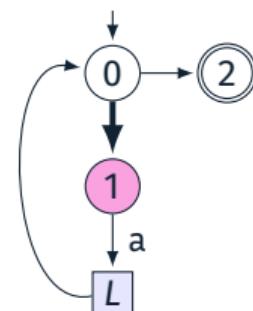
En *inversant* la regex.

Un algorithme en 3 étapes

- Précalculer une table.
- Matcher l'expression principale.

Exemple : $(a \ (?=(a | b)))^*$ sur "aaac".

	a	a	a	c	
(a b)	✓	✓	✓	x	x



Idée clé 1 - Précalcul

En temps linéaire, on peut précalculer chaque position où un lookahead est vrai.

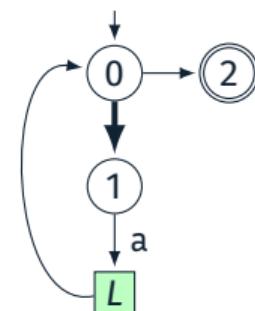
En *inversant* la regex.

Un algorithme en 3 étapes

- Précalculer une table.
- Matcher l'expression principale.

Exemple : $(a \ (?=(a | b)))^*$ sur "aaac".

	a	a	a	c
(a b)	✓	✓	✓	x



Idée clé 1 - Précalcul

En temps linéaire, on peut précalculer chaque position où un lookahead est vrai.

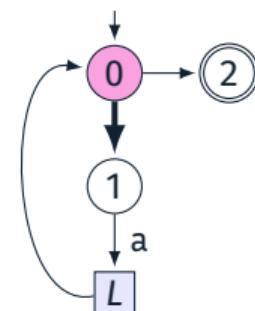
En *inversant* la regex.

Un algorithme en 3 étapes

- Précalculer une table.
- Matcher l'expression principale.

Exemple : $(a \ (?=(a | b)))^*$ sur "aaac".

	a	a	a	c	
(a b)	✓	✓	✓	x	x



Idée clé 1 - Précalcul

En temps linéaire, on peut précalculer chaque position où un lookahead est vrai.

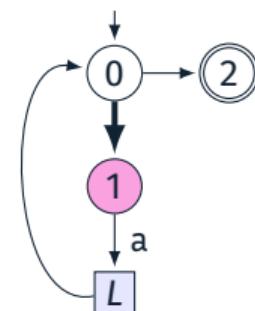
En *inversant* la regex.

Un algorithme en 3 étapes

- Précalculer une table.
- Matcher l'expression principale.

Exemple : $(a \ (?=(a | b)))^*$ sur "aaac".

	a	a	a	c	
(a b)	✓	✓	✓	x	x



Idée clé 1 - Précalcul

En temps linéaire, on peut précalculer chaque position où un lookahead est vrai.

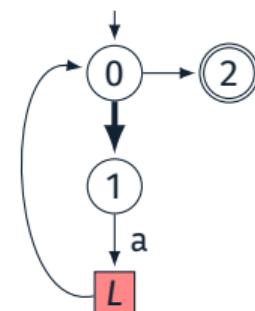
En *inversant* la regex.

Un algorithme en 3 étapes

- Précalculer une table.
- Matcher l'expression principale.

Exemple : $(a \ (?=(a | b)))^*$ sur "aaac".

	a	a	a	c
(a b)	✓	✓	✓	X



Idée clé 1 - Précalcul

En temps linéaire, on peut précalculer chaque position où un lookahead est vrai.

En *inversant* la regex.

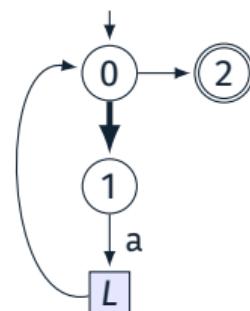
Un algorithme en 3 étapes

- Précalculer une table.
- Matcher l'expression principale.

Et si les lookarounds ont des groupes?

Exemple : $(a \ (?=(a | b)))^*$ sur "aaac".

	a	a	a	c	
(a b)	✓	✓	✓	x	x



Idée clé 1 - Précalcul

En temps linéaire, on peut précalculer chaque position où un lookahead est vrai.

En *inversant* la regex.

Un algorithme en 3 étapes

- Précalculer une table.
- Matcher l'expression principale.

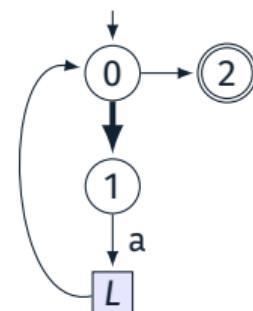
Et si les lookaheads ont des groupes?

Idée clé 2 - Conséquence de la réinitialisation

Chaque groupe dans un lookahead ne peut être défini que par le dernier usage du lookahead.

Exemple : $(a \ (?=(a | b)))^*$ sur "aaac".

	a	a	a	c	
(a b)	✓	✓	✓	x	x



Idée clé 1 - Précalcul

En temps linéaire, on peut précalculer chaque position où un lookahead est vrai.

En *inversant* la regex.

Un algorithme en 3 étapes

- Précalculer une table.
- Matcher l'expression principale.

Et si les lookaheads ont des groupes?

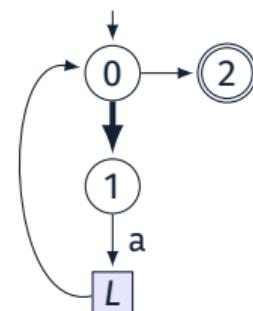
- Reconstruire les groupes des lookahead :
Matcher chaque lookahead une fois à partir de leur dernier usage.

Idée clé 2 - Conséquence de la réinitialisation

Chaque groupe dans un lookahead ne peut être défini que par le dernier usage du lookahead.

Exemple : $(a \ (?=(a | b)))^*$ sur "aaac".

	a	a	a	x	c
(a b)	✓	✓	✓	x	x



Idée clé 1 - Précalcul

En temps linéaire, on peut précalculer chaque position où un lookahead est vrai.

En *inversant* la regex.

Un algorithme en 3 étapes

- Précalculer une table.
- Matcher l'expression principale.

Et si les lookarounds ont des groupes?

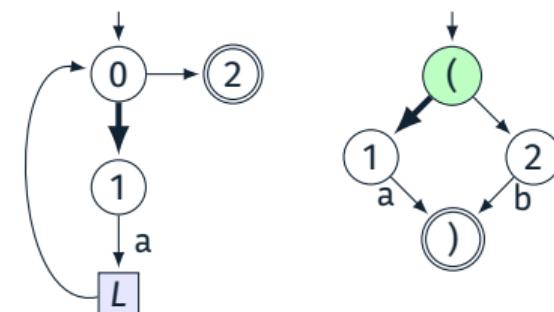
- Reconstruire les groupes des lookahead :
Matcher chaque lookahead une fois à partir de leur dernier usage.

Idée clé 2 - Conséquence de la réinitialisation

Chaque groupe dans un lookahead ne peut être défini que par le dernier usage du lookahead.

Exemple : $(a \ (?=(a | b)))^*$ sur "aaac".

	a	a	a	c
(a b)	✓	✓	✓	x



Comment raisonner sur les regex JavaScript?

Standard JS (pseudocode)

22.2.2 Runtime Semantics: CompilePattern

The syntax-directed operation `CompilePattern` takes argument `rer` (a `RegExp Record`) and returns an `Abstract Closure` that takes a `List` of characters and a non-negative `integer` and returns either a `MatchState` or `FAILURE`. It is defined piecewise over the following productions:

Pattern :: *Disjunction*

1. Let `m` be `CompileSubpattern` of `Disjunction` with arguments `rer` and `FORWARD`.
2. Return a new `Abstract Closure` with parameters `(Input, index)` that captures `rer` and `m` and performs the following steps when called:
 - a. **Assert:** `Input` is a `List` of characters.
 - b. **Assert:** $0 \leq index \leq$ the number of elements in `Input`.
 - c. Let `c` be a new `MatcherContinuation` with parameters `(y)` that captures nothing and performs the following steps when called:
 - i. **Assert:** `y` is a `MatchState`.
 - ii. Return `y`.
 - d. Let `cap` be a `List` of `rer[[CapturingGroupsCount]] undefined` values, indexed 1 through `rer[[CapturingGroupsCount]]`.
 - e. Let `x` be the `MatchState` { `[[Input]]: Input, [[EndIndex]]: index, [[Captures]]: cap` }.
 - f. Return `m(x, c)`.

Comment raisonner sur les regex JavaScript?

Standard JS (pseudocode)

2.2.2.2 Runtime Semantics: CompilePattern

The syntax-directed operation `CompilePattern` takes argument `rer` (a `RegExp Record`) and returns an `Abstract Closure` that takes a `List` of characters and a non-negative `integer` and returns either a `MatchState` or `FAILURE`. It is defined piecewise over the following productions:

Pattern :: *Disjunction*

1. Let `m` be `CompileSubpattern` of `Disjunction` with arguments `rer` and `FORWARD`.
2. Return a new `Abstract Closure` with parameters (`Input`, `index`) that captures `rer` and `m` and performs the following steps when called:
 - a. **Assert:** `Input` is a `List` of characters.
 - b. **Assert:** $0 \leq index \leq$ the number of elements in `Input`.
 - c. Let `c` be a new `MatcherContinuation` with parameters (`y`) that captures nothing and performs the following steps when called:
 - i. **Assert:** `y` is a `MatchState`.
 - ii. Return `y`.
 - d. Let `cap` be a `List` of `rer`.`[[CapturingGroupsCount]]` `undefined` values, indexed 1 through `rer`.`[[CapturingGroupsCount]]`.
 - e. Let `x` be the `MatchState` { `[[Input]]: Input`, `[[EndIndex]]: index`, `[[Captures]]: cap` }.
 - f. Return `m(x, c)`.

Modèles existants

Operation	<i>t</i>	Overapproximate Model for $(w, C_1, \dots, C_{t-1}) \in \mathcal{L}_x(t)$
Alternation	$t_1 t_2$	$\{(w, C_1, \dots, C_{t-1}) \in \mathcal{L}_x(t_1) \wedge C_{t-1} = \dots = C_{t-1} = \emptyset\}$ $\vee \{(w, C_1, \dots, C_{t-1}) \in \mathcal{L}_x(t_2) \wedge C_{t-1} = \dots = C_{t-1} = \emptyset\}$
Concatenation	$t_1 \cdot t_2$	$w = w_1 \leftrightarrow w_2 \wedge (w_1, C_1, \dots, C_{t-1}) \in \mathcal{L}_x(t_1) \wedge (w_2, C_{t-1}, \dots, C_{t-1}) \in \mathcal{L}_x(t_2)$
Backreference-free	t_1^*	$w = w_1 \leftrightarrow w_2 \wedge w_1 \in \mathcal{L}(x) \wedge (w_2, C_1, \dots, C_{t-1}) \in \mathcal{L}_x(t_1)x$ $\wedge \{w_2 = x \wedge C_1 = \dots = C_{t-1} = \emptyset\}$
Quantification		
Positive Lookahead	$(?>t_1)t_2$	$(w, C_1, \dots, C_{t-1}) \in \mathcal{L}_x(t_1) \cdot x \wedge (w, C_{t-1}, \dots, C_{t-1}) \in \mathcal{L}_x(t_2)$
Negative Lookahead	$(?!)t_1)t_2$	$(w, C_1, \dots, C_{t-1}) \notin \mathcal{L}_x(t_1) \cdot x \wedge (w, C_{t-1}, \dots, C_{t-1}) \notin \mathcal{L}_x(t_2)$
Input Start	t_1^+	$(w, C_1, \dots, C_{t-1}) \in \mathcal{L}_x(t_1) \wedge (w, C_1, \dots, C_{t-1}) \in \mathcal{L}_x(t_1) \cdot x$
Input Start (Midline)	t_1^*	$(w, C_1, \dots, C_{t-1}) \in \mathcal{L}_x(t_1) \wedge (w, C_1, \dots, C_{t-1}) \in \mathcal{L}_x(t_1) \cdot (t_1^*)x$
Input End	$t_1^!$	$(w, C_1, \dots, C_{t-1}) \in \mathcal{L}_x(t_1) \wedge (w, C_1, \dots, C_{t-1}) \in \mathcal{L}_x(t_1) \cdot x$
Input End (Midline)	$t_1^!$	$(w, C_1, \dots, C_{t-1}) \in \mathcal{L}_x(t_1) \cdot x \wedge (w, C_1, \dots, C_{t-1}) \in \mathcal{L}_x(t_1) \cdot (t_1^!)x$
Word Boundary	$t_1 \backslash \{b\} t_2$	$w = w_1 \leftrightarrow w_2 \wedge (w_1, C_1, \dots, C_{t-1}) \in \mathcal{L}_x(t_1) \wedge (w_2, C_{t-1}, \dots, C_{t-1}) \in \mathcal{L}_x(t_2)$ $\wedge \{\{w_1 \in \mathcal{L}(x), x \mid 0 \vee w_1 = x\} \wedge w_2 \notin \mathcal{L}(x) \cdot x\} \vee \{w_1 \notin \mathcal{L}(x) \cdot x\} \vee \{w_1 \in \mathcal{L}(x) \cdot x\} \wedge (w_2 \in \mathcal{L}(x) \cdot x) \vee w_2 = x\}$
Non-Word Boundary	$t_1 \backslash \{B\} t_2$	$w = w_1 \leftrightarrow w_2 \wedge (w_1, C_1, \dots, C_{t-1}) \in \mathcal{L}_x(t_1) \wedge (w_2, C_{t-1}, \dots, C_{t-1}) \in \mathcal{L}_x(t_2)$ $\wedge \{\{w_1 \notin \mathcal{L}(x) \cdot x\} \wedge w_2 = x \vee w_2 \notin \mathcal{L}(x) \cdot x\} \vee \{w_1 \notin \mathcal{L}(x) \cdot x\} \wedge \{w_2 \notin \mathcal{L}(x) \cdot x\} \vee (w_1 \notin \mathcal{L}(x) \cdot x) \vee (w_2 \notin \mathcal{L}(x) \cdot x) \wedge w_2 \neq x\}$
Capture Group	(t_1)	$(w, C_1, \dots, C_{t-1}) \in \mathcal{L}_x(t_1) \wedge C_t = w$
Non-Capturing Group	$(?:t_1)$	$(w, C_1, \dots, C_{t-1}) \in \mathcal{L}_x(t_1)$
Base Case	t regular	$w \in \mathcal{L}(t)$

Comment raisonner sur les regex JavaScript?

Standard JS (pseudocode)

22.2.2.2 Runtime Semantics: CompilePattern

The syntax-directed operation `CompilePattern` takes argument `rer` (a `RegExp Record`) and returns an `Abstract Closure` that takes a `List` of characters and a non-negative `integer` and returns either a `MatchState` or `FAILURE`. It is defined piecewise over the following productions:

Pattern = Disjunction

- Let `m` be `CompileSubpattern` of `Disjunction` with arguments `rer` and `FORWARD`.
Return a new `Abstract Closure` with parameters `(Input, index)` that captures `rer` and `m` and performs the following steps when called:
 - `assert: Input` is a List of characters.
 - `assert: 0 ≤ index ≤` the number of elements in `Input`.
 - Let `c` be a new `MatcherContinuation` with parameters `(y)` that captures nothing and performs the following steps when called:
 - `assert: y` is a `MatchState`.
 - Return `y`.
 - Let `cap` be a List of `rer`.`[[CapturingGroupsCount]]` **undefined** values, indexed 1 through `rer`.`[[CapturingGroupsCount]]`.
 - Let `x` be the `MatchState` `[[Input]: Input, [[EndIndex]: index, [[Captures]: cap]]]`.
 - Return `m(x, r)`.

Modèles existants

Operation	t	Overapproximate Model for $(w, C_1, \dots, C_k) \in \mathcal{L}(t)$
Alternation	$t_1 \parallel t_2$	$\{ (w, C_1, \dots, C_k) \in \mathcal{L}(t_1) \mid C_1 \subseteq w \} \cup \{ (w, C_1, \dots, C_k) \in \mathcal{L}(t_2) \mid C_2 \subseteq w \}$
Concatenation	$t_1 \cdot t_2$	$\{ (w, w, C_1, \dots, C_k) \in \mathcal{L}(t_1) \mid C_1 \subseteq w \} \cup \{ (w, C_1, \dots, C_k) \in \mathcal{L}(t_2) \mid C_2 \subseteq w \}$
Backreference-free Quantification	$t_{\forall x}$	$\{ w = w_1 \leftrightarrow w_2 \mid \forall (w_1, \dots, C_k) \in \mathcal{L}(t) \mid C_1 \subseteq w_1 \wedge \dots \wedge C_k \subseteq w_1 \} \cup \{ (w, C_1, \dots, C_k) \in \mathcal{L}(t) \mid C_1 \subseteq w \}$
Positive Lookahead	$(\exists w_i) t_2$	$\{ (w, C_1, \dots, C_k) \in \mathcal{L}(t_1) \mid \forall (w, C_1, \dots, C_k) \in \mathcal{L}(t_2) \mid C_1 \subseteq w \}$
Negative Lookahead	$(\nexists t_2) t_1$	$\{ (w, C_1, \dots, C_k) \notin \mathcal{L}(t_2) \mid \forall (w, C_1, \dots, C_k) \in \mathcal{L}(t_1) \mid C_1 \subseteq w \}$
Quantifier Shunt		$\{ (w, C_1, \dots, C_k) \in \mathcal{L}(t_1) \mid \forall (w, C_1, \dots, C_k) \in \mathcal{L}(t_2) \mid C_1 \subseteq w \}$
1	$p \leq w $	$w[p] = m$
1	$(a, w, p, \Lambda) \rightarrow \emptyset$	$\rightarrow (1, 1, \Lambda)$ (CHARACTER)
1	$p \geq w $ \vee $w[p] \neq a$	$\rightarrow (w, w, p, \Lambda) \rightarrow N'$
1	$\{a, w, p, \Lambda\} \rightarrow \emptyset$	(CHARACTER FAILURE) $\rightarrow ((x_j, w, p, \Lambda) \rightarrow \{p_j, \Lambda_i\}) \mid (p_j, \Lambda_i) \in \text{CAPTURING GR}$
1	$\langle \emptyset, w, p, \Lambda \rangle \rightarrow \emptyset$	(EMPTY SET)
1	$\langle a, w, p, \Lambda \rangle \rightarrow \emptyset$	(CHARACTER FAILURE) $\rightarrow ((y, w, p, \Lambda) \rightarrow N')$
2	$(r_1, w, p, \Lambda) \rightarrow N'$	$\forall y (p, y, \Lambda) \in N' \rightarrow (r_1, w, p, y, \Lambda) \rightarrow N'$
2	$\langle r_1 r_2, w, p, \Lambda \rangle \rightarrow N'$	$\bigcup_{y \in \Sigma^N} \langle r_1 y, w, p, \Lambda \rangle \rightarrow N'$ (CONCATENATION)
2	$\langle r_1, w, p, \Lambda \rangle \rightarrow N'$	$\forall y (p, y, \Lambda) \in N' \rightarrow (r_1 y, w, p, \Lambda) \rightarrow N'$ (UNION)
2	$\forall y (p, y, \Lambda) \rightarrow N'$	$\forall y (p, y, \Lambda) \rightarrow N'$ (POSITIVE LOOKAHEAD)
2	$\forall y (p, y, \Lambda) \in \{N', \Lambda\}$	$N' = \{t \in \{w, \emptyset, \{\}\} \mid (p, t, \Lambda) \in N'\}$
2	$\langle r^*, w, p, \Lambda \rangle \rightarrow \{N', \Lambda\}$	$\bigcup_{y \in \Sigma^N} \langle r^* y, w, p, \Lambda \rangle \rightarrow N'$ (REPETITION)
2	$\forall y (p, y, \Lambda) \rightarrow N'$	$\forall y (p, y, \Lambda) \rightarrow N'$ (NEGATIVE LOOKAHEAD)

Figure 2 Rules of the matching relation \sim

Comment raisonner sur les regex JavaScript?

Problème : les modèles sémantiques existants sont incomplets ou faux.

Standard JS (pseudocode)

22.2.2.2 Runtime Semantics: CompilePattern

The syntax-directed operation `CompilePattern` takes argument `rer` (a `RegExp Record`) and returns an `Abstract Closure` that takes a `List` of characters and a non-negative `integer` and returns either a `MatchState` or `FAILURE`. It is defined piecewise over the following productions:

Pattern = Disjunction

- Let `m` be `CompileSubpattern` of `Disjunction` with arguments `rer` and `FORWARD`.
 - Return a new `Abstract Closure` with parameters `(Input, index)` that captures `rer` and `m` and performs the following steps when called:
 - Assert: `Input` is a List of characters.
 - Assert: `0 ≤ index ≤` the number of elements in `Input`.
 - Let `c` be a new `MatcherContinuation` with parameters `(y)` that captures nothing and performs the following steps when called:
 - Assert: `y` is a `MatchState`.
 - ii. Return `y`.
 - Let `cap` be a List of `rer`.`[[CapturingGroupsCount]]` `undefined` values, indexed 1 through `rer`.`[[CapturingGroupsCount]]`.
 - Let `x` be the `MatchState` `[[Input]]: Input, [[EndIndex]]: index, [[Captures]]: cap`.
 - Return `mx(x, c)`.

Non équivalent



Modèles existants

Operation	t	Overapproximate Model for $(w, C_1, \dots, C_k) \in \mathcal{L}_t(t)$
Alternation	$t_1 \parallel t_2$	$\{w, (C_1, \dots, C_k) \in \mathcal{L}(t_1) \mid C_{i_1} = \emptyset\} \cup \{w, (C_1, \dots, C_k) \in \mathcal{L}(t_2) \mid C_{i_2} = \emptyset\}$
Concatenation	$t_1 \cdot t_2$	$w = w_1 \cdots w_n \wedge (w_1, C_1, \dots, C_k) \in \mathcal{L}(t_1) \wedge \dots \wedge (w_n, C_1, \dots, C_k) \in \mathcal{L}(t_2)$
Backreference-free Quantification	$\forall p_i$	$w = w_1 \cdots w_n \wedge \forall i \in [n] \exists (w_i, C_1, \dots, C_k) \in \mathcal{L}(t) \wedge \forall i \in [n] \exists (w_i, C_1, \dots, C_k) \in \mathcal{L}(t)$
Positive Lookahead	$(C_1 \ldots C_k) t_2$	$(w, C_1, \dots, C_k) \in \mathcal{L}(t_1) \wedge (w, (C_1 \ldots C_k) t_2) \in \mathcal{L}(t_2)$
Negative Lookahead	$(\exists C_1 \ldots C_k) t_2$	$(w, C_1, \dots, C_k) \notin \mathcal{L}(t_1) \wedge (w, (C_1 \ldots C_k) t_2) \in \mathcal{L}(t_2)$
Quantifier Disjunction	$\exists p_i$	$w = w_1 \cdots w_n \wedge \exists i \in [n] \forall j \neq i \exists (w_j, C_1, \dots, C_k) \in \mathcal{L}(t) \wedge \exists p_i \in \mathbb{A}$
1	$p_i \leq w $	$\exists p_i \leq w \wedge \exists i \in [n] \forall j \neq i \exists (w_j, C_1, \dots, C_k) \in \mathcal{L}(t)$
1	$(w, w, p, \lambda) \rightarrow \emptyset$	(CHARACTER)
1	$p \geq w \vee \exists p_i \neq p$	$(w, w, p, \lambda) \rightarrow \mathcal{N}$
1	$(w, w, p, \lambda) \rightarrow \emptyset$	(CHARACTER FAILURE)
1	$(\emptyset, w, p, \lambda) \rightarrow \emptyset$	(CAPTURING GS)
1	$(\emptyset, w, p, \lambda) \rightarrow \emptyset$	(EMPTY SET)
2	$(w, w, p, \lambda) \rightarrow \{\bar{p}\}$	(EMPTY STRING)
2	$(w, w, p, \lambda) \rightarrow \{p_i\}$	$(w, w, p, \lambda) \in N_p \wedge (w, w, p, \lambda) \rightarrow \mathcal{N}$
2	$(w, w, p, \lambda) \rightarrow \bigcup_{p_i \in \mathbb{A}} \{p_i\}$	$(w, w, p, \lambda) \rightarrow \mathcal{N}$
2	$(w, w, p, \lambda) \rightarrow \bigcup_{p_i \in \mathbb{A} \setminus \{p\}}$	(CONCATENATION)
2	$(w, w, p, \lambda) \rightarrow \mathcal{N}'$	$(w_2, p, \lambda) \rightarrow \mathcal{N}' \wedge (w_1, w_2, p, \lambda) \rightarrow \mathcal{N}$
2	$(w, w, p, \lambda) \rightarrow \mathcal{N}'$	(UNION)
2	$(w, w, p, \lambda) \rightarrow \mathcal{N}'$	$(w, w, p, \lambda) \rightarrow \mathcal{N}' \wedge (w, w, p, \lambda) \rightarrow \mathcal{N}$
2	$\exists (p_i, \lambda) \in \{(\bar{p}, \lambda), (\bar{p}, \bar{\lambda})\}$	(POSITIVE LOOKAHEAD)
2	$\exists (p_i, \lambda) \in \{(\bar{p}, \lambda), (\bar{p}, \bar{\lambda})\} \cup \bigcup_{k=1}^n \text{REP}(C_k, \lambda) \cup \mathcal{N}_C$	(NEGATIVE LOOKAHEAD)

■ Figure 3 Rules of the matching game

Comment raisonner sur les regex JavaScript?

Problème : les modèles sémantiques existants sont incomplets ou faux.

Standard JS (pseudocode)

22.2.2 Runtime Semantics: CompilePattern

The syntax-directed operation `CompilePattern` takes argument `rer` (a `RegExp Record`) and returns an `Abstract Closure` that takes a `List` of characters and a non-negative `integer` and returns either a `MatchState` or `FAILURE`. It is defined piecewise over the following productions:

`Pattern :: Disjunction`

1. Let `m` be `CompileSubpattern` of `Disjunction` with arguments `rer` and `FORWARD`.
2. Return a new `Abstract Closure` with parameters `(Input, index)` that captures `rer` and `m` and performs the following steps when called:
 - a. **Assert:** `Input` is a `List` of characters.
 - b. **Assert:** $0 \leq index \leq$ the number of elements in `Input`.
 - c. Let `c` be a new `MatcherContinuation` with parameters `(y)` that captures nothing and performs the following steps when called:
 - i. **Assert:** `y` is a `MatchState`.
 - ii. Return `y`.
 - d. Let `cap` be a `List` of `rer.[[CapturingGroupsCount]] undefined` values, indexed 1 through `rer.[[CapturingGroupsCount]]`.
 - e. Let `x` be the `MatchState` { `[Input]: Input`, `[EndIndex]: index`, `[Captures]: cap` }.
 - f. Return `m(x, c)`.

Équivalent

Nouvelle Mécanisation (Coq)

```
**> Disjunction :: Alternative | Disjunction <**>
Disjunction r1 r2 =>
(*> 1. Let m1 be CompileSubpattern of Alternative with arguments rer and direction. <**)
let! m1 ==> compileSubPattern r1 (Disjunction_left r2 :: ctx) rer direction in
(*> 2. Let m2 be CompileSubpattern of Disjunction with arguments rer and direction. <**)
let! m2 ==> compileSubPattern r2 (Disjunction_right r1 :: ctx) rer direction in
(*> 3. Return a new Matcher with parameters (x, c) that captures m1 and m2 and performs
the following steps when called: <**)
(λ (x: MatchState) (c: MatcherContinuation) =>
(*> a. Assert x is a MatchState. <**)
(*> b. Assert c is a MatcherContinuation. <**)
(*> c. Let r be m1(x, c). <**)
let! r ==> m1 x c in
(*> d. If r is not failure, return r. <**)
if r is not failure then r
(*> e. Return m2(x, c). <**)
else m2 x c): Matcher
```

Une sémantique mécanisée de confiance

Thèse de master encadrée : Mécanisation du chapitre regex du standard JavaScript en Coq/Rocq.
🏅 compétition étudiante de PLDI.

22.2.2.4.1 `IsWordChar(rer, Input, e)`

The abstract operation `IsWordChar` takes arguments `rer` (a `RegExp Record`), `Input` (a `List` of characters), and `e` (an `integer`) and returns a `Boolean`.

It performs the following steps when called:

1. Let `InputLength` be the number of elements in `Input`.
2. If `e = -1` or `e = InputLength`, return `false`.
3. Let `c` be the character `Input[e]`.
4. If `WordCharacters(rer)` contains `c`, return `true`.
5. Return `false`.

```
(** >>  
 22.2.2.4.1 IsWordChar ( rer, Input, e )
```

The abstract operation IsWordChar takes arguments rer (a RegExp Record), Input (a List of characters), and e (an integer) and returns a Boolean.

It performs the following steps when called:

```
<<*>
```

(*>> 1. Let InputLength be the number of elements in Input. <<*)

(*>> 2. If $e = -1$ or $e = \text{InputLength}$, return false. <<*)

(*>> 3. Let c be the character $\text{Input}[e]$. <<*)

(*>> 4. If WordCharacters(rer) contains c , return true. <<*)

(*>> 5. Return false. <<*)

```
(** >>
 22.2.2.4.1 IsWordChar ( rer, Input, e )
```

The abstract operation IsWordChar takes arguments rer (a RegExp Record), Input (a List of characters), and e (an integer) and returns a Boolean.

It performs the following steps when called:

```
<<*>
Definition isWordChar(rer:RegExpRecord)(Input:list Character)(e:integer):Result bool :=
  (*>> 1. Let InputLength be the number of elements in Input. <<*)
  let InputLength:=List.length Input in
  (*>> 2. If e = -1 or e = InputLength, return false. <<*)
  if (e =? -1)%Z || (e =? InputLength)%Z then false
  else
    (*>> 3. Let c be the character Input[ e ]. <<*)
    let! c =<< Input[e] in
    (*>> 4. If WordCharacters(rer) contains c, return true. <<*)
    let! wc =<< wordCharacters rer in
    if CharSet.contains wc c then true
    else
      (*>> 5. Return false. <<*)
      false.
```

- 1/14 Quelles garanties pour le Web?
- 2/14 Mon domaine : compilation formellement vérifiée
- 3/14 Ma méthodologie
- 4/14 Doctorat
- 5/14 Spéculation & Déoptimisation
- 6/14 JITs vérifiés
- 7/14 PostDoc
- 8/14 Nouveaux algorithmes linéaires
- 9/14 Avancements algorithmiques et sémantiques
- 10/14 Vers une plateforme web de confiance
- 11/14 Un moteur vérifié, linéaire, efficace, intégrable
- 12/14 Un sous-ensemble WebAssembly de confiance
- 13/14 Intégration
- 14/14 Vérification formelle pour un Web de confiance

Transparents supplémentaires :
Spéculer dans un langage dynamique
Insérer des instructions spéculatives
Définition Simulations Imbriquées
Simulations imbriquées, exécution
Simulations imbriquées, optimisation
Regex modernes avec priorité
L'étoile JavaScript est unique
Simulation de NFA et étoile
Dupliquer le graphe pour l'étoile JavaScript
Lookarounds et groupes de capture
3 étapes pour les lookarounds
Sémantique formelle pour les regex JavaScript
Une mécanisation de confiance