Une étude fine des choix algorithmiques dans les langages de programmation courants

AAP CNRS 2022

Porteur : **Pablo Rotondo**, LIGM, UMR 8049

Résumé : en s'appuyant sur notre expertise locale en analyse d'algorithmes, ce projet consiste à réunir un petit consortium international pour initier un travail de fond sur l'analyse mathématique des solutions adoptées par les développeurs de langages de programmation, détecter des éventuelles lacunes et proposer des solutions efficaces en pratique.

Description générale : tous les langages de programmation contemporains proposent des implantations d'algorithmes et de structures de données classiques comme les listes, les tables de hachage, les tris, *etc.* Ce sont des briques de base qui sont utilisées pour développer des programmes plus conséquents. Des algorithmes efficaces pour traiter ce genre de questions sont connus depuis plusieurs décennies, depuis le début de l'informatique ou presque, avec souvent de nombreuses variantes efficaces proposées dans la littérature.

Quand les développeurs d'un langage de programmation choisissent d'implanter tel ou tel algorithme, comme leur tri de tableau générique, ils ont ainsi un vaste choix de solutions. En plus des contraintes propres à leur langage, ils utilisent différents tests de performance guider leur décision. Parfois aussi, les ingénieurs innovent et sortent des sentiers bien balisés pour proposer des solutions complètement nouvelles. Les tests de performances ne sont pas simples à réaliser à ce niveau de généricité (les concepteurs d'un langages ne savent pas dans quels contextes leur langage va être utilisé), ils s'appuient soient sur des bases de tests récupérées en interne, soit sur une intuition qu'ont leurs ingénieurs sur les types de scénarios typiques d'utilisation de leurs structures.

Dans ce contexte notre projet adopte l'approche suivante :

1. Examiner le code source de différents langages de programmation (Java, PHP, Python, ...) pour identifier des choix étonnants ou discutables, ou encore des innovations.
2. Proposer des scénarios probabilistes qui modélisent des comportements réalistes d'utilisation de ces algorithmes.
3. Conduire une étude mathématique précise du comportement des algorithmes étudiés dans les différents scénarios retenus.
4. Confirmer l'efficacité de certains algorithmes, révéler les faiblesses d'autres et éventuellement proposer des améliorations substentielles.

Consortium : D'un point de vue international, il y a moins d'une quinzaine de théoriciens dans le monde qui poussent l'analyse d'algorithmes jusqu'au détail de leur implantations dans les langage de programmation, dans des modèles probabilistes. L'objectif de ce projet est d'initier une collaboration internationale sur ce thème, en regroupant un premier consortium de 5 chercheurs, d'apprendre à travailler tous ensemble, en vue de monter des projets plus conséquents en ouvrant aux autres intervenants du domaines. Nous nous connaissons par groupes de 2 ou 3 chercheurs, mais n'avons jamais eu d'action structurante pour développer réellement cet axe de recherche tous ensemble. Notre consortium est composé, dans l'ordre alphabétique, de :

- **Julien Clément**, GREYC, Univ. Caen, France : CR CNRS, spécialisé dans les structures de données à base d'arbres et en algorithmique du texte.
- **Conrado Martínez Parra**, UPC, Barcelone, Espagne : professeur, spécialisé dans l'analyse en moyenne d'algorithmes et des structures de données.
- **Cyril Nicaud**, LIGM, UGE, France : professeur, spécialisé dans l'analyse réalistes d'algorithmes de tri et de hachage, et dans les modèles tenant compte de l'architecture moderne d'ordinateurs.

- **Pablo Rotondo**, LIGM, UGE, France (**P.I.**) : MdC (recruté en 2021), spécialisé en analyse d'algorithmes et en dynamique symbolique, il a notamment révélé des failles dans les tables du langage de programmation LUA.
- **Alfredo Viola**, Univ. República Uruguay, Uruguay : professeur, spécialiste des tables de hachage et des analyses combinatoires d'algorithmes.

Faisabilité et résultats attendus : Tous les membres du consortium sont des spécialistes de l'analyse probabiliste et combinatoire d'algorithmes, et sont des experts dans les structures de données. Nous avons deux résultats importants dans les thèmes du projets :

- Nous sommes les premiers à avoir analysé la complexité de l'algorithme de tri de PYTHONet JAVA [1], et avons détecté une erreur dans l'implantation réalisée en JAVA, qui a été corrigée suite à notre signalement [2].
- Nous avons repéré des défauts d'efficacité dans les tables de LUA, qui sont la structure universelle de ce langage, et proposé des solutions pour y remédier.

Ce sont ces résultats et leur impact qui nous motivent pour continuer à faire profiter de notre expertise les ingénieurs qui développent les langages de programmation. A court terme, nous avons identifié plusieurs sujets à étudier notamment :

- Les *tableaux associatifs* de PHP : une première inspection du code nous laisse à penser qu'il pourrait y avoir des défauts structurels dans leur implantation. C'est une des structures de données principales dans ce langage, qui est très utilisé dans le développement web, l'impact d'une amélioration des performances serait important.
- Les *expressions régulières* en JAVA : le choix effectué est sous-optimal dans le pire cas, mais il reste à étudier si il est bien adapté à des scénarios réalistes.
- Les *ensembles* en PYTHON : la structure de table de hachage derrière leur implantation est assez particulière, et différente de celle utilisée pour les dictionnaires. Une étude fine des raisons de ce choix devraient être éclairante pour proposer des solutions mieux adaptées, éventuellement dans d'autres langages de programmation.

Au-delà de résultats précis comme ceux listés ci-dessus, nous souhaitons commencer à nous structurer internationalement, pour intégrer d'autres chercheurs reconnus comme Sebastien Wild (Univ. Liverpool, UK) ou Markus Nebel (Univ. Bielefeld, Allemagne) dans un projet de plus grande ampleur dans le futur.

Budget demandé – année civile 2023 : 5.5k€

- faire venir Julien Clément et Conrado Martínez une semaine au milieu de l'année (1400€). Organiser une réunion du projet à la fin de l'année en faisant venir tout le monde : Julien Clément, Conrado Martínez pendant une semaine (1400€) et Alfredo Viola pendant 15 jours, une semaine juste avec Pablo Rotondo et Cyril Nicaud, une semaine avec tout le monde (vol 1500€ , hébergement 1200€). Cette réunion permettra également de déterminer la suite du projet.

References

[1] Nicolas Auger, Vincent Jugé, Cyril Nicaud, Carine Pivoteau: On the Worst-Case Complexity of TimSort. *ESA 2018*. `https://arxiv.org/abs/1805.08612`.

[2] Patch note de Java pour corriger l'erreur. `http://hg.openjdk.java.net/jdk/jdk/rev/3a6d47df8239`

[3] Conrado Martínez, Cyril Nicaud, Pablo Rotondo: A Probabilistic Model Revealing Shortcomings in Lua's Hybrid Tables. *Accepted at COCOON 2022*. `https://arxiv.org/abs/2208.13602`