École Doctorale IAEM – Commission de mention informatique

Progress Report and CSI Report – 2019-20

This report contains 4 parts to be filled in French or English:
– an administrative form, to be filled by the PhD student
– the progress report of the PhD thesis, to be filled by the PhD student
– the conclusion of the supervisors
– the report of the Comité de Suivi Individuel (CSI), to be filled by the CSI

For the 4th part, the PhD student must plan a meeting with the "référent scientifique" (first member of his/her CSI) in May. The goal of this meeting is to discuss the progress of the thesis. Prior to this meeting, the other parts must be filled and the full document e-mailed to the référent scientifique.

***After this meeting**, the référent scientifique fills the 4th part and signs it. The PhD student has to make it signed also by the second member of the CSI. The whole document (PDF) has then to be uploaded by the PhD student to her/his ADUM profile and sent to the supervisors and the CMI Doctoral school (ed-iaem-cmi-contact@univ-lorraine.fr) before June 6th.*

In order to facilitate the process, the different parts can be prepared and signed (electronically) as separate documents, but they must be merged again as a single PDF file for uploading to ADUM.

Administrative information *(to be filled by the PhD student)*

PhD student: Matthieu Nicolas
Current PhD year in 2019-20: 3A
Laboratory: Loria
Supervisor(s): Olivier Perrin et Gérald Oster
Title of the PhD thesis: (Ré)Identification efficace dans les types de données répliquées sans conflit (CRDTs)
Fundings for the PhD thesis: Contrat doctoral

Members of the CSI:

- **scientific member:** Stephan Merz
- **auxiliary member:** Ye-Qiong Song

Progress report *(to be filled by the PhD student)*

Short description of the subject: *general context, considered approach, scientific objectives compared to the state of the art*

Dans le cadre de mes travaux de recherche, j'étudie et travaille sur les Conflict-free Replicated Data Types (CRDTs). Conçus pour les systèmes distribués hautement disponibles, ces types de données répliqués intégrent un mécanisme de résolution de conflits directement au sein de leur spécification afin de supporter nativement les modifications concurrentes. Pour résoudre les conflits de manière déterministe, les CRDTs utilisent généralement des identifiants qu'ils associent aux éléments stockés au sein de la structure de données. Cependant, selon le type de CRDT, les identifiants doivent respecter un ensemble de contraintes telles qu'être unique ou appartenir à un espace dense. Dans certains cas, ces contraintes empêchent de borner la taille des identifiants. La taille des identifiants croît alors continuellement avec le nombre de modifications effectuées, aggravant le surcoût lié à l'utilisation des CRDTs par rapport aux structures de données traditionnelles. Ce surcoût décourage l'adoption des CRDTs dans les systèmes distribués. Le but de cette thèse est de proposer des solutions pour pallier ce problème.

Nous nous sommes intéressés à un CRDT souffrant particulièrement du problème de croissance des identifiants : LogootSplit. Nous avons mis en place des benchmarks pour évaluer l'impact de la croissance des identifiants sur cette structure de données. Les mesures que nous avons réalisé ont montré que la proportion de méta-données pouvait augmenter au fil du temps jusqu'à représenter 99% de la structure de données. Ce résultat a confirmé l'intérêt d'apporter des solutions au problème soulevé.

Pour y répondre, nous proposons d'intégrer un mécanisme de renommage au sein des CRDTs. Ce mécanisme a pour but de permettre aux différents noeuds de renommer les identifiants afin de réduire leur taille, tout en respectant les contraintes imposées aux CRDTs. En particulier, le renommage doit se faire sans aucune coordination entre les noeuds. Afin de valider cette approche, nous avons conçu et implémenté un nouveau CRDT, RenamableLogootSplit, qui incorpore un tel mécanisme à LogootSplit.

Main results of the year: *studies and works achieved, results obtained with respect to the objectives of the thesis; difficulties encountered*

Cette année, nous avons procédé à une première validation de RenamablelogootSplit en prenant comme hypothèse l'absence d'opérations de renommage concurrentes. Nous avons effectué cette validation de manière expérimentale, par le biais de simulations reproduisant la rédaction d'un article par un ensemble de collaborateurs. Ces simulations nous ont tout d'abord permis de vérifier que l'ensemble des replicas convergeaient. Ce résultat nous a permis d'accroître la confiance que nous avons en la correction de RenamableLogootSplit, à défaut d'en proposer une preuve formelle.

Nous avons ensuite utilisé les traces issues de ces simulations pour comparer les performances de RenamableLogootSplit par rapport à LogootSplit sur les divers aspects impactés par l'accroissement des méta-données : taille de la structure, consommation de la bande-passante et temps d'intégration des modifications. Les benchmarks que nous avons réalisé montrent que RenamableLogootSplit obtient de meilleurs résultats que LogootSplit sur chacun de ces aspects : le renommage permet à terme de réinitialiser la quantité de méta-données de la structure et la taille des identifiants. De plus, le renommage permet aussi d'optimiser l'état de la structure. Cette optimisation permet d'intégrer plus rapidement les modifications suivantes.

Nous avons aussi mesuré le temps d'intégration de l'opération de renommage elle-même, en fonction de la taille de la structure de données. Les résultats obtenus révèlent que cette opération est coûteuse : son temps d'intégration varie de 100ms jusqu'à 1s en fonction de la taille de la structure. Si cette opération est déclenchée trop tardivement, elle peut être remarquée par les utilisateurs et impacter négativement leur expérience.

Nous avons rédigé un article scientifique décrivant ces travaux et les résultats obtenus. Ce papier a été publié et présenté dans le cadre du workshop PaPoC'20, qui regroupait de nombreux spécialistes du domaine des CRDTs.

Plan for next year: *remaining problems to be considered, approach; precise planning (in particular for those who are at least in 3rd year, for the report writing and the defense)*

Actuellement, nous travaillons sur une seconde validation expérimentale de RenamableLogootSplit. Dans le cadre de cette nouvelle validation, nous déclenchons des opérations concurrentes de renommage. Nous souhaitons ainsi tester le bon fonctionnement de notre mécanisme pour gérer les renommages concurrents et évaluer ses performances. Ce mécanisme consiste à définir un ordre de priorité sur les opérations de renommage concurrentes. Cet ordre permet de choisir de manière déterministe quelle opération appliquer à partir d'un ensemble d'opérations de renommage concurrentes. Si un replica a précédemment appliqué une opération de renommage concurrente non-prioritaire, des fonctions de transformation lui permettent d'annuler l'effet de cette dernière. Les expériences nous permettront aussi de confirmer et d'illustrer nos résultats théoriques de l'impact d'opérations de renommage concurrentes sur la taille de la structure de données. Notre objectif est de compléter ces travaux et de rédiger un article les présentant pour cet automne 2020.

En marge de cette seconde validation, d'autres pistes de recherche restent à explorer. Tout d'abord, il serait intéressant de proposer et d'évaluer plusieurs stratégies pour déterminer quand renommer la structure. L'objectif de ces stratégies serait de déclencher une opération de renommage avant que cette dernière ne soit trop coûteuse à appliquer sur la structure de données, mais tout en minimisant la probabilité que des opérations de renommage ne soient générées de manière concurrente par les autres replicas. Pour ce faire, les replicas pourraient se baser sur des métriques de l'état de la structure (taille de la structure, taille des identifiants...), mais aussi prendre en compte des informations liées à l'état du système (nombre de replicas total, nombre de replicas actuellement connectés...). Nous planifions d'étudier cette piste de recherche de l'automne 2020 au printemps 2021.

Une seconde piste de travail possible consisterait en la proposition et l'évaluation d'une nouvelle stratégie pour déterminer quel renommage privilégier en cas de renommages concurrents. Actuellement, nous définissons et utilisons un ordre de priorité sur les différents replicas pour effectuer ce choix. Bien que fonctionnelle, cette solution autorise des scénarios particulièrement inefficaces. Par exemple, un replica revenant d'une longue absence peut forcer le reste des pairs à annuler les renommages qu'ils ont effectué pendant ce temps s'il a généré une opération de renommage prioritaire. Il serait donc utile de définir une nouvelle stratégie minimisant les traitements effectués du point de vue global du système. La difficulté de sa conception réside dans le fait que l'ensemble des replicas doit converger. Les replicas doivent donc déterminer comme prioritaire une même opération de renommage d'un ensemble d'opérations concurrentes, indépendamment de leur état à la réception de cette dernière et de l'ordre de réception. Cette stratégie doit donc uniquement se baser sur des informations stables. Dans cette optique, nous pourrions par exemple utiliser une métrique du travail effectué avant le renommage (nombre de modifications, taille de la structure, nombre de participants...) pour prendre une décision. Actuellement, nous ne prévoyons pas d'explorer cette piste de recherche dans le cadre de cette thèse.

Aujourd'hui - Automne 2020	Validation de RenamableLogootSplit en cas d'opérations de renommage concurrentes
Automne 2020 - Printemps 2021	Proposition de stratégies pour gérer le déclenchement de l'opération de renommage
Automne 2020 - Été 2021	Rédaction du manuscript de thèse

Table 1: Récapitulatif du planning prévisionnel

Cependant, en fonction de notre progression sur les autres tâches, nous pourrons reconsidérer la question et démarrer des travaux sur ce sujet.

Enfin, l'année à venir sera consacrée à la rédaction du manuscript de thèse. Cette tâche sera effectuée tout au long de l'année, en parallèle des différentes pistes de travail présentés précédemment. L'objectif est de terminer la rédaction à l'été 2021 pour défendre courant automne 2021.

Publications: *if any*

[1] Matthieu Nicolas. Efficient renaming in CRDTs. In *Middleware 2018 - 19th ACM/IFIP International Middleware Conference (Doctoral Symposium)*, Rennes, France, December 2018.

[2] Matthieu Nicolas, Gérald Oster, and Olivier Perrin. Efficient Renaming in Sequence CRDTs. In *7th Workshop on Principles and Practice of Consistency for Distributed Data (PaPoC'20)*, Heraklion, Greece, April 2020.

Project after the thesis: *for 3rd year (and beyond) PhD students: professional project*
J'envisage actuellement soit de poursuivre une carrière dans l'académie en tant qu'enseigneur-chercheur, soit de m'orienter vers une carrière d'ingénieur R&D au sein de l'académie ou de l'industrie.

Scientific and professional modules validated: *list of all modules validated from the beginning of the thesis (do not forget to add them in Adum, together with your publications)*

Scientific modules

- Participation au module Réplication de données (M2 - Parcours SIS - Orientation SIRAV)
- Participation à l'école d'été SATIS 2018
- Participation à l'école d'été VTSA 2019

Professional modules

- Fi4 152 E Sauveteur Secouriste du Travail (SST)
- Fi4 162 C Formation à la communication orale et corporelle en milieu professionnel
- Fi4 282 Outils numériques pour la pédagogie (plateforme Arche, studio professeur)
- Fi4 305 Culture de l'intégrité scientifique
- PA1.1 MDD 14 - Eléments d'innovations pédagogiques

Date and signature of the PhD student
.....

Opinion of the supervisors *(to be filled by the supervisors)*

Opinion on this progress report:

.....

Agreement for an additional year? *Yes/No (if No, please justify)*

.....

Date of the defense: *this can be an approximation*

.....

Date and signature of the PhD supervisors

.....

Report of the Comité de Suivi Individuel *(to be filled by the référent scientifique)*

The questions below are suggestions. For each section, the report can be as short as "Yes" or more detailed according to the feeling of the CSI.

Is the student confident with the PhD progression? *E.g.: How does the PhD student view the progress of her/his thesis? How often does the student meet with her/his supervisors? How is the student's relationship with her/his supervisors? In case of problems, does the student know who to discuss these issues with?*

.....

Is the PhD on good tracks? *E.g: Is the student comfortable with his/her thesis topic? Did she/he embraced the subject? Do you have any comments or advice concerning publications? Does the student have opportunities to present her/his work? Do you have noteworthy concerns about the progression of the thesis?*

.....

Is the professional project sound? *E.g.: What is the professional project of the PhD student for after the PhD? Is she/he aware of the various options and associated expectations (postdoc abroad or in France, industrial R&D, application periods, teaching requirements, etc.)? Have contacts been established? Do you have comments about the PhD student's plans for her/his training courses/schools?*

.....

Conclusion: No problem for an additional registration / Interview with the ED recommended

Date: .....

Signatures:

PhD student	Two members of the CSI	
.....	.....	.....