

Ré-identification sans coordination dans les types de données répliquées sans conflits (CRDTs)

THÈSE

présentée et soutenue publiquement le TODO : Définir une date

pour l'obtention du

Doctorat de l'Université de Lorraine
(mention informatique)

par

Matthieu Nicolas

Composition du jury

<i>Président :</i>	Stephan Merz
<i>Rapporteurs :</i>	Le rapporteur 1 de Paris
	Le rapporteur 2
	suite taratata
	Le rapporteur 3
<i>Examineurs :</i>	L'examineur 1 d'ici
	L'examineur 2
<i>Membres de la famille :</i>	Mon frère
	Ma sœur

Mis en page avec la classe thesul.

Remerciements

Les remerciements.

*Je dédie cette thèse
à ma machine.
Oui, à Pandore,
qui fut la première de toutes.*

Sommaire

Introduction	1
1 Contexte	1
2 Questions de recherche	2
3 Contributions	2
4 Plan du manuscrit	2
Chapitre 1	
État de l’art	3
1.1 Systèmes distribués	3
1.2 Transformées opérationnelles	4
1.3 Types de données répliquées sans conflits	5
1.3.1 Sémantique en cas de concurrence	7
1.3.2 Modèle de synchronisation	11
1.3.3 Adoption dans la littérature et l’industrie	18
1.4 Séquences répliquées sans conflits	19
1.4.1 Approche à pierres tombales	19
1.4.2 Approche à identifiants densément ordonnés	19
1.4.3 Synthèse	19
1.5 LogootSplit	20
1.5.1 Identifiants	20
1.5.2 Aggrégation dynamique d’éléments en blocs	21
1.5.3 Modèle de données	22
1.5.4 Modèle de livraison	24
1.5.5 Limites	26
1.6 Mitigation du surcoût des séquences répliquées sans conflits	27
1.7 Synthèse	28
1.8 Proposition	29

Chapitre 2**Renommage dans une séquence répliquée 31**

2.1	Présentation de l'approche	32
2.1.1	Modèle du système	32
2.1.2	Définition de l'opération de renommage	32
2.2	Introduction de l'opération <i>rename</i>	34
2.2.1	Opération de renommage proposée	34
2.2.2	Gestion des opérations concurrentes au renommage	35
2.2.3	Évolution du modèle de livraison des opérations	38
2.3	Gestion des opérations <i>rename</i> concurrentes	40
2.3.1	Conflits en cas de renommages concurrents	40
2.3.2	Relation de priorité entre renommages	41
2.3.3	Algorithme d'annulation de l'opération de renommage	42
2.3.4	Processus d'intégration d'une opération	47
2.3.5	Règles de récupération de la mémoire des états précédents	51
2.4	Validation	54
2.4.1	Complexité en temps des opérations	54
2.4.2	Expérimentations	58
2.4.3	Résultats	59
2.5	Discussion	66
2.5.1	Stratégie de génération des opérations <i>rename</i>	66
2.5.2	Stockage des états précédents sur disque	67
2.5.3	Compression et limitation de la taille de l'opération <i>rename</i>	67
2.5.4	Définition de relations de priorité pour minimiser les traitements	68
2.5.5	Report de la transition vers la nouvelle époque cible	69
2.5.6	Utilisation de l'opération de renommage comme mécanisme de compression du log d'opérations	70
2.5.7	Implémentation alternative de l'intégration de l'opération <i>rename</i> basée sur le log d'opérations	72
2.6	Comparaison avec les approches existantes	74
2.6.1	Core-Nebula	74
2.6.2	LSEQ	75
2.7	Conclusion	75

Chapitre 3**MUTE, un éditeur de texte web collaboratif P2P temps réel chiffré de bout en bout****77**

3.1	Couche interface utilisateur	78
3.2	Couche réplication	80
3.2.1	Modèle de données du document texte	80
3.2.2	Module de livraison des opérations	80
3.2.3	Collaborateurs	88
3.2.4	Curseurs	92
3.3	Couche réseau	92
3.3.1	Établissement d'un réseau Pair-à-Pair (P2P) entre navigateurs . . .	92
3.3.2	Topologie réseau	94
3.4	Couche sécurité	94
3.5	Conclusion	96

Chapitre 4**Conclusions et perspectives****97**

4.1	Résumé des contributions	98
4.2	Perspectives	98
4.2.1	Définition de relations de priorité pour minimiser les traitements . .	98
4.2.2	Redéfinition de la sémantique du renommage en déplacement d'éléments	98
4.2.3	Définition de types de données répliquées sans conflits plus complexes	98
4.2.4	Étude comparative des différentes familles de CRDTs	98
4.2.5	Définition d'opérations supplémentaires pour fonctionnalités liées à l'édition collaborative	99
4.2.6	Conduction d'expériences utilisateurs d'édition collaborative	99
4.2.7	Comparaison des mécanismes de synchronisation	100
4.2.8	Distance entre versions d'un document	100
4.2.9	Contrôle d'accès	100
4.2.10	Détection et éviction de pairs malhonnêtes	100
4.2.11	Vecteur <i>epoch-based</i>	101
4.2.12	Fusion de versions distantes d'un document collaboratif	102
4.2.13	Rôles et places des bots dans systèmes collaboratifs	102

Annexe A

Algorithmes `RENAMEID`

Annexe B

Algorithmes `REVERTRENAMEID`

Index

107

Bibliographie

Table des figures

1.1	Modifications concurrentes d'un Ensemble répliqué provoquant un conflit provoqué suite à l'ajout et la suppression d'un même élément	6
1.2	Résolution du conflit en utilisant la sémantique <i>Last-Writer-Wins</i> (LWW)	7
1.3	Résolution du conflit en utilisant la sémantique <i>Multi-Value</i> (MV)	8
1.4	Résolution du conflit en utilisant soit la sémantique <i>Add-Wins</i> (AW), soit la sémantique <i>Remove-Wins</i> (RW)	10
1.5	Résolution du conflit en utilisant la sémantique <i>Causal-Length</i> (CL)	10
1.6	Modifications en concurrence d'un Ensemble répliqué par les noeuds A et B	11
1.7	Synchronisation des noeuds A et B en adoptant le modèle de synchronisation par états	13
1.8	Synchronisation des noeuds A et B en adoptant le modèle de synchronisation par opérations	15
1.9	Synchronisation des noeuds A et B en adoptant le modèle de synchronisation par différences d'états	17
1.10	Représentation d'une séquence LogootSplit contenant les éléments "HLO"	22
1.11	Modifications concurrentes d'une séquence répliquée LogootSplit	23
1.12	Résurgence d'un élément supprimé suite à la relivraison de son opération <i>insert</i>	24
1.13	Non-effet de l'opération <i>remove</i> car reçue avant l'opération <i>insert</i> correspondante	25
1.14	Insertion menant à une augmentation de la taille des identifiants	27
2.1	Renommage de la séquence sur le noeud A	34
2.2	Modifications concurrentes menant à une anomalie	35
2.3	Renommage de la modification concurrente avant son intégration en utilisant RENAMEID afin de maintenir l'ordre souhaité	37
2.4	Livraison d'une opération <i>insert</i> sans avoir reçu l'opération <i>rename</i> précédente	38
2.5	Livraison désordonnée d'une opération <i>rename</i> et de l'opération <i>insert</i> qui la précède	39
2.6	Opérations <i>rename</i> concurrentes menant à des états divergents	40
2.7	<i>Arbre des époques</i> correspondant au scénario décrit dans la Figure 2.6 . . .	40
2.8	Sélectionner l'époque cible d'une exécution d'opérations <i>rename</i> concurrentes	42
2.9	Annulation d'une opération <i>rename</i> intégrée précédemment en présence d'un identifiant inséré en concurrence	43

2.10	Annulation d'une opération <i>rename</i> intégrée précédemment en présence d'un identifiant inséré causalement après	45
2.11	Intégration d'une opération <i>insert</i> distante	47
2.12	Intégration d'une opération <i>rename</i> distante	51
2.13	Suppression des époques obsolètes et récupération de la mémoire des <i>anciens états</i> associés	52
2.14	Évolution de la taille du document en fonction du Conflict-free Replicated Data Type (CRDT) utilisé et du nombre de <i>renaming bots</i> dans la collaboration	60
2.15	Temps d'intégration des opérations standards	61
2.16	Progression du nombre d'opérations du log rejouées en fonction du temps	64
2.17	Livraison d'une opération <i>rename</i> d'un noeud	69
3.1	Architecture système de l'application MUTE	78
3.2	Architecture logicielle de l'application MUTE	79
3.3	Capture d'écran d'une session d'édition collaborative avec MUTE	79
3.4	Gestion de la livraison <i>exactly-once</i> des opérations	82
3.5	Gestion de la livraison <i>causale-remove</i> des opérations	83
3.6	Gestion de la livraison <i>epoch based</i> des opérations	85
3.7	Utilisation du mécanisme d'anti-entropie par le noeud C pour se synchroniser avec le noeud B	87
3.8	Exécution du mécanisme de détection des défaillances par le noeud C pour tester le noeud B	88
3.9	Architecture système pour la couche réseau de MUTE	93
3.10	Architecture système pour la couche sécurité de MUTE	95

Introduction

1 Contexte

- Systèmes collaboratifs (wikis, plateformes de contenu, réseaux sociaux) et leurs bienfaits (qualité de l'info, vitesse de l'info (exemple de crise ?), diffusion de la parole). Démocratisation (sic) de ces systèmes au cours de la dernière décennie.
- En raison du volume de données et de requêtes, adoptent architecture décentralisée. Permet ainsi de garantir disponibilité, tolérance aux pannes et capacité de passage à l'échelle.
- Mais échoue à adresser problèmes non-techniques : confidentialité, souveraineté, protection contre censure, dépendance et nécessité de confiance envers autorité centrale.
- À l'heure où les entreprises derrière ces systèmes font preuve d'ingérence et d'intérêts contraires à ceux de leurs utilisateur-rices (Cambridge Analytica, Prism, non-modération/mise en avant de contenus racistes, invisibilisation de contenus féministes), paraît fondamental de proposer les moyens technologiques accessibles pour concevoir et déployer des alternatives.
- *Matthieu: TODO : Voir si angle écologique/réduction consommation d'énergie peut être pertinent.*
- Systèmes pair-à-pair sont une direction intéressante pour répondre à ces problématiques, de part leur absence d'autorité centrale, la distribution des tâches et leur conception mettant le pair au centre. Mais posent de nouvelles problématiques de recherche.
- Ces systèmes ne disposent d'aucun contrôle sur les noeuds qui les composent. Le nombre de noeuds peut donc croître de manière non-bornée et atteindre des centaines de milliers de noeuds. La complexité des algorithmes de ces systèmes ne doit donc pas dépendre de ce paramètre, ou alors de manière logarithmique.
- De plus, ces noeuds n'offrent aucune garantie sur leur stabilité. Ils peuvent donc rejoindre et participer au système de manière éphémère. S'agit du phénomène connu sous le nom de churn. Les algorithmes de ces systèmes ne peuvent donc pas reposer sur des mécanismes nécessitant une coordination synchrone d'une proportion des noeuds.
- Finalement, ces noeuds n'offrent aucune garanties sur leur fiabilité et intentions. Les noeuds peuvent se comporter de manière byzantine. Pour assurer la confidentialité, l'absence de confiance requise et le bon fonctionnement du système, ce dernier doit être conçu pour résister aux comportements byzantins de ses acteurs.

- Ainsi, il est nécessaire de faire progresser les technologies existantes pour les rendre compatible avec ce nouveau modèle de système. Dans le cadre de cette thèse, nous écartons la problématique des adversaires byzantins et nous intéressons aux mécanismes de réplication de données dans les systèmes collaboratifs pair-à-pair temps réel.

2 Questions de recherche

3 Contributions

4 Plan du manuscrit

Chapitre 1

État de l’art

Sommaire

1.1	Systèmes distribués	3
1.2	Transformées opérationnelles	4
1.3	Types de données répliquées sans conflits	5
1.3.1	Sémantique en cas de concurrence	7
1.3.2	Modèle de synchronisation	11
1.3.3	Adoption dans la littérature et l’industrie	18
1.4	Séquences répliquées sans conflits	19
1.4.1	Approche à pierres tombales	19
1.4.2	Approche à identifiants densément ordonnés	19
1.4.3	Synthèse	19
1.5	LogootSplit	20
1.5.1	Identifiants	20
1.5.2	Aggrégation dynamique d’éléments en blocs	21
1.5.3	Modèle de données	22
1.5.4	Modèle de livraison	24
1.5.5	Limites	26
1.6	Mitigation du surcoût des séquences répliquées sans conflits	27
1.7	Synthèse	28
1.8	Proposition	29

1.1 Systèmes distribués

- Contexte des systèmes distribués à large échelle
- Réplique les données afin de pouvoir supporter les pannes
- Adopte le paradigme de la réplication optimiste [66]
- Autorise les noeuds à consulter et à modifier la donnée sans aucune coordination entre eux

- Autorise alors les noeuds à diverger temporairement
- Permet d'être toujours disponible, de toujours répondre aux requêtes même en cas de partition réseau
- Permet aussi, en temps normal, de réduire le temps de réponse (privilégie la latence) [1]
- Comme ce modèle autorise les noeuds à modifier la donnée sans se coordonner, possible d'effectuer des modifications concurrentes
- Généralement, un mécanisme de résolution de conflits est nécessaire afin d'assurer la convergence des noeuds dans une telle situation
- Plusieurs approches ont été proposées pour implémenter un tel mécanisme

1.2 Transformées opérationnelles

- Approche permettant de gérer des modifications concurrentes sur un type de données
- Consiste à transformer les opérations par rapport aux effets des opérations concurrentes pour rendre les rendre commutatives. Permet de rendre l'ordre d'intégration des opérations sans importance par rapport à l'état final obtenu
- Se décompose en 2 parties : algorithmes (génériques) et fonctions de transformations (spécifiques au type de données)
- Plusieurs algorithmes OT adoptent une architecture centralisée (trouver citations)
- Cette architecture pose des problèmes de performances (bottleneck), sécurité (SPOF), coût, d'utilisabilité (mode offline), pérennité (disparition du service), vie privée et de résistance à la censure.
- Pour ces raisons, des algorithmes reposant sur une architecture décentralisée ont été proposés
- Mais ne règlent qu'en partie ces limites
- Notamment, ne sont pas adaptés à des systèmes P2P dynamiques
- Besoin de vector clocks sur chaque opération pour détecter la concurrence. Vector clocks adaptés dans systèmes à nombre de pairs fixe, mais pas aux systèmes dynamiques (revoir causal barrier pour p-e nuancer ce propos).
- Néanmoins, cette approche a permis de démocratiser les systèmes collaboratifs via son adoption par des services tels que Google Docs, Overleaf, Framapad
- De plus, dans le cadre de ces travaux, ont été définies les propriétés CCI [70].
- Remettre en question la propriété Causalité des CCI. Généralement, confond causalité et happen-before et exprime en finalité une contrainte trop forte. Cette contrainte peut réduire la réactivité du système (exemple avec 2 insertions sans liens mais qui force d'attendre la 1ère pour intégrer la 2nde). Causalité pose aussi des problèmes de passage à l'échelle car repose sur vector clocks. IMO, doit relaxer cette propriété pour pouvoir construire systèmes à large échelle.

Matthieu: TODO : Mentionner TP1 et TP2

Matthieu: TODO : Spécification faible et forte des séquences répliquées

1.3 Types de données répliquées sans conflits

Pour limiter la coordination entre les noeuds, les systèmes distribués adoptent le paradigme de la réplication optimiste. Ce paradigme consiste à ce que chaque noeud possède une copie de la donnée. Chaque noeud possède le droit de la modifier sans se coordonner avec les autres noeuds. Noeuds peuvent alors temporairement diverger, c.-à-d. posséder des états différents. Un mécanisme de synchronisation leur permet ensuite de partager leurs modifications respectives et de nouveau converger. Ce paradigme offre ainsi aux noeuds une haute disponibilité *Matthieu: TODO : insérer ref sticky model* ainsi qu'une faible latence.

Afin d'ordonner les modifications effectués dans un système, la littérature repose généralement sur la relation *happens-before* [41]. Nous l'adaptions ci-dessous à notre contexte :

Définition 1 (Relation *happens-before*) *La relation happens-before indique qu'une modification m_1 a eu lieu avant une modification m_2 , notée $m_1 \rightarrow m_2$, si et seulement si une des conditions suivantes est respectée :*

- (i) m_1 a eu lieu avant m_2 sur le même noeud.
- (ii) m_1 a été délivrée au noeud auteur de m_2 avant la génération de m_2 .
- (iii) Il existe une modification m telle que $m_1 \rightarrow m \wedge m \rightarrow m_2$.

Dans le cadre d'un système distribué, on note que la relation *happens-before* ne permet pas d'établir un ordre total entre les modifications¹. En effet, deux modifications m_1 et m_2 peuvent être effectuées en parallèle par deux noeuds différents, sans avoir connaissance de la modification de leur pair respectif. De telles modifications sont alors dites *concurrentes* :

Définition 2 (Concurrence) *Deux modifications m_1 et m_2 sont concurrentes, noté $m_1 \parallel m_2$, si et seulement si $m_1 \not\rightarrow m_2 \wedge m_1 \not\leftarrow m_2$.*

Lorsque les modifications possibles sur un type de données sont commutatives, l'intégration des modifications effectuées par les autres noeuds, même concurrentes, ne nécessite aucun mécanisme particulier. Cependant, les modifications permises par un type de données ne sont généralement pas commutatives car de sémantiques contraires. Ainsi, une exécution dans un système distribué suivant le paradigme de réplication optimiste peut mener à la génération de modifications concurrentes non commutatives. Nous parlons alors de conflits. La figure Figure 1.1 présente un scénario où des modifications de sémantiques opposées sont générées en concurrence.

Dans cet exemple, deux noeuds A et B répliquent et partagent un même Ensemble. Les deux noeuds possèdent le même état initial : $\{a\}$. Le noeud A retire l'élément a de l'ensemble, $rmv(a)$. Puis, le noeud A ré-ajoute l'élément a dans l'ensemble via l'opération de modification $add(a)$. En concurrence, le noeud B retire lui aussi l'élément a de l'ensemble. Les deux noeuds se synchronisent ensuite.

À l'issue de ce scénario, l'état à produire n'est pas trivial : le noeud A a exprimé son intention d'ajouter l'élément a à l'ensemble, tandis que le noeud B a exprimé son intention

1. Nous utilisons le terme *modifications* pour désigner les *opérations de modifications* des types abstraits de données afin d'éviter une confusion avec le terme *opération* introduit ultérieurement.

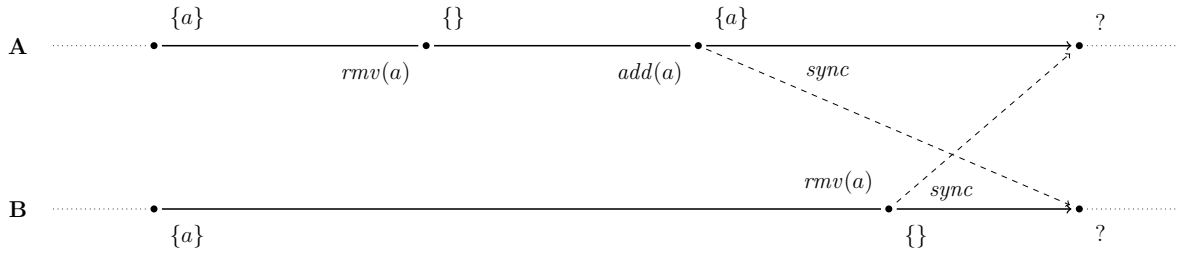


FIGURE 1.1 – Modifications concurrentes d'un Ensemble répliqué provoquant un conflit provoqué suite à l'ajout et la suppression d'un même élément

contraire de retirer l'élément a de ce même ensemble. Ainsi, les états $\{a\}$ et $\{\}$ semblent tous les deux corrects et légitimes dans cette situation. Il est néanmoins primordial que les noeuds choisissent et convergent vers un même état pour leur permettre de poursuivre leur collaboration. Pour ce faire, il est nécessaire de mettre en place un mécanisme de résolution de conflits, potentiellement automatique.

Les CRDTs répondent à ce besoin. Les CRDTs sont de nouvelles spécifications des types de données existants, e.g. l'Ensemble ou la Séquence. Ces nouvelles spécifications sont conçues pour être utilisées dans des systèmes distribués adoptant la réplication optimiste. Ainsi, elles offrent les deux propriétés suivantes :

- (i) Les CRDTs peuvent être modifiés sans coordination avec les autres noeuds.
- (ii) Tous les noeuds ayant observés le même ensemble de modifications obtiendront des états équivalents. Cette propriété est connue sous le nom de *convergence forte* Matthieu: TODO : Ajouter ref.

Pour offrir la propriété de *convergence forte*, la spécification des CRDTs reposent sur la théorie des treillis :

- (i) Les différents états possibles d'un CRDT forment un sup-demi-treillis, possédant une relation d'ordre partiel \leq .
- (ii) Les modifications génèrent par inflation un nouvel état supérieur ou égal à l'état original d'après \leq .
- (iii) Il existe une fonction de fusion qui, pour toute paire d'états, génère l'état minimal supérieur d'après \leq aux deux états fusionnés. On parle aussi de borne supérieure ou de Least Upper Bound (LUB) pour catégoriser l'état résultant de cette fusion.

Malgré leur spécification différente, les CRDTs partagent la même sémantique et interface que les types non-répliqués correspondant du point de vue des utilisateurs. Ainsi, les CRDTs partagent le comportement des types non-répliqués dans le cadre d'exécutions séquentielles. Cependant, ils définissent aussi une nouvelle sémantique pour chaque nouveau scénario ne pouvant se produire que dans le cadre d'une exécution distribuée.

Plusieurs sémantiques valides peuvent être proposées pour ces nouveaux scénarios. Un CRDT se doit donc de préciser quelle sémantique il choisit.

L'autre aspect définissant un CRDT donné est le modèle qu'il adopte pour propager les modifications. Au fil des années, la littérature a établi et défini plusieurs modèles dit de synchronisation, chacun ayant ses propres besoins et avantages. De fait, plusieurs CRDTs

peuvent être proposés pour un même type donné en fonction du modèle de synchronisation choisi.

Ainsi, ce qui définit un CRDT est sa sémantique en cas de concurrence et son modèle de synchronisation. Dans les prochaines sections, nous présenterons les différentes sémantiques possibles pour un type donné, l'Ensemble, en guise d'exemple. Puis nous présenterons les différents modèles de synchronisation proposés dans la littérature, et détaillerons leurs contraintes et impact sur les CRDT les adoptant, toujours en utilisant le même exemple.

Matthieu: TODO : Faire le lien avec les travaux de Burckhardt [13] et les MRDTs [37]

1.3.1 Sémantique en cas de concurrence

Plusieurs sémantiques peuvent être proposées pour résoudre les conflits. Certaines de ces sémantiques ont comme avantage d'être générique, c.-à-d. applicable à l'ensemble des types de données. En contrepartie, elles souffrent de cette même généralité, en ne permettant que des comportements simples en cas de conflits.

À l'inverse, la majorité des sémantiques proposées dans la littérature sont spécifiques à un type de données. Elles visent ainsi à prendre plus finement en compte l'intention des modifications pour proposer des comportements plus précis.

Dans la suite de cette section, nous présentons ces sémantiques génériques ainsi que celles spécifiques à l'Ensemble et, à titre d'exemple, les illustrons à l'aide du scénario présenté dans la Figure 1.1.

Sémantique *Last-Writer-Wins*

Une manière simple pour résoudre un conflit consiste à trancher de manière arbitraire et de sélectionner une modification parmi l'ensemble des modifications en conflit. Pour faire cela de manière déterministe, une approche est de reproduire et d'utiliser l'ordre total sur les modifications qui serait instauré par une horloge globale pour choisir la modification à prioriser.

Cette approche, présentée dans [36], correspond à la sémantique nommée *Last-Writer-Wins* (LWW). De par son fonctionnement, cette sémantique est générique et est donc utilisée par une variété de CRDTs pour des types différents. La Figure 1.2 illustre son application à l'Ensemble pour résoudre le conflit de la Figure 1.1.

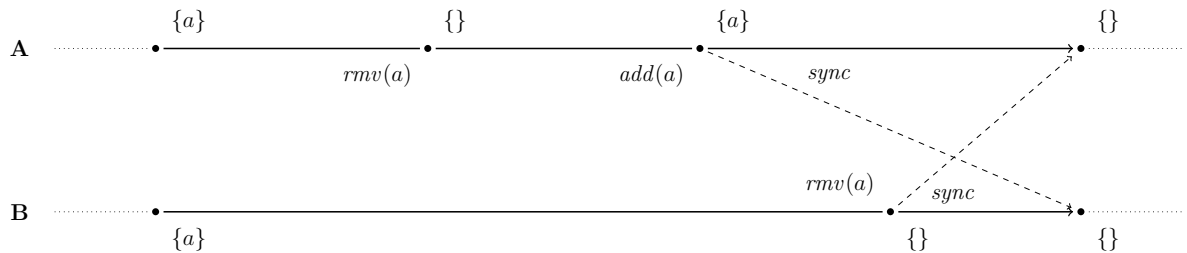


FIGURE 1.2 – Résolution du conflit en utilisant la sémantique LWW

Comme indiqué précédemment, le scénario illustré dans la Figure 1.2 présente un conflit entre les modifications concurrentes $add(a)$ et $rmv(a)$ générées de manière concurrente respectivement par les noeuds A et B. Pour le résoudre, la sémantique LWW associe à chaque modification une estampille. L'ordre créé entre les modifications par ces dernières permet de déterminer quelle modification désigner comme prioritaire. Ici, nous considérons que $add(a)$ a eu lieu plus tôt que $rmv(a)$. La sémantique LWW désigne donc $rmv(a)$ comme prioritaire et ignore $add(a)$. L'état obtenu à l'issue de cet exemple par chaque noeud est donc $\{\}$.

Il est à noter que si la modification $rmv(a)$ du noeud B avait eu lieu plus tôt dans notre exemple, l'état final obtenu aurait été $\{a\}$. Ainsi, des exécutions reproduisant le même ensemble de modifications produiront des résultats différents en fonction de l'ordre créé par les estampilles associées à chaque modification. Ces estampilles étant des métadonnées du mécanisme de résolution de conflits, elles sont dissimulées aux utilisateur-rices. Le comportement de cette sémantique peut donc être perçu comme aléatoire et s'avérer perturbant pour les utilisateur-rices.

La sémantique LWW repose sur l'horloge de chaque noeud pour attribuer une estampille à chacune de leurs modifications. Les horloges physiques étant sujettes à des imprécisions et notamment des décalages, utiliser les estampilles qu'elles fournissent peut provoquer des anomalies vis-à-vis de la relation *happens-before*. Les systèmes distribués préfèrent donc généralement utiliser des horloges logiques [41]. *Matthieu: TODO : Ajouter refs des horloges logiques plus intelligentes (Interval Tree Clock, Hybrid Clock...)*

Sémantique *Multi-Value*

Une seconde sémantique générique² est la sémantique *Multi-Value* (MV). Cette approche propose de gérer les conflits de la manière suivante : plutôt que de prioriser une modification par rapport aux autres modifications concurrentes, la sémantique MV maintient l'ensemble des états résultant possibles. Nous présentons son application à l'Ensemble dans la Figure 1.3.

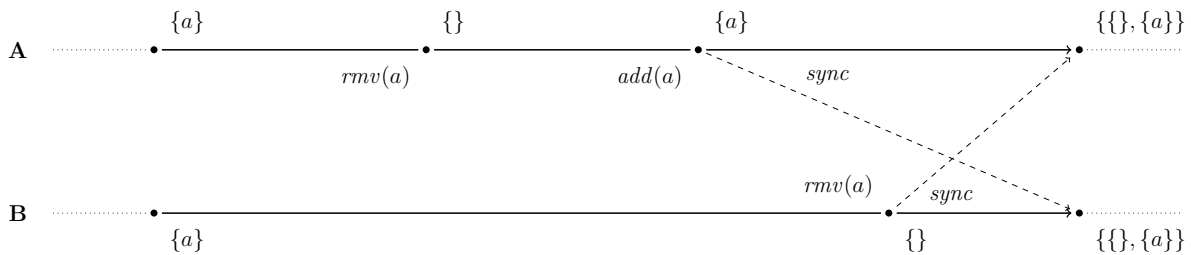


FIGURE 1.3 – Résolution du conflit en utilisant la sémantique MV

La Figure 1.3 présente la gestion du conflit entre les modifications concurrentes $add(a)$ et $rmv(a)$ par la sémantique MV. Devant ces modifications contraires, chaque noeud calcule chaque état possible, c.-à-d. un état sans l'élément a , $\{\}$, et un état avec ce dernier, $\{a\}$. Le CRDT maintient alors l'ensemble de ces états en parallèle. L'état obtenu est donc $\{\{\}, \{a\}\}$.

2. Bien que généralement associée au type *Registre* uniquement.

Ainsi, la sémantique MV expose les conflits aux utilisateur-rices lors de leur prochaine consultation de l'état du CRDT. Les utilisateur-rices peuvent alors prendre connaissance des intentions de chacun-e et résoudre le conflit manuellement. Dans la Figure 1.3, résoudre le conflit revient à re-effectuer une modification $add(a)$ ou $rmv(a)$ selon l'état choisi. Ainsi, si plusieurs personnes résolvent en concurrence le conflit de manière contraire, la sémantique MV exposera de nouveau les différents états proposés sous la forme d'un conflit.

Il est intéressant de noter que cette sémantique mène à un changement du domaine du CRDT considéré : en cas de conflit, la valeur retournée par le CRDT correspond à un Ensemble de valeurs du type initialement considéré. E.g. si nous considérons que le type correspondant au CRDT dans la Figure 1.3 est le type $Set\langle V \rangle$, nous observons que la valeur finale obtenue a pour type $Set\langle Set\langle V \rangle \rangle$. Il s'agit à notre connaissance de la seule sémantique opérant ce changement.

Sémantiques *Add-Wins* et *Remove-Wins*

Comme évoqué précédemment, d'autres sémantiques sont spécifiques au type de données concerné. Ainsi, nous abordons à présent des sémantiques spécifiques au type de l'Ensemble.

Dans le cadre de l'Ensemble, un conflit est provoqué lorsque des modifications add et rmv d'un même élément sont effectuées en concurrence. Ainsi, deux approches peuvent être proposées pour résoudre le conflit :

- (i) Une sémantique où la modification add d'un élément prend la précedence sur les modifications concurrentes rmv du même élément, nommée *Add-Wins* (AW). L'élément est alors présent dans l'état obtenu à l'issue de la résolution du conflit.
- (ii) Une sémantique où la modification rmv d'un élément prend la précedence sur les opérations concurrentes add du même élément, nommée *Remove-Wins* (RW). L'élément est alors absent de l'état obtenu à l'issue de la résolution du conflit.

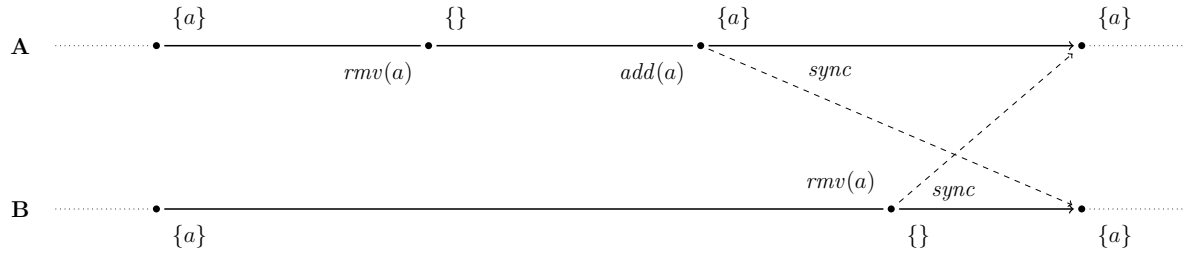
La Figure 1.4 illustre l'application de chacune de ces sémantiques sur notre exemple.

Sémantique *Causal-Length*

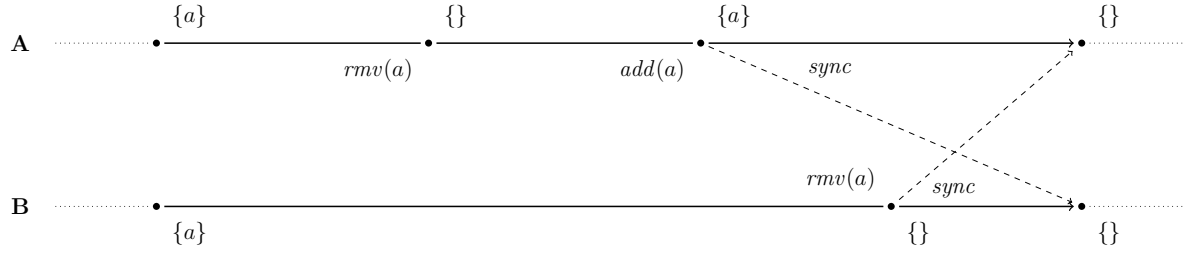
Une nouvelle sémantique pour l'Ensemble fut proposée [80] récemment. Cette sémantique se base sur les observations suivantes :

- (i) add et rmv d'un élément prennent place à tour de rôle, chaque modification invalidant la précédente.
- (ii) add (resp. rmv) concurrents d'un même élément représentent la même intention. Prendre en compte une de ces modifications concurrentes revient à prendre en compte leur ensemble.

À partir de ces observations, YU et ROSTAD proposent de déterminer pour chaque élément la chaîne d'ajouts et retraits la plus longue. C'est cette chaîne, et précisément son dernier maillon, qui indique si l'élément est présent ou non dans l'ensemble final. La Figure 1.5 illustre son fonctionnement.



(a) Application de la sémantique AW



(b) Application de la sémantique RW

FIGURE 1.4 – Résolution du conflit en utilisant soit la sémantique AW, soit la sémantique RW

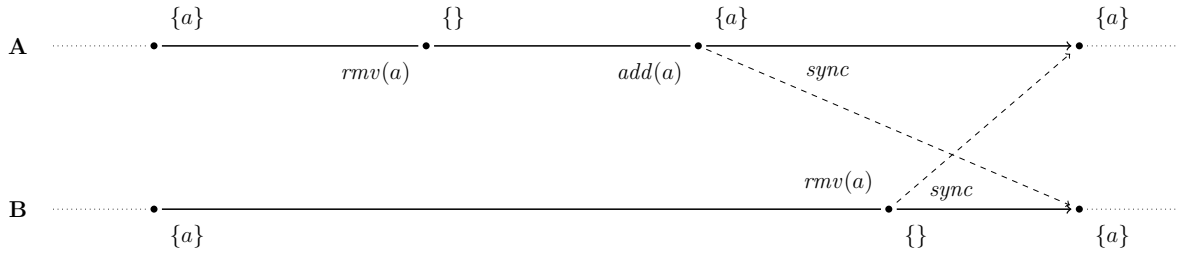


FIGURE 1.5 – Résolution du conflit en utilisant la sémantique CL

Dans notre exemple, la modification $rmv(a)$ effectuée par B est en concurrence avec une modification identique effectuée par A. La sémantique CL définit que ces deux modifications partagent la même intention. Ainsi, A ayant déjà appliqué sa propre modification préalablement, il ne prend pas en compte *de nouveau* cette modification lorsqu'il la reçoit de B. Son état reste donc inchangé.

À l'inverse, la modification $add(a)$ effectuée par A fait suite à sa modification $rmv(a)$. La sémantique CL définit alors qu'elle fait suite à toute autre modification $rmv(a)$ concurrente. Ainsi, B intègre cette modification lorsqu'il la reçoit de A. Son état évolue donc pour devenir $\{a\}$.

Synthèse

Dans cette section, nous avons mis en lumière l'existence de solutions différentes pour résoudre un même conflit. Chacune de ces solutions correspond à une sémantique spécifique de résolution de conflits. Ainsi, pour un même type de données, différents CRDTs

peuvent être spécifiés. Chacun de ces CRDTs est spécifié par la combinaison de sémantiques qu'il adopte, chaque sémantique servant à résoudre un des types de conflits du type de données.

Il est à noter qu'aucune sémantique n'est intrinsèquement meilleure et préférable aux autres. Il revient aux concepteur-rices d'applications de choisir les CRDTs adaptés en fonction des besoins et des comportements attendus en cas de conflits.

Par exemple, pour une application collaborative de listes de courses, l'utilisation d'un MV-Registre pour représenter le contenu de la liste se justifie : cette sémantique permet d'exposer les modifications concurrentes aux utilisateur-rices. Ainsi, les personnes peuvent détecter et résoudre les conflits provoqués par ces éditions concurrentes, e.g. l'ajout de l'élément *lait* à la liste, pour cuisiner des crêpes, tandis que les *oeufs* nécessaires à ces mêmes crêpes sont retirés. En parallèle, cette même application peut utiliser un LWW-Registre pour représenter et indiquer aux utilisateur-rices la date de la dernière modification effectuée.

1.3.2 Modèle de synchronisation

Dans le modèle de réplication optimiste, les noeuds divergent momentanément lorsqu'ils effectuent des modifications locales. Pour ensuite converger vers des états équivalents, les noeuds doivent propager et intégrer l'ensemble des modifications. La Figure 1.6 illustre ce point.

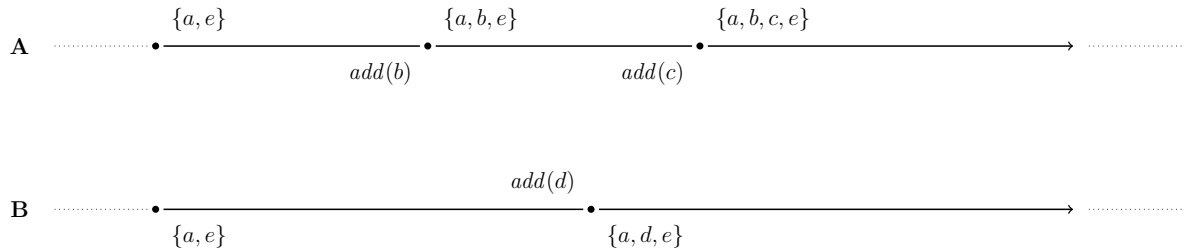


FIGURE 1.6 – Modifications en concurrence d'un Ensemble répliqué par les noeuds A et B

Dans cet exemple, deux noeuds A et B partagent et éditent un même Ensemble à l'aide d'un CRDT. Les deux noeuds possèdent le même état initial : $\{a, e\}$.

Le noeud A effectue les modifications $add(b)$ puis $add(c)$. Il obtient ainsi l'état $\{a, b, c, e\}$. De son côté, le noeud B effectue la modification suivante : $add(d)$. Son état devient donc $\{a, d, e\}$. Ainsi, les noeuds doivent encore s'échanger leur modifications pour converger vers l'état souhaité³, c.-à-d. $\{a, b, c, d, e\}$.

Dans le cadre des CRDTs, le choix de la méthode pour synchroniser les noeuds n'est pas anodin. En effet, ce choix impacte la spécification même du CRDT et ses prérequis.

Initialement, deux approches ont été proposées : une méthode de synchronisation par états [68, 67] et une méthode de synchronisation par opérations [68, 67, 9, 8]. Une troisième

3. Le scénario ne comportant uniquement des modifications add , aucun conflit n'est produit malgré la concurrence des modifications.

approche, nommée synchronisation par différence d'états [4, 3], fut spécifiée par la suite. Le but de cette dernière est d'allier le meilleur des deux approches précédentes.

Dans la suite de cette section, nous présentons ces approches ainsi que leurs caractéristiques respectives. Pour les illustrer, nous complétons l'exemple décrit ici. Cependant, nous nous focalisons uniquement sur les messages envoyés par les noeuds et n'évoquons seulement les métadonnées introduites par chaque modèle de synchronisation, par soucis de clarté et de simplicité.

Synchronisation par états

L'approche de la synchronisation par états propose que les noeuds diffusent leurs modifications en transmettant leur état. Les CRDTs adoptant cette approche doivent définir une fonction `merge`. Cette fonction correspond à la fonction de fusion mentionnée précédemment en (iii) : elle prend en paramètres une paire d'états et génère en retour l'état correspondant à leur LUB. Cette fonction doit être associative, commutative et idempotente.

Ainsi, lorsqu'un noeud reçoit l'état d'un autre noeud, il fusionne ce dernier avec son état courant à l'aide de la fonction `merge`. Il obtient alors un nouvel état intégrant l'ensemble des modifications ayant été effectuées sur les deux états.

La nature croissante des états des CRDTs couplée aux propriétés d'associativité, de commutativité et d'idempotence de la fonction `merge` permettent de reposer sur la couche réseau sans lui imposer de contraintes fortes : les messages peuvent être perdus, réordonnés ou même dupliqués. Les noeuds convergeront tant que la couche réseau garantit que les noeuds seront capables de transmettre leur état aux autres à terme. Il s'agit là de la principale force des CRDTs synchronisés par états.

Néanmoins, la définition de la fonction `merge` offrant ces propriétés peut s'avérer complexe et a des répercussions sur la spécification même du CRDT. Notamment, les états doivent conserver une trace de l'existence des éléments et de leur suppression afin d'éviter qu'une fusion d'états ne les fassent ressurgir. Ainsi, les CRDTs synchronisés par états utilisent régulièrement des pierres tombales.

En plus de l'utilisation de pierres tombales, la taille de l'état peut croître de manière non-bornée dans le cas de certains types de données, e.g. l'Ensemble ou la Séquence. Ainsi, ces structures peuvent atteindre à terme des tailles conséquentes. Dans de tels cas, diffuser l'état complet à chaque modification induirait alors un coût rédhibitoire. L'approche de la synchronisation par états s'avère donc inadaptée aux systèmes temps réel et repose généralement sur une synchronisation périodique.

Nous illustrons le fonctionnement de cette approche avec la Figure 1.7. Dans cet exemple, après que les noeuds aient effectués leurs modifications respectives, le mécanisme de synchronisation périodique de chaque noeud se déclenche. Le noeud A (resp. B) diffuse alors son état $\{a, b, c, e\}$ (resp. $\{a, d, e\}$) à B (resp. A).

À la réception de l'état, chaque noeud utilise la fonction `merge` pour intégrer les modifications de l'état reçu dans son propre état. Dans le cadre de l'Ensemble répliqué, cette fonction consiste généralement à faire l'union des états, en prenant en compte l'estampille et le statut (présent ou non) associé à chaque élément. Ainsi la fusion de leur état respectif, $\{a, b, c, e\} \cup \{a, d, e\}$, permet aux noeuds de converger à l'état souhaité : $\{a, b, c, d, e\}$.

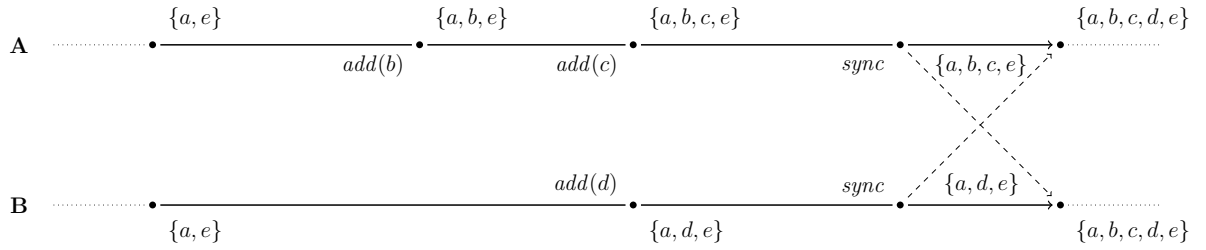


FIGURE 1.7 – Synchronisation des noeuds A et B en adoptant le modèle de synchronisation par états

Avant de conclure, il est intéressant de noter que ce modèle de synchronisation propose par nature une diffusion des modifications suivant le modèle de cohérence causale *Mathieu: TODO : Ajouter ref.* En effet, ce modèle de synchronisation assure une livraison soit de toutes les modifications connues d'un noeud, soit d'aucune. Par exemple, dans la Figure 1.7, le noeud B ne peut pas recevoir et intégrer l'élément c sans l'élément b . Ainsi, ce modèle permet naturellement d'éviter ce qui pourrait être interprétées comme des anomalies par les utilisateur-rices.

Synchronisation par opérations

L'approche de la synchronisation par opérations propose quant à elle que les noeuds diffusent leurs modifications sous la forme d'opérations. Pour chaque modification possible, les CRDTs synchronisés par opérations doivent définir deux fonctions : **prepare** et **effect**.

La fonction **prepare** a pour but de générer une opération correspondant à la modification effectuée, et commutative avec les potentielles opérations concurrentes. Elle prend en paramètres la modification ainsi que ses paramètres, et l'état courant du noeud. Cette fonction n'a pas d'effet de bord, c.-à-d. ne modifie pas l'état courant, et génère en retour l'opération à diffuser à l'ensemble des noeuds.

Une opération est un message. Son rôle est d'encoder la modification sous la forme d'un ou plusieurs éléments irréductibles du sup-demi-treillis.

Définition 3 (Élément irréductible) *Un élément irréductible d'un sup-demi-treillis est un élément atomique de ce dernier. Il ne peut être obtenu par la fusion d'autres états.*

Il est à noter que dans le cas des CRDTs purs synchronisés par opérations [8], les modifications labellisées avec leur information de causalité correspondent à des éléments irréductibles, c.-à-d. à des opérations. La fonction **prepare** peut donc être omise pour cette sous-catégorie de CRDTs synchronisés par opérations.

La fonction **effect** permet quant à elle d'intégrer les effets d'une opération générée ou reçue. Elle prend en paramètre l'état courant et l'opération, et retourne un nouvel état. Ce nouvel état correspond à la LUB entre l'état courant et le ou les éléments irréductibles encodés par l'opération.

La diffusion des modifications par le biais d'opérations présentent plusieurs avantages. Tout d'abord, la taille des opérations est généralement fixe et inférieure à la taille de

l'état complet du CRDT, puisque les opérations servent à encoder un de ses éléments irréductibles. Ensuite, l'expressivité des opérations permet de proposer plus simplement des algorithmes efficaces pour leur intégration par rapport aux modifications équivalentes dans les CRDTs synchronisés par états. Par exemple, la suppression d'un élément dans un Ensemble se traduit en une opération de manière presque littérale, tandis que pour les CRDTs synchronisés par états, c'est l'absence de l'élément dans l'état qui va rendre compte de la suppression effectuée. Ces avantages rendent possible la diffusion et l'intégration une à une des modifications et rendent ainsi utilisables les CRDTs synchronisés par opérations pour construire des systèmes temps réels.

Il est à noter que la seule contrainte imposée aux CRDTs synchronisés par opérations est que leurs opérations concurrentes soient commutatives. Ainsi, il n'existe aucune contrainte sur la commutativité des opérations liées causalement. De la même manière, aucune contrainte n'est définie sur l'idempotence des opérations. Ces libertés impliquent qu'il peut être nécessaire que les opérations soient délivrées au CRDT en respectant un ordre donné et en garantissant leur livraison en exactement une fois pour garantir la convergence. Ainsi, un intergiciel chargé de la diffusion et de la livraison des opérations est usuellement associé aux CRDTs synchronisés par opérations pour respecter ces contraintes.

Généralement, les CRDTs synchronisés par opérations sont présentés dans la littérature comme nécessitant une livraison causale des opérations. Ce modèle de livraison permet de respecter le modèle de cohérence causale et ainsi de simplifier raisonnement sur exécutions.

Ce modèle introduit néanmoins plusieurs effets négatifs. Tout d'abord, ce modèle peut provoquer un délai dans la diffusion des modifications. En effet, la perte d'une opération par le réseau provoque la mise en attente de la livraison des opérations suivantes. Les opérations mises en attente ne sont délivrées qu'une fois l'opération perdue re-diffusée et délivrée.

De plus, il nécessite que des informations de causalité précises soient attachées à chaque opération. Pour cela, les systèmes reposent généralement sur l'utilisation de vecteurs de version *Matthieu: TODO : Insérer ref.* Or, la taille de cette structure de données croît de manière linéaire avec le nombre de noeuds du système. Les métadonnées de causalité peuvent ainsi représenter la majorité des données diffusées sur le réseau⁴. *Matthieu: TODO : Ajouter mention que OT a été abandonné à cause de cette contrainte même.* Cependant, nous observons que la livraison dans l'ordre causal de toutes les opérations n'est pas toujours nécessaire. Par exemple, l'ordre d'intégration de deux opérations d'ajout d'éléments différents dans un Ensemble n'a pas d'importance. Nous pouvons alors nous affranchir de la livraison dans l'ordre causal pour accélérer la vitesse d'intégration des modifications et pour réduire les métadonnées envoyées.

Une autre contrainte généralement associée aux CRDTs synchronisés par opérations est la nécessité d'une livraison en exactement un exemplaire de chaque opération. Cette contrainte dérive de :

4. La relation de causalité étant transitive, les opérations et leurs relations de causalité forment un DAG. [61] propose d'ajouter en dépendances causales d'une opération seulement les opérations correspondant aux extrémités du DAG au moment de sa génération. Ce mécanisme plus complexe permet de réduire la consommation réseau, mais induit un surcoût en calculs et en mémoire utilisée.

- (i) La potentielle non-idempotence des opérations.
- (ii) La nécessité de la livraison de chaque opération pour la livraison causale.

Toutefois, nous observons que des opérations peuvent être sémantiquement rendues obsolètes par d'autres opérations, e.g. une opération d'ajout d'un élément dans un Ensemble est rendue obsolète par une opération de suppression ultérieure du même élément. Ainsi, l'intergiciel de livraison peut se contenter d'assurer une livraison en un exemplaire au plus des opérations non-obsolètes. Ce choix permet de réduire la consommation réseau en évitant la diffusion d'opérations désormais non-pertinentes.

Pour compenser la perte d'opérations par le réseau et ainsi garantir la livraison à terme des opérations pertinentes, l'intergiciel de livraison des opérations doit mettre en place un mécanisme d'anti-entropie. Plusieurs mécanismes de ce type ont été proposés dans la littérature [58, 21, 40, 28] *Matthieu: TODO : Ajouter refs Scuttlebutt si applicable à Op-based* et proposent des compromis variés entre complexité en temps, complexité spatiale et consommation réseau.

Nous illustrons le modèle de synchronisation par opérations à l'aide de la Figure 1.8. Dans ce nouvel exemple, les noeuds diffusent les modifications qu'ils effectuent sous la forme d'opérations. Nous considérons que le CRDT utilisé est un CRDT pur synchronisé par opérations, c.-à-d. que les modifications et opérations sont confondues.

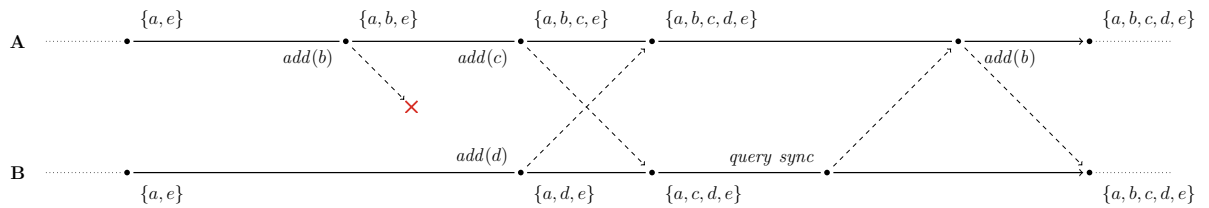


FIGURE 1.8 – Synchronisation des noeuds A et B en adoptant le modèle de synchronisation par opérations

Le noeud A diffuse donc les opérations $add(b)$ et $add(c)$. Il reçoit ensuite l'opération $add(d)$ de B, qu'il intègre à sa copie. Il obtient alors l'état $\{a, b, c, d, e\}$.

De son côté, le noeud B ne reçoit initialement pas l'opération $add(b)$ suite à une défaillance réseau. Il génère et diffuse $add(d)$ puis reçoit l'opération $add(c)$. Nous considérons que la livraisons des opérations add n'est pas obligatoire dans cet exemple, cette opération est alors intégrée. Le noeud B obtient alors l'état $\{a, c, d, e\}$.

Ensuite, le mécanisme d'anti-entropie du noeud B se déclenche. Le noeud B envoie alors à A une demande de synchronisation contenant un résumé de son état, e.g. son vecteur de version. À partir de cette donnée, le noeud A détermine que B n'a pas reçu l'opération $add(a)$. Il génère alors une réponse contenant cette opération et lui envoie. À la réception de l'opération, le noeud B l'intègre. Il obtient l'état $\{a, b, c, d, e\}$ et converge ainsi avec A.

Avant de conclure, nous noterons qu'il est nécessaire pour les noeuds de maintenir leur journal d'opérations. En effet, les noeuds l'utilisent pour renvoyer les opérations manquées lors de l'exécution du mécanisme d'anti-entropie évoqué ci-dessus. Ceci se traduit par une augmentation perpétuelle des métadonnées des CRDTs synchronisés par opérations. Pour

y pallier, des travaux [8, 11] proposent de tronquer le journal des opérations pour en supprimer les opérations connues de tous. Les noeuds reposent alors sur la notion de stabilité causale pour déterminer les opérations supprimables de manière sûre.

Définition 4 (Stabilité causale) *Une opération est stable causalement lorsqu'elle a été observée par l'ensemble des noeuds du système. Ainsi, toute opération future dépend causalement des opérations causalement stables, c.-à-d. les noeuds ne peuvent plus générer d'opérations concurrentes aux opérations causalement stables de manière honnête .*

Un mécanisme d'instantané *Matthieu: TODO : Ajouter refs* doit néanmoins être associé au mécanisme de troncage du journal pour générer un état équivalent à celui résultant des opérations supprimées. Ce mécanisme est en effet nécessaire pour permettre un nouveau noeud de rejoindre le système et d'obtenir l'état courant à partir de l'instantané et du journal tronqué.

Pour résumer, cette approche permet de mettre en place simplement un système fonctionnel à l'aide d'un CRDT synchronisé par opérations et d'un intergiciel de diffusion et de livraison RCB. Mais comme illustré ci-dessus, chaque CRDT synchronisé par opérations établit les propriétés de ses différentes opérations et délègue potentiellement des responsabilités à l'intergiciel de diffusion et livraison. La complexité de cette approche réside ainsi dans l'ajustement du couple $\langle CRDT, intergiciel \rangle$ pour régler finement et optimiser leur fonctionnement en tandem. Des approches [8, 11] ont été proposées ces dernières années pour concevoir et structurer plus proprement ces composants et leurs relations, mais reposent encore sur une livraison causale des opérations. *Matthieu: TODO : Vérifier que c'est bien le cas dans [11]*

Synchronisation par différences d'états

Dans [4], ALMEIDA, SHOKER et BAQUERO introduisent un nouveau modèle de synchronisation pour CRDTs. La proposition de ce modèle est nourrie par les observations suivantes :

- (i) Les CRDTs synchronisés par opérations sont vulnérables aux défaillances du réseau et nécessitent généralement pour pallier à cette vulnérabilité une livraison des opérations en exactement un exemplaire et respectant l'ordre causal.
- (ii) Les CRDTs synchronisés par états pâtissent du surcoût induit par la diffusion de leurs états complets, généralement croissant continuellement.

Pour pallier aux faiblesses de chaque approche et allier le meilleur des deux mondes, les auteurs proposent les CRDTs synchronisés par différences d'états [4, 3, 24]. Il s'agit en fait d'une sous-famille des CRDTs synchronisés par états. Ainsi, comme ces derniers, ils disposent d'une fonction `merge` permettant de produire la LUB entre deux états, et qui est associative, commutative et idempotente.

La spécificité des CRDTs synchronisés par différences d'états est qu'une modification locale produit en retour un delta. Un delta encode la modification effectuée sous la forme d'un état du lattice. Les deltas étant des états, ils peuvent être diffusés puis intégrés par les autres noeuds à l'aide de la fonction `merge`. Ceci permet de bénéficier des propriétés

d'associativité, de commutativité et d'idempotence offertes par cette fonction. Les CRDTs synchronisés par différences d'états offrent ainsi :

- (i) Une diffusion des modifications avec un surcoût pour le réseau proche de celui des CRDTs synchronisés par opérations.
- (ii) Une résistance aux défaillances réseaux similaire celle des CRDTs synchronisés par états.

Cette définition des CRDTs synchronisés par différences d'états, introduite dans [4, 3], fut ensuite précisée dans [24]. Dans cet article, ENES et al. précisent qu'utiliser des éléments irréductibles (cf. Définition 3) comme deltas est optimal du point de vue de la taille des deltas produits.

Concernant la diffusion des modifications, les CRDTs synchronisés par différences d'états autorisent un large éventail de possibilités. Par exemple, les deltas peuvent être diffusés et intégrés de manière indépendante. Une autre approche possible consiste à tirer avantage du fait que les deltas sont des états : il est possible d'agréger plusieurs deltas à l'aide la fonction `merge`, éliminant leurs éventuelles redondances. Ainsi, la fusion de deltas permet ensuite de diffuser un ensemble de modifications par le biais d'un seul et unique delta, minimal. Et en dernier recours, les CRDTs synchronisés par différences d'états peuvent adopter le même schéma de diffusion que les CRDTs synchronisés par états, c.-à-d. diffuser leur état complet de manière périodique. Plusieurs de ces approches sont évaluées et comparées de manière expérimentale dans [24].

Nous illustrons cette approche avec la Figure 1.9. Dans cet exemple, nous considérons que les noeuds adoptent la seconde approche évoquée, c.-à-d. que périodiquement les noeuds agrègent les deltas issus de leurs modifications et diffusent le delta résultant.

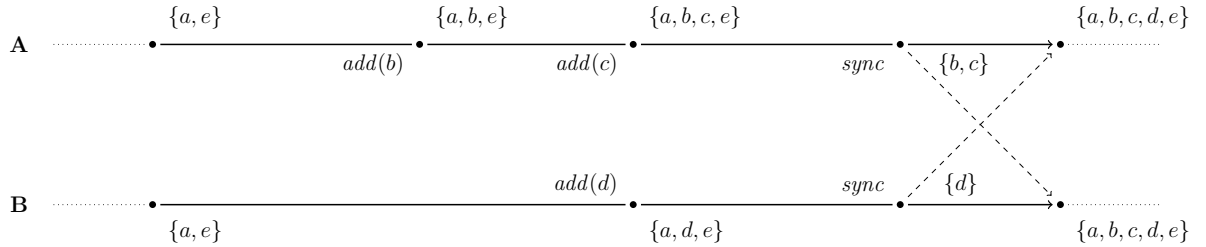


FIGURE 1.9 – Synchronisation des noeuds A et B en adoptant le modèle de synchronisation par différences d'états

Le noeud A effectue les modifications $add(b)$ et $add(c)$, qui retournent respectivement les deltas $\{b\}$ et $\{c\}$. Le noeud A agrège ces deltas et diffuse donc le delta suivant b, c . Quant au noeud B, il effectue la modification $add(d)$ qui produit le delta $\{d\}$. S'agissant de son unique modification, il diffuse ce delta inchangé.

Quand A (resp. B) reçoit le delta $\{d\}$ (resp. $\{b, c\}$), il l'intègre à sa copie en utilisant la fonction `merge`. Les deux noeuds convergent alors à l'état $\{a, b, c, d, e\}$.

La synchronisation par différences d'états permet donc de réduire la taille des messages diffusés sur le réseau par rapport à la synchronisation par états. Cependant, il est important de noter que la décomposition en deltas entraîne la perte d'une des propriétés

intrinsèques des CRDTs synchronisés par états : le respect du modèle de cohérence causale. En effet, sans mécanisme supplémentaire, la perte ou le ré-ordonnement de deltas par le réseau peut mener à une livraison dans le désordre des modifications à l'un des noeuds. S'ils souhaitent assurer une diffusion causale des modifications, les CRDTs synchronisés par différences d'états doivent donc définir et intégrer dans leur spécification un mécanisme similaire à l'intergiciel de livraison des CRDTs synchronisés par opérations. En revanche et à l'instar des CRDTs synchronisés par opérations, les CRDTs synchronisés par différences d'états peuvent aussi faire le choix inverse : s'affranchir du modèle de cohérence causale pour accélérer la diffusion des modifications et minimiser le surcoût du type répliqué.

Ainsi, les CRDTs synchronisés par différences d'états sont une évolution prometteuse des CRDTs synchronisés par états. Ce modèle de synchronisation rend ces CRDTs utilisables dans les systèmes temps réels sans introduire de contraintes sur la fiabilité du réseau. Mais pour cela, il ajoute une couche supplémentaire de complexité à la spécification des CRDTs synchronisés par états, c.-à-d. le mécanisme dédié à la livraison des deltas, qui peut s'avérer source d'erreurs et de coûts supplémentaires.

Synthèse

Nous récapitulons les principales propriétés et différences des modèles de synchronisations pour CRDTs dans Tableau 1.1.

TABLE 1.1 – Récapitulatif comparatif des différents modèles de synchronisation pour CRDTs

	State-based	Op-based	Delta-based
États forment sup-demi-treillis	✓	✓	✓
Intègre modifications par fusion d'états	✓	✗	✓
Intègre modifications de manière atomique	✗	✓	✓
Résistant aux défaillances réseau	✓	✗	✓
Peut s'affranchir de la cohérence causale	✗	✓	✓
Adapté pour systèmes temps réel	✗	✓	✓

1.3.3 Adoption dans la littérature et l'industrie

- Proposition et conception de CRDTs pour une variété de types de données : Registre, Compteur *Matthieu: TODO : Ajouter IPA*, Ensemble, Liste/Sequence, Graphe, JSON, Filesystem, Access Control. Propose généralement plusieurs sémantiques de résolution de conflits par type de données.
- Conception et développement de bibliothèques mettant à disposition des développeurs d'applications des types de données composés [52, 51, 77, 38, 7] *Matthieu: TODO : Revoir et ajouter Melda (PaPoC'22) si fitting*
- Conception de langages de programmation intégrant des CRDTs comme types primitifs, destinés au développement d'applications distribuées [43, 20]

- Conception et implémentation de bases de données distribuées, relationnelles ou non, privilégiant la disponibilité et la minimisation de la latence à l'aide des CRDTs [64, 16, 75, 15, 79] *Matthieu: TODO : Ajouter Redis et Akka*
- Conception d'un nouveau paradigme d'applications, Local-First Software, dont une des fondations est les CRDTs [39, 32] *Matthieu: TODO : Vérifier et ajouter l'article avec Digital Garden (PaPoC'22 ?) si fitting*
- Éditeurs collaboratifs temps réel à large échelle et offrant de nouveaux scénarios de collaboration grâce aux CRDTs [45, 55]

1.4 Séquences répliquées sans conflits

- Présentation du type Sequence et de son utilisation
- Spécification du type : données, modifications
- Présentation de conflits
 - dOpt puzzle
- Introduction des principales approches pour Sequence CRDTs

1.4.1 Approche à pierres tombales

- WOOT [57, 74, 2]
- RGA [65]
- RGASplit [12]

1.4.2 Approche à identifiants densément ordonnés

- Treedoc [62]
- Logoot [72, 73]

Matthieu: NOTE : Ajouter LogootSplit de manière sommaire aussi à cet endroit ?

Matthieu: TODO : Autres Sequence CRDTs à considérer : String-wise CRDT [78], Chronofold [31]

1.4.3 Synthèse

- Deux approches différentes pour la résolution de conflits ont été proposées pour CRDTs pour Séquence. Chacune de ces approches visent à minimiser surcoût du type répliqué, que ce soit d'un point de vue mémoire, computations et réseau.
- Au fil des années, ces approches ont été raffinées avec de nouveaux CRDTs de plus en plus en efficace.
- Néanmoins, malgré les évaluations et comparaisons, la littérature n'a pas établi une supériorité d'une approche sur l'autre. Les approches proposent seulement des compromis différents sur la nature du surcoût, que nous récapitulons dans Tableau 1.2.

L'approche basée sur pierres tombales offre une consommation réseau constante grâce à ses identifiants de taille fixe, mais souffre d'une consommation mémoire ne pouvant qu'augmenter. L'approche basée sur identifiants densément ordonnés bénéficie d'un meilleur délai de diffusion des modifications, les modifications pouvant être livrées dans le désordre, mais souffre d'une empreinte réseau augmentant avec la taille de ses identifiants.

TABLE 1.2 – Récapitulatif comparatif des différents approches pour CRDTs pour Séquence

	Dense ids-based	Tombstoned-based
Identifiants de taille fixe	✗	✓
Éléments réellement supprimés	✓	✗
Empreinte réseau fixe	✗	✓
Peut s'affranchir de la cohérence causale	✓	✗

- Pour la suite de ce manuscrit, nous prenons LogootSplit comme base de travail. Nous détaillons donc son fonctionnement dans la section suivante.

1.5 LogootSplit

LogootSplit [6] est l'état de l'art des séquences répliquées à identifiants densément ordonnés. Comme expliqué précédemment, LogootSplit utilise des identifiants provenant d'un ordre total dense pour positionner les éléments dans la séquence répliquée.

1.5.1 Identifiants

Pour ce faire, LogootSplit assigne des identifiants composés d'une liste de tuples aux éléments. Les tuples sont définis de la manière suivante :

Définition 5 (Tuple) *Un Tuple est un quadruplet $\langle position, nodeId, nodeSeq, offset \rangle$ où*

- *position incarne la position souhaitée de l'élément.*
- *nodeId est l'identifiant unique du noeud qui a généré le tuple.*
- *nodeSeq est le numéro de séquence courant du noeud à la génération du tuple.*
- *offset indique la position de l'élément au sein d'un bloc. Nous reviendrons plus en détails sur ce composant dans la sous-section 1.5.2.*

Matthieu: TODO : Ajouter une relation d'ordre sur les tuples

Dans ce manuscrit, nous représentons les tuples par le biais de la notation suivante : $position^{nodeId}_{nodeSeq}_{offset}$ où *position* est une lettre minuscule, *nodeId* une lettre majuscule et *nodeSeq* et *offset* des entiers, e.g. i_0^{B0} .

À partir de là, les identifiants LogootSplit sont définis de la manière suivante :

Définition 6 (Identifiant) *Un Identifiant est une liste de Tuples.*

Matthieu: TODO : Définir la notion de base (et autres fonctions utiles sur les identifiants ? genre isPrefix, concat, getTail...)

Nous représentons les identifiants en listant les tuples qui les composent. Par exemple, l'identifiant composé des tuples $\langle i, B, 0, 0 \rangle \langle f, A, 0, 0 \rangle$ est présenté de la manière suivante : $i_0^{B0} f_0^{A0}$.

Les identifiants ont pour rôle d'ordonner les éléments relativement les uns par rapport aux autres. Pour ce faire, une relation d'ordre total aux identifiants est associée à l'ensemble des identifiants :

Définition 7 (Relation $<_{id}$) *La relation $<_{id}$ est un ordre strict total sur l'ensemble des identifiants. Elle permet aux noeuds de comparer n'importe quelle paire d'identifiants. Elle est définie en utilisant l'ordre lexicographique sur les composants des différents tuples des identifiants comparés.*

- En utilisant cette relation d'ordre, les noeuds peuvent ordonner les éléments grâce à leur identifiant.
- Par exemple, déterminent que $i_0^{A1} <_{id} i_0^{B0}$ car les positions sont identiques et que le *nodeId* (A) du premier est plus petit que le *nodeId* (B) du second
- et que $i_0^{B0} <_{id} i_0^{B0} f_0^{A0}$ car le premier est un préfixe du second

Matthieu: TODO : Montrer que cet ensemble d'identifiants est un ensemble dense

1.5.2 Aggrégation dynamique d'éléments en blocs

Au lieu de stocker les identifiants de chaque élément de la séquence, LogootSplit propose d'aggréger de façon dynamique les éléments dans des blocs. Pour cela, LogootSplit introduit la notion d'intervalle d'identifiants :

Définition 8 (IdInterval) *Un IdInterval est un couple $\langle idBegin, offsetEnd \rangle$ où*

- *idBegin est l'identifiant du premier élément de l'intervalle.*
- *offsetEnd est l'offset du dernier identifiant de l'intervalle.*

Les intervalles d'identifiants permettent à LogootSplit d'assigner logiquement un identifiant à un ensemble d'éléments, tout en ne stockant réellement que l'identifiant de son premier élément et le dernier offset de son dernier élément.

LogootSplit regroupe les éléments avec des identifiants *contigus* dans un interval. Nous appelons *contigus* deux identifiants qui partagent une même base (c.-à-d. qui sont identiques à l'exception de leur dernier offset) et dont les *offsets* sont consécutifs. Nous représentons un intervalle d'identifiants à l'aide du formalisme suivant : $position_{begin..end}^{nodeId\ nodeSeq}$ où *begin* est l'offset du premier identifiant de l'intervalle et *end* du dernier.

Les blocs permettent d'associer un intervalle d'identifiants aux éléments correspondant. Les blocs sont définis de la manière suivante :

Définition 9 (Bloc) *Un Bloc est un quadruplet $\langle idInterval, elts, isAppendable, isPrependable \rangle$ où*

- *idInterval* est l'intervalle d'identifiants formant le bloc
- *elts* sont les éléments contenus dans le bloc
- *isAppendable* (resp. *isPrependable*) est un booléen indiquant si l'auteur du bloc peut ajouter un nouvel élément en fin (resp. début) de bloc

La Figure 1.10 présente un exemple de séquence LogootSplit : dans la Figure 1.10a, les identifiants i_0^{B0} , i_1^{B0} , i_2^{B0} forment une chaîne d'identifiants contigus. LogootSplit est donc capable de regrouper ces éléments en un bloc représentant l'intervalle d'identifiants $i_{0..2}^{B0}$ pour minimiser les métadonnées stockées, comme montré dans la Figure 1.10b.

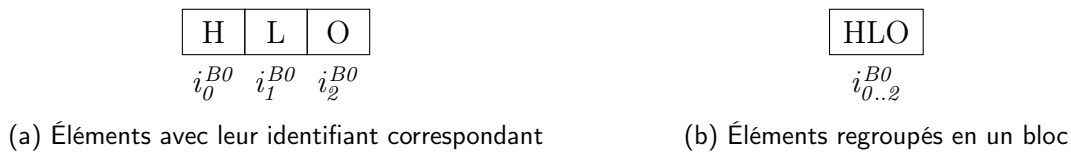


FIGURE 1.10 – Représentation d'une séquence LogootSplit contenant les éléments "HLO"

Cette fonctionnalité réduit le nombre d'identifiants stockés au sein de la structure de données, puisque les identifiants sont conservés à l'échelle des blocs plutôt qu'à l'échelle de chaque élément. Ceci permet de réduire de manière significative le surcoût en métadonnées de la structure de données. L'utilisation de blocs améliore aussi les performances de la structure de données. En effet, l'utilisation de blocs permet de parcourir plus efficacement la structure de données. Les blocs permettent aussi d'effectuer des modifications à l'échelle de la chaîne de caractères et non plus seulement caractère par caractère.

Matthieu: TODO : indiquer que le couple $\langle \text{nodeId}, \text{nodeSeq} \rangle$ permet d'identifier de manière unique la base d'un bloc ou d'un identifiant

Notons que pour une séquence donnée, nous pouvons identifier chacun de ses identifiants par le triplet $\langle \text{nodeId}, \text{nodeSeq}, \text{offset} \rangle$ issue de leur dernier Tuple. Par exemple, le triplet $\langle B, 0, 2 \rangle$ désigne de manière unique l'identifiant i_2^{B0} dans Figure 1.10.

1.5.3 Modèle de données

ANDRÉ et al. [6] définissent une séquence LogootSplit de la manière suivante :

Définition 10 (Séquence LogootSplit) Une séquence Séquence LogootSplit est un triplet $\langle \text{nodeId}, \text{nodeSeq}, \text{blocks} \rangle$ où

- *nodeId* est l'identifiant du noeud.
- *nodeSeq* est le numéro de séquence courant du noeud.
- *blocks* est une liste de Blocs correspondant à l'état actuel de la séquence répliquée.

Plusieurs fonctions sont définies sur cette structure de données et permettent de l'interroger et de la modifier :

- $\text{ins}(S, \text{index}, \text{elts})$ permet d'insérer les éléments elts à la position index dans la séquence S . Cette fonction génère et associe un intervalle d'identifiants valide aux éléments insérés. Elle retourne une opération *insert* permettant aux autres noeuds d'intégrer la modification à leur état.

Définition 11 (insert) Une opération *insert* est un couple $\langle id, \text{elts} \rangle$ où

- id est l'identifiant du premier élément inséré par cette opération.
- elts est la liste des éléments insérés par cette opération.
- $\text{rem}(S, \text{index}, \text{length})$ permet de supprimer length éléments à partir la position index dans la séquence S . Cette fonction répertorie les éléments supprimés sous la forme d'intervalles d'identifiants. Elle retourne une opération *remove* permettant aux autres noeuds d'intégrer la modification à leur état.

Définition 12 (remove) Une opération *remove* est une liste d'intervalles d'identifiants où chaque intervalle désigne un ensemble d'éléments à supprimer.

Nous présentons dans la Figure 1.11 un exemple d'utilisation de cette séquence répliquée.

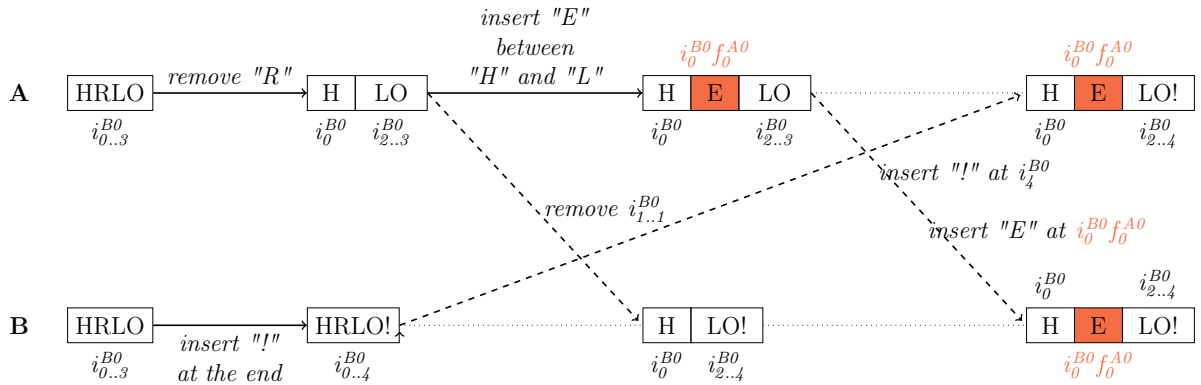


FIGURE 1.11 – Modifications concurrentes d'une séquence répliquée LogootSplit

Dans cet exemple, deux noeuds A et B répliquent et éditent collaborativement un document texte en utilisant LogootSplit. Ils partagent initialement le même état : une séquence composée d'un seul bloc associant les identifiants $i_{0..3}^{B0}$ aux éléments "HRLO". Les noeuds se mettent ensuite à éditer le document.

Le noeud A commence par supprimer l'élément "R" de la séquence. LogootSplit génère l'opération *remove* correspondante en utilisant l'identifiant de l'élément supprimé (i_1^{B0}). Cette opération est envoyée au noeud B pour qu'il intègre cette modification.

Le noeud A insère ensuite un élément "E" dans la séquence, entre le "H" et le "L". LogootSplit doit alors générer un identifiant id à associer à ce nouvel élément. Ce nouvel identifiant id doit respecter la contrainte suivante : $i_0^{B0} <_{id} id <_{id} i_2^{B0}$. Cependant, LogootSplit ne peut pas générer un identifiant composé d'un seul tuple respectant cet ordre. LogootSplit génère alors id en recopiant le premier tuple (i_0^{B0}) et en y ajoutant un nouveau

tuple (f_0^{A0}) . LogootSplit génère l'opération *insert* correspondante, indiquant l'élément à insérer et sa position grâce à son identifiant. Cette opération est ensuite diffusée sur le réseau.

En parallèle, le noeud B insère un élément "!" à la fin de la séquence. Comme le noeud B est l'auteur du bloc $i_{0..3}^{B0}$, il peut y ajouter de nouveaux éléments. LogootSplit associe donc l'identifiant i_4^{B0} à l'élément "!" et l'ajoute au bloc existant.

Les noeuds se synchronisent ensuite. Le noeud A reçoit l'opération *insert* de l'élément "!" à la position i_4^{B0} . Le noeud A détermine que cet élément doit être inséré à la fin de la séquence (puisque $i_3^{B0} <_{id} i_4^{B0}$) et qu'il peut être ajouté au bloc $i_{2..3}^{B0}$ (puisque i_3^{B0} et i_4^{B0} sont contigus).

De son côté, le noeud B reçoit tout d'abord l'opération *remove* des éléments identifiés par l'intervalle $i_{1..1}^{B0}$, c.-à-d. l'élément attaché à l'identifiant i_1^{B0} . Le noeud B supprime donc l'élément "R" de son état.

Il reçoit ensuite l'opération *insert* de l'élément "E" à la position $i_0^{B0}f_0^{A0}$. Le noeud B insère cet élément entre les éléments "H" et "L" (puisque $i_0^{B0} <_{id} i_0^{B0}f_0^{A0} <_{id} i_2^{B0}$), respectant ainsi l'intention du noeud A.

Matthieu: NOTE : Pourrait définir dans cette sous-section la notion de séquence bien-formée

1.5.4 Modèle de livraison

Afin de garantir son bon fonctionnement, LogootSplit doit être associé à une couche de livraison de messages garantissant plusieurs propriétés.

Livraison des opérations en exactement un exemplaire

Tout d'abord, la couche de livraison de messages doit assurer que toutes les opérations soient délivrées aux noeuds, mais qu'une seule et unique fois. La Figure 1.12 représente un exemple illustrant la nécessité de cette contrainte.

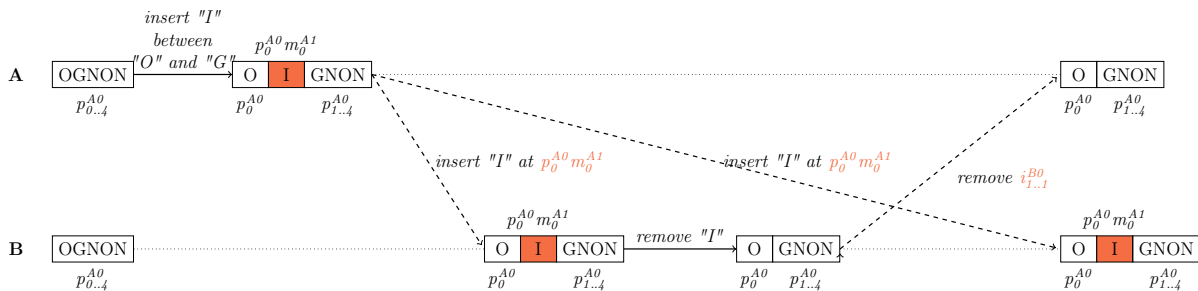


FIGURE 1.12 – Résurgence d'un élément supprimé suite à la relivraison de son opération *insert*

Dans cet exemple, deux noeuds A et B répliquent et éditent collaborativement une séquence. La séquence répliquée contient initialement les éléments "OGNON", qui sont associés à l'intervalle d'identifiants $p_{0..4}^{A0}$.

Le noeud A commence par insérer un nouvel élément, "I", dans la séquence entre les éléments "O" et "G". L'opération *insert* résultante, insérant l'élément "I" à la position $p_0^{A0} m_0^{A1}$, est diffusée au noeud B.

À la réception de l'opération *insert*, le noeud B l'intègre à son état. Puis il supprime dans la foulée ce nouvel élément. L'opération *remove* générée est envoyée au noeud A.

Le noeud A intègre l'opération *remove*, ce qui a pour effet de supprimer l'élément "I" associé à l'identifiant $p_0^{A0} m_0^{A1}$. Il obtient alors un état équivalent à celui du noeud B.

Cependant, l'opération *insert* insérant l'élément "I" à la position $p_0^{A0} m_0^{A1}$ est de nouveau délivrée au noeud B. De multiples raisons peuvent être à l'origine de cette nouvelle livraison : perte du message d'*acknowledgment*, utilisation d'un protocole de diffusion épidémique des messages, déclenchement du mécanisme d'anti-entropie en concurrence... Le noeud B ré-intègre alors l'opération *insert*, ce qui fait revenir l'élément "I" et l'identifiant associé. L'état du noeud B diverge désormais de celui-ci du noeud A.

Pour se prémunir de ce type de scénarios, LogootSplit requiert que la couche de livraison des messages assure une livraison en exactement un exemplaire des opérations. Cette contrainte permet d'éviter que d'anciens éléments et identifiants ressurgissent après leur suppression chez certains noeuds uniquement à cause d'une livraison multiple de l'opération *insert* correspondante.

Matthieu: QUESTION : Ajouter quelques lignes ici sur comment faire ça en pratique (Ajout d'un dot aux opérations, maintien d'un dot store au niveau de la couche livraison, vérification que dot pas encore présent dans dot store avant de passer opération à la structure de données) ? Ou je garde ça pour le chapitre sur MUTE ?

Livraison de l'opération *remove* après l'opération *insert*

Une autre propriété que doit assurer la couche de livraison de messages est que les opérations *remove* doivent être livrées au CRDT après les opérations *insert* correspondantes. La Figure 1.13 présente un exemple justifiant cette contrainte.

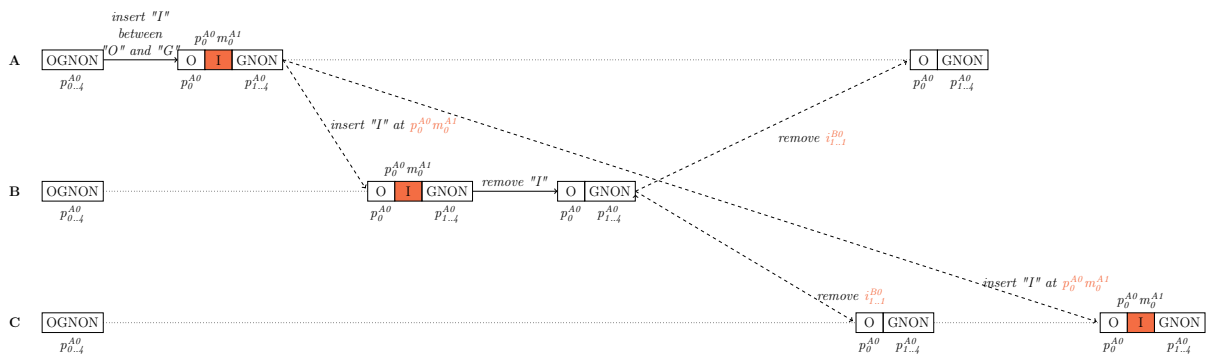


FIGURE 1.13 – Non-effet de l'opération *remove* car reçue avant l'opération *insert* correspondante

Dans cet exemple, trois noeuds A, B et C répliquent et éditent collaborativement une séquence. Le noeud A commence par insérer un nouvel élément, "I", dans la séquence entre les éléments "O" et "G". L'opération *insert* résultante, insérant l'élément "I" à la position $p_0^{A0} m_0^{A1}$, est diffusée aux autres noeuds.

À la réception de l'opération *insert*, le noeud B l'intègre à son état. Cependant, le noeud B supprime dans la foulée l'élément nouvellement ajouté. Il diffuse ensuite l'opération *remove* générée.

Toutefois, suite à un aléa du réseau, l'opération *remove* supprimant l'élément "I" est livrée au noeud C avant l'opération *insert* l'ajoutant à son état. Lorsque le noeud C reçoit l'opération *remove*, il parcourt son état à la recherche de l'élément "I" pour le supprimer. Cependant, celui-ci n'est pas présent dans son état courant. L'intégration de l'opération s'achève donc sans effectuer de modification.

Le noeud C reçoit ensuite l'opération *insert*. Le noeud C intègre ce nouvel élément dans la séquence en utilisant son identifiant ($p_0^{A0} <_{id} p_0^{A0} m_0^{A1} <_{id} p_1^{A0}$).

Ainsi, l'état du noeud C diverge de celui-ci des autres noeuds à terme, et cela malgré que les noeuds A, B et C aient intégré le même ensemble d'opérations. Ce résultat transgresse la propriété de Cohérence forte à terme (SEC) que doivent assurer les CRDTs. Afin d'empêcher ce scénario de se produire, LogootSplit impose donc la livraison causale des opérations *remove* par rapport aux opérations *insert* correspondantes.

Matthieu: QUESTION : Même que pour la exactly-once delivery, est-ce que j'explique ici comment assurer cette contrainte plus en détails (Ajout des dots des opérations insert en dépendances de l'opération remove, vérification que dots présents dans dot store avant de passer l'opération remove à la structure de données) ou je garde ça pour le chapitre sur MUTE ?

Définition du modèle de livraison

Pour résumer, la couche de livraison des opérations associée à LogootSplit doit respecter le modèle de livraison suivant :

Définition 13 (Exactly-once + Causal remove) *Le modèle de livraison Exactly-once + Causal remove définit les 3 règles suivantes sur la livraison des opérations :*

1. Une opération doit être délivrée à l'ensemble des noeuds à terme,
2. Une opération doit être délivrée qu'une seule et unique fois aux noeuds,
3. Une opération *remove* doit être délivrée à un noeud une fois que les opérations *insert* des éléments concernés par la suppression ont été délivrées à ce dernier.

Il est à noter que ELVINGER [22] a récemment proposé dans ses travaux de thèse Dotted LogootSplit, un nouveau Sequence CRDT basée sur les différences. Inspiré de Logoot et LogootSplit, ce nouveau CRDT associe une séquence à identifiants densément ordonnés à un contexte causal. Le contexte causal est une structure de données permettant à Dotted LogootSplit de représenter et de maintenir efficacement les informations des modifications déjà intégrées à l'état courant. Cette association permet à Dotted LogootSplit de fonctionner de manière autonome, sans imposer de contraintes particulières à la couche livraison autres que la livraison à terme.

1.5.5 Limites

Comme indiqué précédemment, la taille des identifiants provenant d'un ordre total dense est variable. Quand les noeuds insèrent de nouveaux éléments entre deux autres

1.6. Mitigation du surcoût des séquences répliquées sans conflits

ayant la même valeur de *position*, LogootSplit n'a pas d'autre choix que d'augmenter la taille de l'identifiant résultant. La Figure 1.14 illustre de tels cas. Dans cet exemple, puisque le noeud A insère un nouvel élément entre deux identifiants contigus i_0^{B0} et i_1^{B0} , LogootSplit ne peut pas générer un identifiant adapté de la même taille. Pour respecter l'ordre souhaité, LogootSplit génère un identifiant en ajoutant un nouveau tuple à l'identifiant du prédecesseur : $i_0^{B0} f_0^{A0}$.

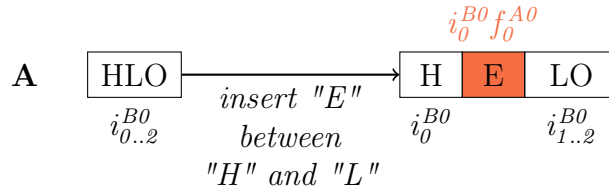


FIGURE 1.14 – Insertion menant à une augmentation de la taille des identifiants

Par conséquent, la taille des identifiants a tendance à croître alors que le système progresse. Cette croissance impacte négativement les performances de la structure de données sur plusieurs aspects. Puisque les identifiants attachés aux éléments deviennent plus long, le surcoût en métadonnées augmente. Ceci augmente aussi la consommation en bande-passante puisque les noeuds doivent diffuser les identifiants aux autres.

Matthieu: TODO : Ajouter une phrase pour expliquer que la croissance des identifiants impacte aussi le temps d'intégration des modifications

De plus, le nombre de blocs composant la séquence répliquée augmente au fil du temps. En effet, plusieurs contraintes sur la génération d'identifiants empêchent les noeuds d'ajouter des nouveaux éléments aux blocs existants. Par exemple, seul le noeud qui a généré un bloc peut ajouter un élément à ce dernier. Ces limitations provoquent la génération de nouveau blocs. La séquence se retrouve finalement fragmentée en de nombreux blocs de seulement quelques caractères chacun. Cependant, aucun mécanisme pour fusionner les blocs à posteriori n'est fourni. L'efficacité de la structure décroît donc puisque chaque bloc entraîne un surcoût.

Comme illustré plus loin, nous avons mesuré au cours de nos évaluations que le contenu représente à terme moins de 1% de taille de la structure de données. Les 99% restants correspondent aux métadonnées utilisées par la séquence répliquée. Il est donc nécessaire de proposer des mécanismes et techniques afin de mitiger les problèmes soulignés précédemment.

1.6 Mitigation du surcoût des séquences répliquées sans conflits

- Plusieurs approches ont été proposées pour réduire croissance des métadonnées dans Sequence CRDTs
- RGA (et RGASplit) propose un mécanisme de GC des pierres tombales. Nécessite cependant stabilité causale des opérations de suppression. S'agit d'une contrainte

forte, peu adaptée aux systèmes dynamiques à large échelle. *Matthieu: TODO : Trouver référence sur la stabilité causale dans systèmes dynamiques*

- Core & Nebula propose un mécanisme de ré-équilibrage de l'arbre pour Treedoc. Le ré-équilibrage a pour effet de supprimer des potentielles pierres tombales et de réduire la taille des identifiants. Repose sur un algorithme de consensus. S'agit de nouveau d'une contrainte forte pour systèmes dynamique à large échelle. Pour y pallier, propose de séparer les pairs entre deux ensembles : Core et Nebula. Permet de limiter le nombre participant au consensus. Un protocole de rattrapage permet aux noeuds de la Nebula de mettre à jour leurs modifications concurrentes à un ré-équilibrage.
- LSEQ adopte une autre approche. Part du constat que les identifiants dans Logoot croissent de manière linéaire. Vise une croissance logarithmique des identifiants. Pour cela, propose de nouvelles fonctions d'allocation des identifiants visant à maximiser le nombre d'identifiants insérés avant de devoir augmenter la taille de l'identifiant. Propose aussi d'utiliser une base exponentielle pour la valeur *position* des identifiants. Atteint ainsi la croissance polylogarithmique des identifiants, sans coordination requise entre les noeuds et mécanisme supplémentaire. Solution adaptée aux systèmes distribués à large échelle. Conjecture cependant que cette approche se marie mal avec les Sequence CRDTs utilisant des blocs. En effet, ajoute une raison supplémentaire à la croissance des identifiants : l'insertion entre identifiants contigus. Force alors la croissance des identifiants.

1.7 Synthèse

- Systèmes distribués adoptent le modèle de la réplication optimiste pour offrir de meilleures performances, c.-à-d. disponibilité et latence, et assurer la résilience du système, c.-à-d. accroître la capacité de tolérance aux pannes.
- Ce modèle autorise les noeuds à modifier leur copie sans coordination, provoquant ainsi des divergences temporaires. Pour résoudre les potentiels conflits et assurer la convergence à terme des copies, systèmes utilisent les CRDTs en place et lieu des types de données séquentiels.
- CRDTs pour Séquence ont été proposés pour conception d'éditeurs collaboratifs pair-à-pair. Deux approches sont utilisées pour concevoir leur mécanismes de résolution de conflits : l'approche basée sur les pierres tombales et l'approche basée sur les identifiants densément ordonnés.
- Chacune de ces approches introduit un surcoût croissant, pénalisant leurs performances à terme. Plusieurs travaux ont été proposés pour limiter ce surcoût, notamment [42, 81] qui présentent un mécanisme de renommage des identifiants pour les CRDTs pour Séquence basés sur identifiants densément ordonnés.
- Mais cette approche requiert un mécanisme de consensus, des renommages concurrents provoquant un nouveau conflit. Cette contrainte empêche son utilisation dans les systèmes pair-à-pair ne disposant pas de noeuds suffisamment stables et bien connectés pour exécuter le mécanisme de consensus.

1.8 Proposition

- Dans ce manuscrit, nous proposons et présentons un nouveau mécanisme de renommage pour CRDTs pour Séquence, ne nécessitant pas de coordination synchrone entre les noeuds.
- Concevons ce mécanisme pour le CRDT pour Séquence LogootSplit, mais principe de notre approche est générique. Pourrait ainsi l'adapter et proposer un équivalent pour autres CRDTs pour Séquence, e.g. RGASplit.
- Présentons et détaillons notre contribution dans le chapitre suivant.

Chapitre 2

Renommage dans une séquence répliquée

Sommaire

2.1	Présentation de l'approche	32
2.1.1	Modèle du système	32
2.1.2	Définition de l'opération de renommage	32
2.2	Introduction de l'opération <i>rename</i>	34
2.2.1	Opération de renommage proposée	34
2.2.2	Gestion des opérations concurrentes au renommage	35
2.2.3	Évolution du modèle de livraison des opérations	38
2.3	Gestion des opérations <i>rename</i> concurrentes	40
2.3.1	Conflits en cas de renommages concurrents	40
2.3.2	Relation de priorité entre renommages	41
2.3.3	Algorithme d'annulation de l'opération de renommage	42
2.3.4	Processus d'intégration d'une opération	47
2.3.5	Règles de récupération de la mémoire des états précédents	51
2.4	Validation	54
2.4.1	Complexité en temps des opérations	54
2.4.2	Expérimentations	58
2.4.3	Résultats	59
2.5	Discussion	66
2.5.1	Stratégie de génération des opérations <i>rename</i>	66
2.5.2	Stockage des états précédents sur disque	67
2.5.3	Compression et limitation de la taille de l'opération <i>rename</i>	67
2.5.4	Définition de relations de priorité pour minimiser les traitements	68
2.5.5	Report de la transition vers la nouvelle époque cible	69
2.5.6	Utilisation de l'opération de renommage comme mécanisme de compression du log d'opérations	70
2.5.7	Implémentation alternative de l'intégration de l'opération <i>rename</i> basée sur le log d'opérations	72

2.6	Comparaison avec les approches existantes	74
2.6.1	Core-Nebula	74
2.6.2	LSEQ	75
2.7	Conclusion	75

2.1 Présentation de l'approche

Nous proposons un nouveau CRDT pour la *Sequence* appartenant à l'approche des identifiants densément ordonnés : *RenamableLogootSplit* [53, 54]. Cette structure de données permet aux pairs d'insérer et de supprimer des éléments au sein d'une séquence répliquée. Nous introduisons une opération *rename* qui permet de (i) réassigner des identifiants plus courts aux différents éléments de la séquence (ii) fusionner les blocs composant la séquence. Ces deux actions permettent à l'opération *rename* de produire un nouvel état minimisant son surcoût en métadonnées.

2.1.1 Modèle du système

Le système est composé d'un ensemble dynamique de noeuds, les noeuds pouvant rejoindre puis quitter la collaboration tout au long de sa durée. Les noeuds collaborent afin de construire et maintenir une séquence à l'aide de *RenamableLogootSplit*. Chaque noeud possède une copie de la séquence et peut l'éditer sans se coordonner avec les autres. Les modifications des noeuds prennent la forme d'opérations qui sont appliquées immédiatement à leur copie locale. Les opérations sont ensuite transmises de manière asynchrone aux autres noeuds pour qu'ils puissent à leur tour appliquer les modifications à leur copie.

Les noeuds communiquent par l'intermédiaire d'un réseau P2P. Ce réseau est non-fiable : les messages peuvent être perdus, ré-ordonnés ou même livrés à plusieurs reprises. Le réseau peut aussi être sujet à des partitions, qui séparent alors les noeuds en des sous-groupes disjoints. Afin de compenser les limitations du réseau, les noeuds reposent sur une couche de livraison de messages.

Puisque *RenamableLogootSplit* est une extension de *LogootSplit*, cette structure de données partage les mêmes contraintes sur la livraison de messages. La couche de livraison de messages sert donc à livrer les messages à l'application exactement une fois. La couche de livraison de messages a aussi pour tâche de garantir la livraison des opérations de suppression après les opérations d'insertion correspondantes. Aucune autre contrainte n'existe sur l'ordre de livraison des opérations. Finalement, la couche de livraison intègre aussi un mécanisme d'anti-entropie [58]. Ce mécanisme permet aux noeuds de se synchroniser par paires, en détectant et ré-échangeant les messages perdus.

2.1.2 Définition de l'opération de renommage

L'objectif de l'opération *rename* est de réassigner de nouveaux identifiants aux éléments de la séquence répliquée sans modifier son contenu. Puisque les identifiants sont des métadonnées utilisées par la structure de données uniquement afin de résoudre les

conflits, les utilisateurs ignorent leur existence. Les opérations *rename* sont donc des opérations systèmes : elles sont émises et appliquées par les noeuds en coulisses, sans aucune intervention des utilisateurs.

Afin de garantir le respect du modèle de cohérence SEC, nous définissons plusieurs propriétés de sécurité que l'opération *rename* doit respecter. Ces propriétés sont inspirées principalement par celles proposées dans [81].

Propriété 1 (*Déterminisme*) *Les opérations rename sont intégrées par les noeuds sans aucune coordination. Pour assurer que l'ensemble des noeuds atteigne un état équivalent à terme, une opération rename donnée doit toujours générer le même nouvel identifiant à partir de l'identifiant courant.*

Propriété 2 (*Préservation de l'intention de l'utilisateur*) *Bien que l'opération rename n'est pas elle-même n'incarne pas une intention de l'utilisateur, elle ne doit pas entrer en conflit avec les actions des utilisateurs. Notamment, les opérations rename ne doivent pas annuler ou altérer le résultat d'opérations insert et remove du point de vue des utilisateurs.*

Propriété 3 (*Séquence bien formée*) *La séquence répliquée doit être bien formée. Appliquée une opération rename sur une séquence bien formée doit produire une nouvelle séquence bien formée. Une séquence bien formée doit respecter les propriétés suivantes :*

Propriété 3.1 (*Préservation de l'unicité*) *Chaque identifiant doit être unique. Donc, pour une opération rename donnée, chaque identifiant doit être associé à un nouvel identifiant distinct.*

Propriété 3.2 (*Préservation de l'ordre*) *Les éléments de la séquence doivent être triés en fonction de leur identifiants. L'ordre existant entre les identifiants initiaux doit donc être préservé par l'opération rename.*

Propriété 4 (*Commutativité avec les opérations concurrentes*) *Les opérations concurrentes peuvent être délivrées dans des ordres différents à chaque noeud. Afin de garantir la convergence des répliquas, l'ordre d'application d'un ensemble d'opérations concurrentes ne doit pas avoir d'impact sur l'état obtenu. L'opération rename doit donc être commutative avec n'importe quelle opération concurrente.*

La Propriété 4 est particulièrement difficile à assurer. Cette difficulté est due au fait que les opérations *rename* modifient les identifiants assignés aux éléments. Cependant, les autres opérations telles que les opérations *insert* et *remove* reposent sur ces identifiants pour spécifier où insérer les éléments ou lesquels supprimer. Les opérations *rename* sont donc intrinsèquement incompatibles avec les opérations *insert* et *remove* concurrentes. De la même manière, des opérations *rename* concurrentes peuvent réassigner des identifiants différents à des mêmes éléments. Les opérations *rename* concurrentes ne sont donc pas commutatives. Par conséquent, il est nécessaire de concevoir et d'utiliser des méthodes de résolution de conflits pour assurer la Propriété 4.

Dans un souci de simplicité, la présentation de l'opération *rename* est divisée en deux parties. Dans la section 2.2, nous présentons l'opération *rename* proposée avec l'hypothèse qu'aucune opération *rename* concurrente ne peut être générée. Cette hypothèse nous

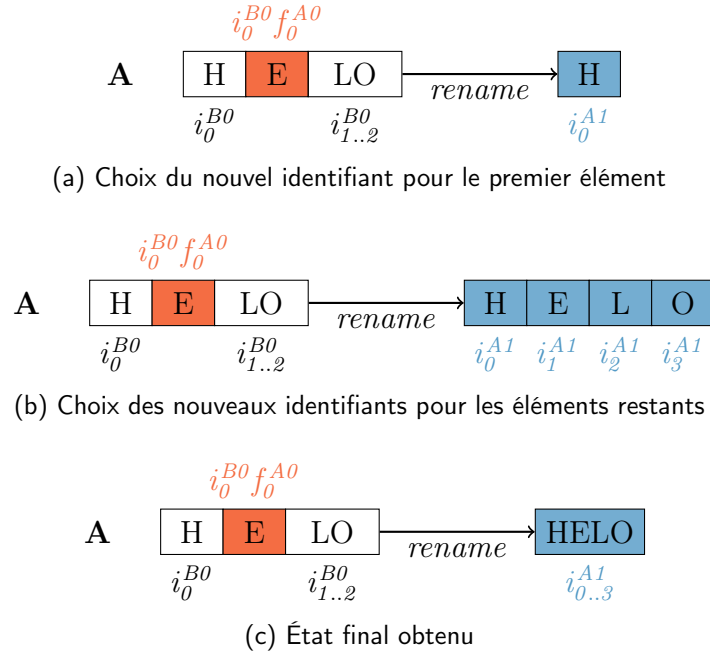


FIGURE 2.1 – Renommage de la séquence sur le noeud A

permet de nous concentrer sur le fonctionnement de l'opération *rename* elle-même ainsi que sur comment gérer les opérations *insert* et *remove* concurrentes. Ensuite, dans la section 2.3, nous supprimons cette hypothèse. Nous présentons alors notre approche pour gérer les scénarios avec des opérations *rename* concurrentes.

2.2 Introduction de l'opération *rename*

2.2.1 Opération de renommage proposée

Notre opération de renommage permet à RenamableLogootSplit de réduire le surcoût en métadonnées des séquences répliquées. Pour ce faire, elle réassigne des identifiants arbitraires aux éléments de la séquence.

Son comportement est illustré dans la Figure 2.1. Dans cet exemple, le noeud A initie une opération *rename* sur son état local. Tout d'abord, le noeud A génère un nouvel identifiant à partir du premier tuple de l'identifiant du premier élément de la séquence (i_0^{B0}). Pour générer ce nouvel identifiant, le noeud A reprend la position de ce tuple (i) mais utilise son propre identifiant de noeud (A) et numéro de séquence actuel (1). De plus, son offset est mis à 0. Le noeud A réassigne l'identifiant résultant (i_0^{A1}) au premier élément de la séquence, comme décrit dans la Figure 2.1a. Ensuite, le noeud A dérive des identifiants contigus pour tous les éléments restants en incrémentant de manière successive l'offset (i_1^{A1} , i_2^{A1} , i_3^{A1}), comme présenté dans la Figure 2.1b. Comme nous assignons des identifiants consécutifs à tous les éléments de la séquence, nous pouvons au final agréger ces éléments en un seul bloc, comme illustré en Figure 2.1c. Ceci permet aux noeuds de bénéficier au mieux de la fonctionnalité des blocs et de minimiser le surcoût en métadonnées.

de l'état résultat.

Pour converger, les autres noeuds doivent renommer leur état de manière identique. Cependant, ils ne peuvent pas simplement remplacer leur état courant par l'état généré par le renommage. En effet, ils peuvent avoir modifié en concurrence leur état. Afin de ne pas perdre ces modifications, les noeuds doivent traiter l'opération *rename* eux-mêmes. Pour ce faire, le noeud qui a généré l'opération *rename* diffuse son *ancien état* aux autres.

Définition 14 (Ancien état) *Un ancien état est la liste des intervalles d'identifiants qui composent l'état courant de la séquence répliquée au moment du renommage.*

De ce fait, nous définissons l'opération *rename* de la manière suivante :

Définition 15 (rename) *Une opération *rename* est un triplet $\langle nodeId, nodeSeq, formerState \rangle$ où*

- *nodeId* est l'identifiant du noeud qui a générée l'opération *rename*.
- *nodeSeq* est le numéro de séquence du noeud au moment de la génération de l'opération *rename*.
- *formerState* est l'ancien état du noeud au moment du renommage.

En utilisant ces données, les autres noeuds calculent le nouvel identifiant de chaque identifiant renommé. Concernant les identifiants insérés de manière concurrente au renommage, nous expliquons dans la sous-section 2.2.2 comment les noeuds peuvent les renommer de manière déterministe.

2.2.2 Gestion des opérations concurrentes au renommage

Après avoir appliqué des opérations *rename* sur leur état local, les noeuds peuvent recevoir des opérations concurrentes. La Figure 2.2 illustre de tels cas.

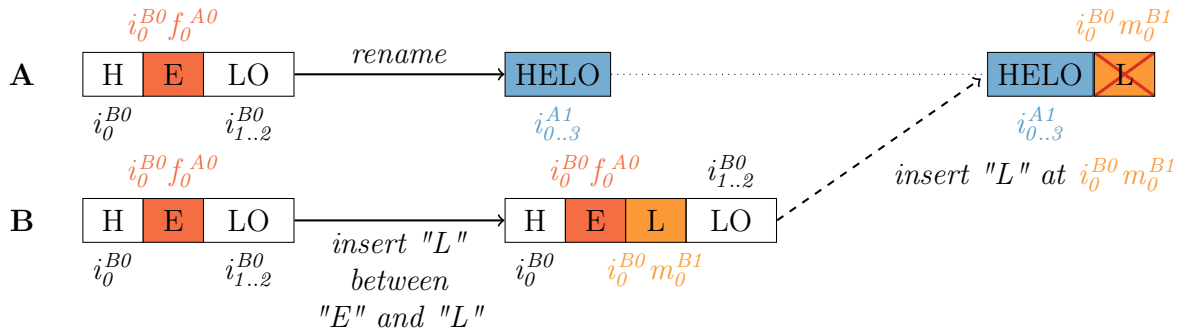


FIGURE 2.2 – Modifications concurrentes menant à une anomalie

Dans cet exemple, le noeud B insère un nouvel élément "L", lui assigne l'identifiant $i_0^{B0} m_0^{B1}$ et diffuse cette modification, de manière concurrente à l'opération *rename* décrite dans la Figure 2.2. À la réception de l'opération *insert*, le noeud A ajoute l'élément inséré au sein de sa séquence, en utilisant l'identifiant de l'élément pour déterminer sa position.

Cependant, puisque les identifiants ont été modifiés par l'opération *rename* concurrente, le noeud A insère le nouvel élément à la fin de sa séquence (puisque $i_3^{A1} <_{id} i_0^{B0} m_0^{B1}$) au lieu de l'insérer à la position souhaitée. Comme illustré par cet exemple, appliquer naïvement les modifications concurrentes provoquerait des anomalies. Il est donc nécessaire de traiter les opérations concurrentes aux opérations *rename* de manière particulière.

Tout d'abord, les noeuds doivent détecter les opérations concurrentes aux opérations *rename*. Pour cela, nous utilisons un système basé sur des *époques*. Initialement, la séquence répliquée débute à l'époque *origine* notée ε_0 . Chaque opération *rename* introduit une nouvelle époque et permet aux noeuds d'y avancer depuis l'époque précédente. Par exemple, l'opération *rename* décrite dans Figure 2.2 permet aux noeuds de faire progresser leur état de ε_0 à ε_{A1} . Nous définissons les époques de la manière suivante :

Définition 16 (Époque) Une époque est un couple $\langle nodeId, nodeSeq \rangle$ où

- *nodeId* est l'identifiant du noeud qui a générée cette époque.
- *nodeSeq* est le numéro de séquence du noeud au moment de la génération de cette époque.

Notons que l'époque générée est caractérisée et identifiée de manière unique par son couple $\langle nodeId, nodeSeq \rangle$.

Au fur et à mesure que les noeuds reçoivent des opérations *rename*, ils construisent et maintiennent localement la *chaîne des époques*. Cette structure de données ordonne les époques en fonction de leur relation *parent-enfant* et associe à chaque époque l'*ancien état* correspondant (c.-à-d. l'*ancien état* inclus dans l'opération *rename* qui a généré cette époque). De plus, les noeuds marquent chaque opération avec leur époque courante au moment de génération de l'opération. À la réception d'une opération, les noeuds comparent l'époque de l'opération à l'époque courante de leur séquence.

Si les époques diffèrent, les noeuds doivent transformer l'opération avant de pouvoir l'intégrer. Les noeuds déterminent par rapport à quelles opérations *rename* doit être transformée l'opération reçue en calculant le chemin entre l'époque de l'opération et leur époque courante en utilisant la *chaîne des époques*.

Les noeuds utilisent la fonction `RENAMEID`, décrite dans l'Algorithme 1, pour transformer les opérations *insert* et *remove* par rapport aux opérations *rename*. Cet algorithme associe les identifiants d'une époque *parente* aux identifiants correspondant dans l'époque *enfant*. L'idée principale de cet algorithme est de renommer les identifiants inconnus au moment de la génération de l'opération *rename* en utilisant leur prédecesseur. Un exemple est présenté dans la Figure 2.3. Cette figure décrit le même scénario que la Figure 2.2, à l'exception que le noeud A utilise `RENAMEID` pour renommer les identifiants générés de façon concurrente avant de les insérer dans son état.

L'algorithme procède de la manière suivante. Tout d'abord, le noeud récupère le prédecesseur de l'identifiant donné $i_0^{B0} m_0^{B1}$ dans l'ancien état : $i_0^{B0} f_0^{A0}$. Ensuite, il calcule l'équivalent de $i_0^{B0} f_0^{A0}$ dans l'état renommé : i_1^{A1} . Finalement, le noeud A concatène cet identifiant et l'identifiant donné pour générer l'identifiant correspondant dans l'époque *enfant* : $i_1^{A1} i_0^{B0} m_0^{B1}$. En réassignant cet identifiant à l'élément inséré de manière concurrente, le noeud A peut l'insérer à son état tout en préservant l'ordre souhaité.

Algorithme 1 Fonctions principales pour renommer un identifiant

```

function RENAMEID(id, renamedIds, nId, nSeq)
  length  $\leftarrow$  renamedIds.length
  firstId  $\leftarrow$  renamedIds[0]
  lastId  $\leftarrow$  renamedIds[length - 1]
  pos  $\leftarrow$  position(firstId)

  if id < firstId then
    newFirstId  $\leftarrow$  new Id(pos, nId, nSeq, 0)
    return renIdLessThanFirstId(id, newFirstId)
  else if id  $\in$  renamedIds then
    index  $\leftarrow$  findIndex(id, renamedIds)
    return new Id(pos, nId, nSeq, index)
  else if lastId < id then
    newLastId  $\leftarrow$  new Id(pos, nId, nSeq, length - 1)
    return renIdGreaterThanLastId(id, newLastId)
  else
    return renIdFromPredId(id, renamedIds, pos, nId, nSeq)
  end if
end function

function RENIDFROMPREDID(id, renamedIds, pos, nId, nSeq)
  index  $\leftarrow$  findIndexOfPred(id, renamedIds)
  newPredId  $\leftarrow$  new Id(pos, nId, nSeq, index)

  return concat(newPredId, id)
end function

```

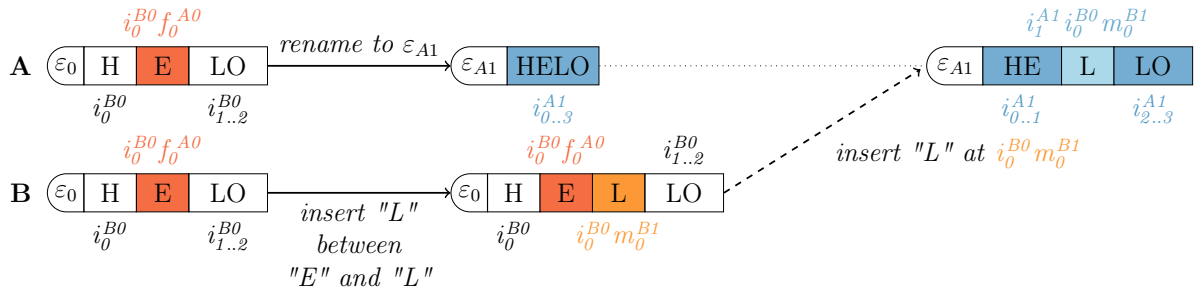


FIGURE 2.3 – Renommage de la modification concurrente avant son intégration en utilisant RENAMEID afin de maintenir l'ordre souhaité

RENAMEID permet aussi aux noeuds de gérer le cas contraire : intégrer des opérations *rename* distantes sur leur copie locale alors qu'ils ont précédemment intégré des modifications concurrentes. Ce cas correspond à celui du noeud B dans la Figure 2.3. À la réception de l'opération *rename* du noeud A, le noeud B utilise RENAMEID sur chacun des identifiants de son état pour le renommer et atteindre un état équivalent à celui du noeud A.

L'Algorithme 1 présente seulement le cas principal de RENAMEID, c.-à-d. le cas où l'identifiant à renommer appartient à l'intervalle des identifiants formant l'ancien état ($firstId \leq_{id} id \leq_{id} lastId$). Les fonctions pour gérer les autres cas, c.-à-d. les cas où

l'identifiant à renommer n'appartient pas à cet intervalle ($id <_{id} firstId$ ou $lastId <_{id} id$), sont présentées dans l'Annexe A.

L'algorithme que nous présentons ici permet aux noeuds de renommer leur état identifiant par identifiant. Une extension possible est de concevoir `RENAMEBLOCK`, une version améliorée qui renomme l'état bloc par bloc. `RENAMEBLOCK` réduirait le temps d'intégration des opérations *rename*, puisque sa complexité en temps ne dépendrait plus du nombre d'identifiants (c.-à-d. du nombre d'éléments) mais du nombre de blocs. De plus, son exécution réduirait le temps d'intégration des prochaines opérations *rename* puisque le mécanisme de renommage regroupe les éléments en moins de blocs.

2.2.3 Évolution du modèle de livraison des opérations

L'introduction de l'opération *rename* nécessite de faire évoluer le modèle de livraison des opérations associé à `RenamableLogootSplit`. Afin d'illustrer cette nécessité, considérons l'exemple suivant :

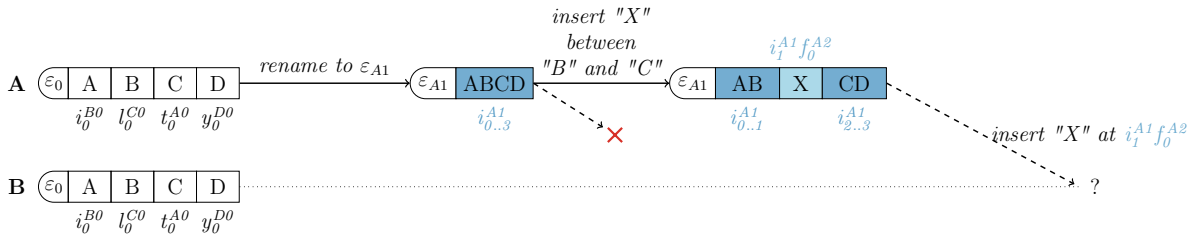


FIGURE 2.4 – Livraison d'une opération *insert* sans avoir reçu l'opération *rename* précédente

Dans la Figure 2.4, les noeuds A et B répliquent tous deux une même séquence, contenant les éléments "ABCD". Tout d'abord, le noeud A procède au renommage de cet état. Puis il insère un nouvel élément, "X", entre "B" et "C". Les opérations correspondantes aux actions du noeud A sont diffusées sur le réseau.

Cependant, l'opération *rename* n'est pas livrée au noeud B, par exemple suite à un problème réseau. L'opération *insert* est quant à elle correctement livrée à ce dernier. Le noeud B doit alors intégrer dans son état un élément et l'identifiant qui lui est attaché. Mais cet identifiant est issu d'une époque (ε_{A1}) différente de son époque actuelle (ε_0) et dont le noeud n'avait pas encore connaissance. Il convient de s'interroger sur l'état à produire dans cette situation.

Comme nous l'avons déjà illustré par la Figure 2.2, les identifiants d'une époque ne peuvent être comparés qu'aux identifiants de la même époque. Tenter d'intégrer une opération *insert* ou *remove* provenant d'une époque encore inconnue ne résulterait qu'en un état incohérent et une transgression de l'intention utilisateur. Il est donc nécessaire d'empêcher ce scénario de se produire.

Pour cela, nous proposons de faire évoluer le modèle de livraison des opérations de `RenamableLogootSplit`. Celui-ci repose sur celui de `LogootSplit`, que nous avons défini dans la Définition 13. Pour rappel, ce modèle requiert que (i) les opérations soient livrées

qu'une seule et unique fois au CRDT, (ii) les opérations *remove* soient livrées au CRDT qu'après les opérations *insert* ajoutant les éléments à supprimer.

Pour prévenir les scénarios tels que celui illustré par la Figure 2.4 nous y ajoutons la règle suivante : les opérations *rename* doivent être livrées à la structure de données avant les opérations qui ont une dépendance causale vers ces dernières. Nous obtenons donc le modèle de livraison suivant :

Définition 17 (Exactly-once + Causal remove + Epoch-based) *Le modèle de livraison Exactly-once + Causal remove + Epoch-based définit les 4 règles suivantes sur la livraison des opérations :*

1. Une opération doit être délivrée à l'ensemble des noeuds à terme,
2. Une opération doit être délivrée qu'une seule et unique fois aux noeuds,
3. Une opération *remove* doit être délivrée à un noeud une fois que les opérations *insert* des éléments concernés par la suppression ont été délivrées à ce dernier.
4. Une opération doit être délivrée à un noeud une fois que l'opération *rename* une fois que l'opération *rename* qui introduit son époque de génération a été délivrée à ce dernier.

Il est cependant intéressant de noter que la livraison de l'opération *rename* ne requiert pas de contraintes supplémentaires. Notamment, une opération *rename* peut être livrée dans le désordre par rapport aux opérations *insert* et *remove* dont elle dépend causalement. La Figure 2.5 présente un exemple de ce cas figure.

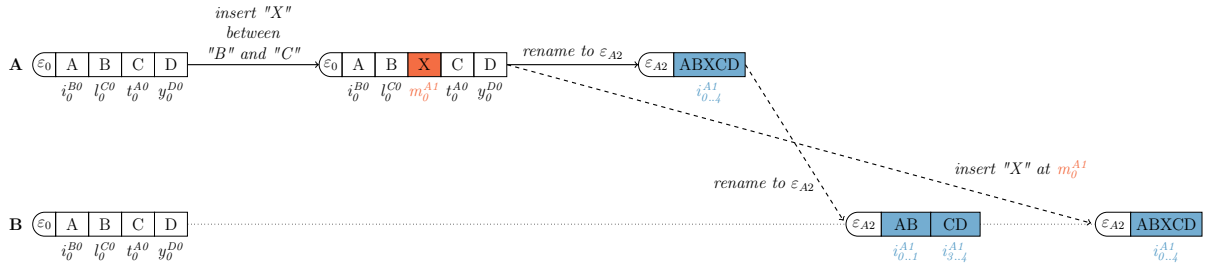


FIGURE 2.5 – Livraison désordonnée d'une opération *rename* et de l'opération *insert* qui la précède

Dans cet exemple, les noeuds A et B répliquent tous deux une même séquence, contenant les éléments "ABCD". Le noeud A commence par insérer un nouvel élément, "X", entre les éléments "B" et "C". Puis il procède au renommage de son état. Les opérations correspondantes aux actions du noeud A sont diffusées sur le réseau.

Cependant, suite à un aléa du réseau, le noeud B reçoit les deux opérations *insert* et *rename* dans le désordre. L'opération *rename* est donc livrée en première au noeud B. En utilisant les informations contenues dans l'opération, le noeud B est renommé chaque identifiant composant son état.

Ensuite, le noeud B reçoit l'opération *insert*. Comme l'époque de génération de l'opération *insert* (ϵ_0) est différente de celle de son état courant (ϵ_{A2}), le noeud B utilise *RENAMEID* pour renommer l'identifiant avant de l'insérer. m_0^{A1} faisant partie de l'*ancien*

état, le noeud B utilise l'index de cet identifiant dans l'*ancien état* (2) pour calculer son équivalent à l'époque ε_{A2} (i_2^{A2}). Le noeud B insère l'élément "X" avec ce nouvel identifiant et converge alors avec le noeud A, malgré la livraison dans le désordre des opérations.

2.3 Gestion des opérations *rename* concurrentes

2.3.1 Conflits en cas de renommages concurrents

Nous considérons à présent les scénarios avec des opérations *rename* concurrentes. Figure 2.6 développe le scénario décrit précédemment dans Figure 2.3.

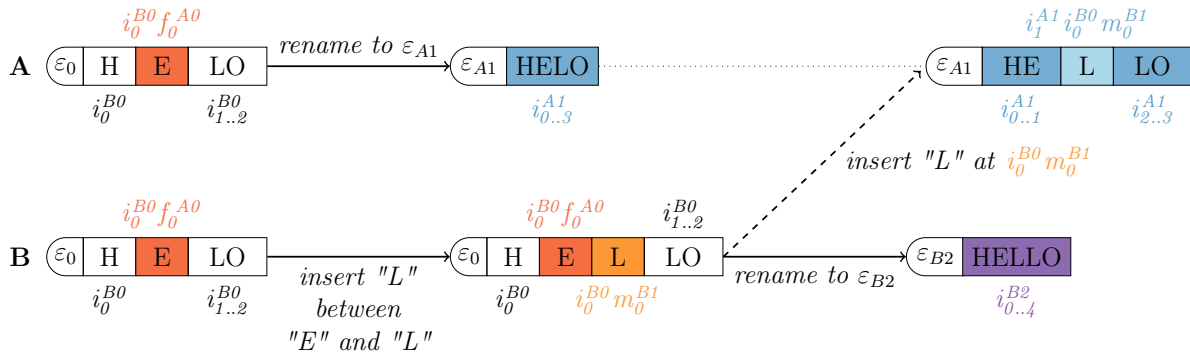


FIGURE 2.6 – Opérations *rename* concurrentes menant à des états divergents

Après avoir diffusé son opération *insert*, le noeud B effectue une opération *rename* sur son état. Cette opération réassigne à chaque élément un nouvel identifiant à partir de l'identifiant du premier élément de la séquence (i_0^{B0}), de l'identifiant du noeud (**B**) et de son numéro de séquence courant (2). Cette opération introduit aussi une nouvelle époque : ε_{B2} . Puisque l'opération *rename* de A n'a pas encore été délivrée au noeud B à ce moment, les deux opérations *rename* sont concurrentes.

Puisque des époques concurrentes sont générées, les époques forment désormais l'*arbre des époques*. Nous représentons dans la Figure 2.7 l'*arbre des époques* que les noeuds obtiennent une fois qu'ils se sont synchronisés à terme. Les époques sont représentées sous la forme de noeuds de l'arbre et la relation *parent-enfant* entre elles est illustrée sous la forme de flèches noires.

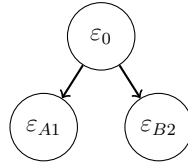


FIGURE 2.7 – *Arbre des époques* correspondant au scénario décrit dans la Figure 2.6

À l'issue du scénario décrit dans la Figure 2.6, les noeuds A et B sont respectivement aux époques ε_{A1} et ε_{B2} . Pour converger, tous les noeuds devraient atteindre la même époque à terme. Cependant, la fonction `RENAMEID` décrite dans l'Algorithme 1 permet

seulement aux noeuds de progresser d'une époque *parente* à une de ses époques *enfants*. Le noeud A (resp. B) est donc dans l'incapacité de progresser vers l'époque du noeud B (resp. A). Il est donc nécessaire de faire évoluer notre mécanisme de renommage pour sortir de cette impasse.

Tout d'abord, les noeuds doivent se mettre d'accord sur une époque commune de l'*arbre des époques* comme époque cible. Afin d'éviter des problèmes de performances dus à une coordination synchrone, les noeuds doivent sélectionner cette époque de manière non-coordonnée, c.-à-d. en utilisant seulement les données présentes dans l'*arbre des époques*. Nous présentons un tel mécanisme dans la sous-section 2.3.2.

Ensuite, les noeuds doivent se déplacer à travers l'*arbre des époques* afin d'atteindre l'époque cible. La fonction `RENAMEID` permet déjà aux noeuds de descendre dans l'arbre. Les cas restants à gérer sont ceux où les noeuds se trouvent actuellement à une époque *soeur* ou *cousine* de l'époque cible. Dans ces cas, les noeuds doivent être capable de remonter dans l'*arbre des époques* pour retourner au Plus Petit Ancêtre Commun (PPAC) de l'époque courante et l'époque cible. Ce déplacement est en fait similaire à annuler l'effet des opérations *rename* précédemment appliquées. Nous proposons un algorithme, `REVERTRENAMEID`, qui remplit cet objectif dans la sous-section 2.3.3.

2.3.2 Relation de priorité entre renommages

Pour que chaque noeud sélectionne la même époque cible de manière non-coordonnée, nous définissons la relation *priority*.

Définition 18 (Relation *priority* $<_{\epsilon}$) La relation *priority* $<_{\epsilon}$ est un ordre strict total sur l'ensemble des époques. Elle permet aux noeuds de comparer n'importe quelle paire d'époques.

En utilisant la relation *priority*, nous définissons l'époque cible de la manière suivante :

Définition 19 (Époque cible) L'époque cible est l'époque de l'ensemble des époques vers laquelle les noeuds doivent progresser. Les noeuds sélectionnent comme époque cible l'époque maximale d'après l'ordre établi par *priority*.

Pour définir la relation *priority*, nous pouvons choisir entre plusieurs stratégies. Dans le cadre de ce travail, nous utilisons l'ordre lexicographique sur le chemin des époques dans l'*arbre des époques*. La Figure 2.8 fournit un exemple.

La Figure 2.8a décrit une exécution dans laquelle trois noeuds A, B et C génèrent plusieurs opérations avant de se synchroniser à terme. Comme seules les opérations *rename* sont pertinentes pour le problème qui nous occupe, nous représentons seulement ces opérations dans cette figure. Initialement, le noeud A génère une opération *rename* qui introduit l'époque ϵ_{A1} . Cette opération est délivrée au noeud C, qui génère ensuite sa propre opération *rename* qui introduit l'époque ϵ_{C6} . De manière concurrente à ces opérations, le noeud B génère deux opérations *rename*, introduisant ϵ_{B2} et ϵ_{B7} .

Une fois que les noeuds se sont synchronisés, ils obtiennent l'*arbre des époques* représenté dans la Figure 2.8b. Dans cette figure, la flèche tireté rouge représente l'ordre entre

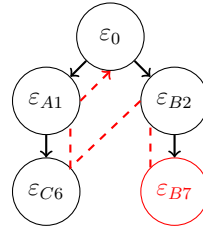
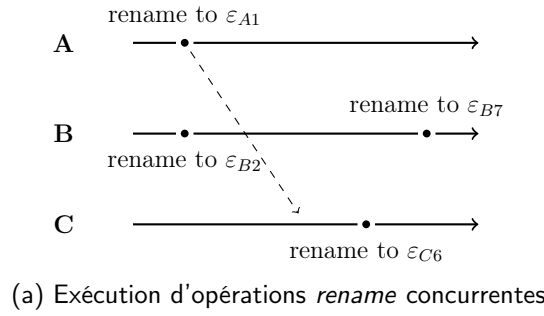


FIGURE 2.8 – Sélectionner l'époque cible d'une exécution d'opérations *rename* concurrentes

les époques d'après la relation *priority* tandis que l'époque cible choisie est représentée sous la forme d'un noeud rouge.

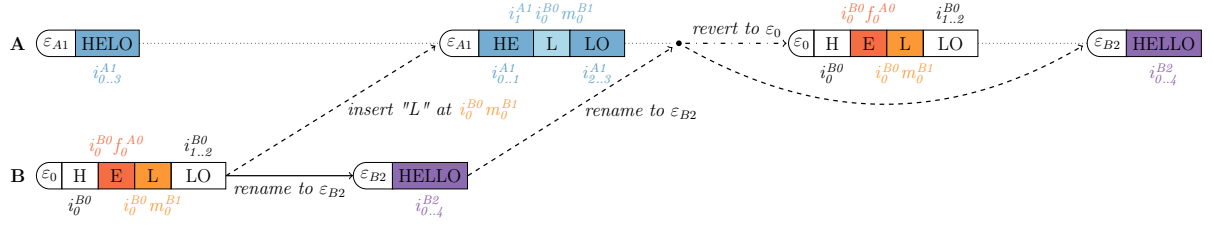
Pour déterminer l'époque cible, les noeuds reposent sur la relation *priority*. D'après l'ordre lexicographique sur le chemin des époques dans l'*arbre des époques*, nous avons $\varepsilon_0 < \varepsilon_0\varepsilon_{A1} < \varepsilon_0\varepsilon_{A1}\varepsilon_{C6} < \varepsilon_0\varepsilon_{B2} < \varepsilon_0\varepsilon_{B2}\varepsilon_{B7}$. Chaque noeud sélectionne donc ε_{B7} comme époque cible de manière non-coordonnée.

D'autres stratégies pourraient être proposées pour définir la relation *priority*. Par exemple, *priority* pourrait reposer sur des métriques intégrées au sein des opérations *rename* pour représenter le travail accumulé sur le document. Cela permettrait de favoriser la branche de l'*arbre des époques* avec le plus de collaborateurs actifs pour minimiser la quantité globale de calculs effectués par les noeuds du système. Nous approfondissons ce sujet dans la sous-section 2.5.4.

2.3.3 Algorithme d'annulation de l'opération de renommage

À présent, nous développons le scénario présenté dans la Figure 2.6. Dans la Figure 2.9, le noeud A reçoit l'opération *rename* du noeud B. Cette opération est concurrente à l'opération *rename* que le noeud A a appliqué précédemment. D'après la relation *priority* proposée, le noeud A sélectionne l'époque introduite ε_{B2} comme l'époque cible ($\varepsilon_{A1} <_{\varepsilon} \varepsilon_{B2}$). Mais pour pouvoir renommer son état vers l'époque ε_{B2} , il doit au préalable faire revenir son état courant de l'époque ε_{A1} à un état équivalent à l'époque ε_0 . Nous devons définir un mécanisme permettant aux noeuds d'annuler les effets d'une opération *rename* appliquée précédemment.

C'est précisément le but de `REVERTRENAMEID`, qui associe les identifiants de l'époque *enfant* aux identifiants correspondant dans l'époque *parente*. Nous décrivons cette fonction

2.3. Gestion des opérations *rename* concurrentesFIGURE 2.9 – Annulation d’une opération *rename* intégrée précédemment en présence d’un identifiant inséré en concurrence

dans l’Algorithme 2.

Les objectifs de REVERTRENAMEID sont les suivants : (i) Restaurer à leur ancienne valeur les identifiants générés causalement avant l’opération *rename* annulée (ii) Restaurer à leur ancienne valeur les identifiants générés de manière concurrente à l’opération *rename* annulée (iii) Assigner de nouveaux identifiants respectant l’ordre souhaité aux éléments qui ont été insérés causalement après l’opération *rename* annulée.

Le cas (i) est le plus trivial. Pour retrouver la valeur de *id* à partir de *newId*⁵, REVERTRENAMEID utilise simplement la valeur de offset de *newId*. En effet, cette valeur correspond à l’index de *id* dans l’ancien état (c.-à-d. $renamedIds[offset] = id$). Par exemple, dans la Figure 2.9, l’identifiant i_0^{A1} a pour offset 0, REVERTRENAMEID renvoie donc $renamedIds[0] = i_0^{B0}$.

Les cas (ii) et (iii) sont gérés en utilisant les stratégies suivantes. Le motif générique pour l’identifiant *newId* est de la forme *newPredId tail*. Deux invariants sont associés à ce motif. D’après la Propriété 3.2, nous avons :

$$newId \in]newPredId, newSuccId[$$

et nous devons obtenir :

$$id \in]predId, succId[$$

Le premier sous-cas se produit quand nous avons $tail \in]predId, succId[$. Dans ce cas, *newId* peut résulter d’une opération *insert* concurrent à l’opération *rename* (c.-à-d. le cas (ii)). Nous avons alors :

$$newId \in]newPredId \ predId, newPredId \ succId[$$

Dans cette situation, *newId* a été obtenu en utilisant RENIDFROMPREDID et nous avons $id = tail$. Nous observons qu’en renvoyant *tail*, REVERTRENAMEID valident les deux contraintes, c.-à-d. préserver l’ordre souhaité et restaurer à sa valeur initiale l’identifiant. Pour illustrer ce cas, considérons l’identifiant $i_1^{A1} i_0^{B0} m_0^{B1}$ dans Figure 2.9. Pour cet identifiant, nous avons :

- $newPredId = i_1^{A1}$, donc $predId = i_0^{B0} f_0^{A0}$ d’après le cas (i)
- $newSuccId = i_2^{A1}$, donc $succId = i_1^{B0}$ d’après le cas (i)

5. Nous appelons *newX* les identifiants dans l’époque résultant de l’application d’une opération *rename*, tandis que *X* décrit leur équivalent à l’époque initiale.

Algorithme 2 Fonctions principales pour annuler le renommage appliqué précédemment à un identifiant

```

function REVERTRENAMEID(id, renamedIds, nId, nSeq)
  length ← renamedIds.length
  firstId ← renamedIds[0]
  lastId ← renamedIds[length - 1]
  pos ← position(firstId)

  newFirstId ← new Id(pos, nId, nSeq, 0)
  newLastId ← new Id(pos, nId, nSeq, length - 1)

  if id < newFirstId then
    return revRenIdLessThanNewFirstId(id, firstId, newFirstId)
  else if isRenamedId(id, pos, nId, nSeq, length) then
    index ← getFirstOffset(id)
    return renamedIds[index]
  else if newLastId < id then
    return revRenIdGreaterThanNewLastId(id, lastId)
  else
    index ← getFirstOffset(id)
    return revRenIdfromPredId(id, renamedIds, index)
  end if
end function

function REVRENIDFROMPREDID(id, renamedIds, index)
  predId ← renamedIds[index]
  succId ← renamedIds[index + 1]
  tail ← getTail(id, 1)

  if tail < predId then
    ▷ id has been inserted causally after the rename op
    return concat(predId, MIN_TUPLE, tail)
  else if succId < tail then
    ▷ id has been inserted causally after the rename op
    offset ← getLastOffset(succId) - 1
    predOfSuccId ← createIdFromBase(succId, offset)
    return concat(predOfSuccId, MAX_TUPLE, tail)
  else
    return tail
  end if
end function

```

Nous avons donc bien :

$$i_1^{A1} i_0^{B0} m_0^{B1} \in]i_1^{A1} i_0^{B0} f_0^{A0}, i_1^{A1} i_1^{B0}]$$

et $tail = i_0^{B0} m_0^{B1}$. Renvoyer cette valeur nous permet ainsi de conserver l'ordre entre les identifiants puisque :

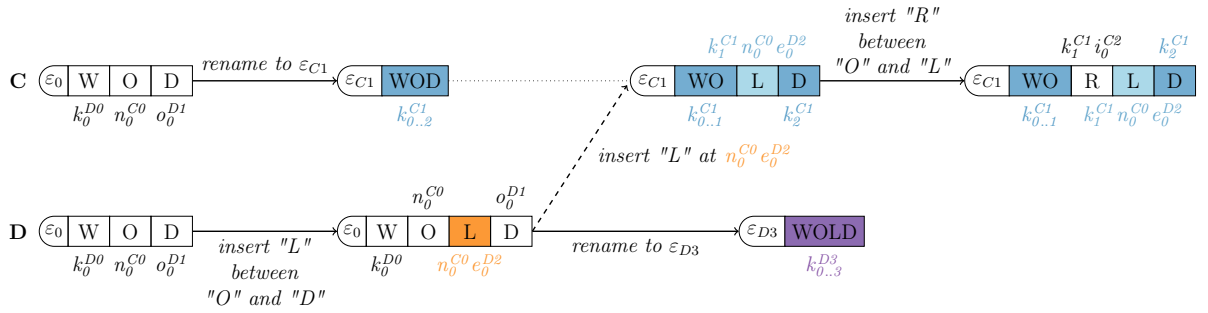
$$i_0^{B0} f_0^{A0} <_{id} i_0^{B0} m_0^{B1} <_{id} i_1^{B0}$$

Le second sous-cas correspond au cas où nous avons $tail < predId$. $newId$ ne peut avoir été inséré que causalement après l'opération *rename* (c.-à-d. le cas (iii)). Nous avons

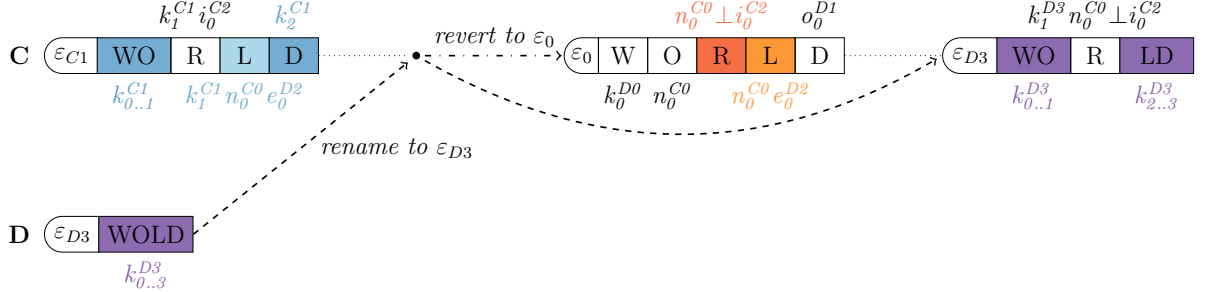
alors :

$$newId \in]newPredId, newPredId \text{ } predId[$$

Puisque $newId$ a été inséré causalement après l'opération *rename*, il n'existe pas de contrainte sur la valeur à retourner autre que la Propriété 3.2. Pour gérer ce cas, nous introduisons deux nouveaux tuples exclusifs au mécanisme de renommage : MIN_TUPLE et MAX_TUPLE , notés respectivement \perp et \top . Ils sont respectivement le tuple minimal et maximal utilisables pour générer des identifiants. En utilisant MIN_TUPLE , REVERTRENAMEID est capable de renvoyer une valeur pour id adaptée à l'ordre souhaité (avec $id = predId \perp tail$). Nous justifions ce comportement à l'aide de la Figure 2.10.



(a) Génération d'une opération *insert* dépendante causalement d'une opération *rename*



(b) Annulation de l'opération *rename* précédente au profit d'une opération *rename* concurrente

FIGURE 2.10 – Annulation d'une opération *rename* intégrée précédemment en présence d'un identifiant inséré causalement après

Dans la Figure 2.10, les noeuds C et D répliquent une même séquence contenant les éléments "WOD". Dans la Figure 2.10a, le noeud C commence par renommer son état. En concurrence, le noeud D insère l'élément "L" entre les éléments "O" et "D". L'opération *insert* correspondante est délivrée au noeud C, qui l'intègre en suivant le comportement défini en sous-section 2.2.2. Le noeud C procède ensuite à l'insertion de l'élément "R" entre les éléments "O" et "L". Cette insertion dépend donc causalement de l'opération *rename* effectuée précédemment par C. En parallèle, le noeud D effectue un renommage de son état. Cette opération *rename* est donc concurrente à l'opération *rename* générée précédemment par C.

Dans la Figure 2.10b, l'opération *rename* de D est délivrée au noeud C. L'époque introduite par cette opération étant prioritaire par rapport à l'époque actuelle de C ($\varepsilon_{C1} <_{\varepsilon} \varepsilon_{D3}$), le noeud C procède à l'annulation de son opération *rename*.

L'identifiant qui nous intéresse ici est l'identifiant inséré causalement après l'opération *rename* annulée : $k_1^{C1} i_0^{C2}$. Cet identifiant est compris entre les identifiants suivants :

$$k_1^{C1} <_{id} k_1^{C1} i_0^{C2} <_{id} k_1^{C1} n_0^{C0} e_0^{D2}$$

D'après les règles présentées précédemment :

- k_1^{C1} est transformé en n_0^{C0} (cas (i))
- $k_1^{C1} n_0^{C0} e_0^{D2}$ est transformé en $n_0^{C0} e_0^{D2}$ (cas (ii))

Nous devons générer un identifiant *id* à partir de $k_1^{C1} i_0^{C2}$ tel que :

$$n_0^{C0} <_{id} id <_{id} n_0^{C0} e_0^{D2}$$

Utiliser *predId* (n_0^{C0}) en tant que préfixe de *id* nous permet de garantir que $n_0^{C0} <_{id} id$. Cependant, appliquer la même stratégie que pour le cas (ii) pour générer *id* transgresserait la Propriété 3.2. En effet, nous obtiendrions $id = n_0^{C0} i_0^{C2}$, or $n_0^{C0} i_0^{C2} \not<_{id} n_0^{C0} e_0^{D2}$.

Ainsi, nous devons choisir un autre préfixe dans cette situation, notamment pour garantir que l'identifiant résultant sera plus petit que les identifiants suivants. C'est pour cela que nous introduisons *MIN_TUPLE*. En concaténant *predId* et le tuple minimal, nous obtenons un préfixe nous permettant à la fois de garantir que $n_0^{C0} <_{id} id$ et que $id <_{id} n_0^{C0} e_0^{D2}$. Nous obtenons donc $id = n_0^{C0} \perp i_0^{C2}$, ce qui respecte la Propriété 3.2.

Finalement, le dernier sous-cas est le pendant du sous-cas précédent et se produit lorsque nous avons $succId < tail$. Nous avons alors :

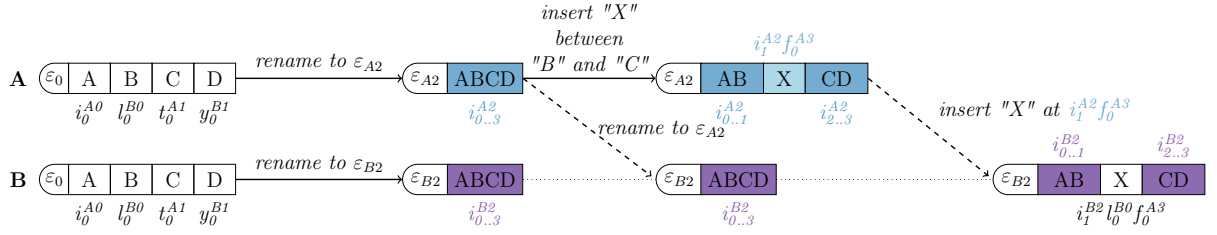
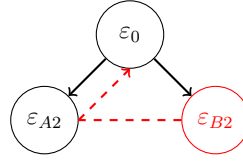
$$newId \in]newPredId \ succId, newSuccId[$$

La stratégie pour gérer ce cas est similaire et consiste à ajouter un préfixe pour créer l'ordre souhaité. Pour générer ce préfixe, REVERTRENAMEID utilise *predOfSuccId* et *MAX_TUPLE*. *predOfSuccId* est obtenu en décrémentant le dernier offset de *succId*. Ainsi, pour préserver l'ordre souhaité, REVERTRENAMEID renvoie *id* avec $id = predOfSuccId \top tail$.

Comme pour l'Algorithme 1, l'Algorithme 2 ne présente seulement que le cas principal de REVERTRENAMEID. Il s'agit du cas où l'identifiant à restaurer appartient à l'intervalle des identifiants renommés ($newFirstId \leq_{id} id \leq_{id} newLastId$). Les fonctions pour gérer les cas restants sont présentées dans l'Annexe B.

Notons que RENAMEID et REVERTRENAMEID ne sont pas des fonctions réciproques. REVERTRENAMEID restaure à leur valeur initiale les identifiants insérés causalement avant ou de manière concurrente à l'opération *rename*. Par contre, RENAMEID ne fait pas de même pour les identifiants insérés causalement après l'opération *rename*. Rejouer une opération *rename* précédemment annulée altère donc ces identifiants. Cette modification peut entraîner une divergence entre les noeuds, puis qu'un même élément sera désigné par des identifiants différents.

Ce problème est toutefois évité dans notre système grâce à la relation *priority* utilisée. Puisque la relation *priority* est définie en utilisant l'ordre lexicographique sur le chemin

2.3. Gestion des opérations *rename* concurrentes(a) Exécution nécessitant l'intégration d'une opération *insert* provenant d'une époque concurrente(b) Arbre des époques de B à la réception de l'opération *insert*FIGURE 2.11 – Intégration d'une opération *insert* distante

des époques dans l'*arbre des époques*, les noeuds se déplacent seulement vers l'époque la plus à droite de l'*arbre des époques* lorsqu'ils changent d'époque. Les noeuds évitent donc d'aller et revenir entre deux mêmes époques, et donc d'annuler et rejouer les opérations *rename* correspondantes.

2.3.4 Processus d'intégration d'une opération

Le processus d'intégration d'une opération distante distingue deux cas différents : (i) le cas de figure où l'opération reçue est une opération *insert* ou *remove* (ii) le cas de figure où l'opération reçue est une opération *rename*.

Intégration d'une opération *insert* ou *remove* distante

Dans l'Algorithme 3, nous présentons l'algorithme d'intégration d'une opération *insert* distante dans RenamableLogootSplit.

Cet algorithme se décompose en de multiples étapes. Afin d'illustrer chacune d'entre elles, nous utilisons l'exemple représenté par la Figure 2.11.

Dans la Figure 2.11a, deux noeuds A et B éditent une séquence répliquée via RenamableLogootSplit. Initialement, les deux noeuds possèdent des répliques identiques. Le noeud A commence par effectuer une opération *rename*. Il génère alors l'état équivalent à son état précédent, à la nouvelle époque ε_{A2} . Puis il effectue une opération *insert*, insérant un nouvel élément "X" entre les éléments "B" et "C". L'identifiant $i_1^{A2} f_0^{A3}$ est attribué à ce nouvel élément. Chacune des opérations du noeud A est diffusée sur le réseau.

De son côté, le noeud B génère en concurrence sa propre opération *rename* sur l'état initial. Il obtient alors un état équivalent, à l'époque ε_{B2} . Il reçoit ensuite l'opération *rename* du noeud A, qu'il intègre. Puisque $\varepsilon_{A2} < \varepsilon_{B2}$, le noeud B ne modifie pas son époque courante (ε_{B2}). Le noeud B obtient toutefois l'*arbre des époques* représenté dans la Figure 2.11b.

Algorithme 3 Algorithme d'intégration d'une opération *insert* distante

```

function INSREMOTE(seq, epochTree, currentEpoch, insOp)
  if currentEpoch = opEpoch then
    insert(seq, getIdBegin(insertOp), getContent(insertOp))
  else
5:    insertedIdInterval  $\leftarrow$  getInsertedIdInterval(insOp)
    ids  $\leftarrow$  expand(insertedIdInterval)

    opEpoch  $\leftarrow$  getEpoch(insOp)
     $\langle$ epochsToRevert, epochsToApply $\rangle \leftarrow$  getPathBetweenEpochs(epochTree, opEpoch, currentE-
10:    poch)

    for epoch in epochsToRevert do
      renamedIds  $\leftarrow$  getRenamedIds(epochTree, epoch)
      nId  $\leftarrow$  getNodeId(epochTree, epoch)
      nSeq  $\leftarrow$  getNodeSeq(epochTree, epoch)
15:      revertRenameIdpartial  $\leftarrow$  papply(revertRenameId, renamedIds, nId, nSeq)
      ids  $\leftarrow$  map(ids, revertRenameIdpartial)
    end for

    for epoch in epochsToApply do
20:      renamedIds  $\leftarrow$  getRenamedIds(epochTree, epoch)
      nId  $\leftarrow$  getNodeId(epochTree, epoch)
      nSeq  $\leftarrow$  getNodeSeq(epochTree, epoch)
      renameIdpartial  $\leftarrow$  papply(renameId, renamedIds, nId, nSeq)
      ids  $\leftarrow$  map(ids, renameIdpartial)
25:    end for

    content  $\leftarrow$  getContent(insOp)
    newIdIntervals  $\leftarrow$  aggregate(ids)
    insertOps  $\leftarrow$  generateInsertOps(newIdIntervals, content)
30:    for insertOp in insertOps do
      insert(seq, getIdBegin(insertOp), getContent(insertOp))
    end for
  end if
end function

```

Puis le noeud B reçoit l'opération *insert* de l'élément "X" à la position $i_1^{A2}f_0^{A3}$. C'est le traitement de cette opération que nous allons détailler ici.

Tout d'abord, le noeud B compare l'époque de l'opération avec l'époque courante de la séquence. Si les deux époques correspondaient, le noeud B pourrait intégrer l'opération directement en utilisant l'algorithme de LogootSplit dénommé ici INSERT. Mais dans le cas présent, l'époque de l'opération (ε_{A2}) est différente de l'époque courante (ε_{B2}). Il lui est donc nécessaire de transformer l'opération avant de pouvoir l'appliquer.

Pour cela, le noeud doit identifier les transformations à appliquer à l'opération. Pour ce faire, le noeud calcule le chemin entre l'époque de l'opération et l'époque courante à l'aide de la fonction GETPATHBETWEENEPOCHS (ligne 9).

La fonction GETPATHBETWEENEPOCHS applique l'algorithme suivant : (i) elle calcule le chemin entre l'époque de l'opération et la racine de l'*arbre des époques* ($[\varepsilon_{A2}, \varepsilon_0]$) (ii) elle calcule le chemin entre l'époque courante et la racine de l'*arbre des époques*

($[\varepsilon_{B2}, \varepsilon_0]$) (iii) elle détermine la première intersection entre ces deux chemins (ε_0). Cette époque correspond au Plus Petit Ancêtre Commun (PPAC) entre l'époque de l'opération et l'époque courante. (iv) elle tronque les deux chemins au niveau du PPAC ($[\varepsilon_{A2}]$ et $[\varepsilon_{B2}]$) (v) elle inverse l'ordre des époques du chemin entre l'époque courante et la racine ($[\varepsilon_{B2}]$) (vi) elle retourne les deux chemins obtenus ($([\varepsilon_{A2}], [\varepsilon_{B2}])$).

Le chemin entre l'époque de l'opération et l'époque PPAC ($[\varepsilon_{A2}]$) correspond aux renommages dont les effets doivent être retirés de l'opération. Pour cela, le noeud récupère les informations de chaque renommage via l'*arbre des époques* (lignes 12-14). Puis il applique REVERTRENAMEID sur chaque identifiant de l'opération (ligne 16). Le noeud procède ensuite de manière similaire pour les époques appartenant au chemin entre l'époque PPAC et l'époque courante ($[\varepsilon_{B2}]$), qui correspondent aux renommages dont les effets doivent être intégrés à l'opération (lignes 19-25).

À ce stade, le noeud obtient la liste des identifiants à insérer à l'époque courante. Il peut alors réutiliser la fonction INSERT pour les intégrer à son état. Pour minimiser le nombre de parcours de la séquence, le noeud aggrège les identifiants en intervalles d'identifiants au préalable à l'aide de la fonction AGGREGATE (ligne 28). Cette fonction regroupe simplement les identifiants contigus en intervalles d'identifiants et retourne la liste des intervalles obtenus.

À partir des intervalles d'identifiants obtenus et du contenu initial de l'opération *insert*, le noeud régénère une liste d'opérations *insert*. Ces opérations sont ensuite successivement intégrées à la séquence.

L'algorithme d'intégration d'une opération *remove* distante est très similaire à l'algorithme d'intégration d'une opération *insert* que nous venons de présenter. Seules les lignes permettant de récupérer les identifiants supprimés (5), de générer l'opération *remove* transformée (29) et de l'appliquer (3 et 31) diffèrent.

Intégration d'une opération *rename* distante

L'autre cas de figure que RenamableLogootSplit doit gérer est l'intégration d'une opération *rename* distante. Pour cela, RenamableLogootSplit repose sur l'algorithme présenté dans l'Algorithme 4.

Comme précédemment, nous utilisons l'exemple illustré dans la Figure 2.12 pour présenter le fonctionnement de cet algorithme.

La Figure 2.12 reprend le scénario décrit précédemment dans la Figure 2.11. Elle complète ce dernier en faisant apparaître la réception de l'opération *rename* vers l'époque ε_{B2} par le noeud A. C'est sur ce point que nous allons nous focaliser ici.

À la réception de l'opération *rename* vers l'époque ε_{B2} , le noeud A utilise RENREMOTE pour intégrer cette opération. Tout d'abord, le noeud A ajoute l'époque ε_{B2} et les métadonnées associées (ancien état, auteur de l'opération *rename*, numéro de séquence de l'auteur de l'opération *rename*) à son propre arbre des époques (ligne 6).

Le noeud compare ensuite l'époque introduite (ε_{B2}) à son époque courante (ε_{A2}) en utilisant la relation $<_{\varepsilon}$. Si l'époque introduite était plus petite que l'époque courante, aucun traitement supplémentaire ne serait nécessaire. RENREMOTE se contenterait de renvoyer comme résultats la séquence et l'époque courante, inchangées, et le nouvel *arbre des époques* (ligne 9).

Algorithme 4 Algorithme d'intégration d'une opération *rename* distante

```

function RENREMOTE(seq, epochTree, currentEpoch, renOp)
  opEpoch  $\leftarrow$  getEpoch(renOp)
  renamedIds  $\leftarrow$  getRenamedIds(renOp)
  introducedEpoch  $\leftarrow$  getIntroducedepoch(renOp)

5:   newEpochTree  $\leftarrow$  addEpoch(epochTree, introducedEpoch, opEpoch, renamedIds)

  if introducedEpoch  $<_{\varepsilon}$  currentEpoch then
    return  $\langle$ seq, newEpochTree, currentEpoch $\rangle$ 
10:  else
    idIntervals  $\leftarrow$  getIdIntervals(seq)
    ids  $\leftarrow$  flatMap(idIntervals, expand)

     $\langle$ epochsToRevert, epochsToApply $\rangle \leftarrow$  getPathBetweenEpochs(newEpochTree, currentEpoch,
    introducedEpoch)
15:    for epoch in epochsToRevert do
      renamedIds  $\leftarrow$  getRenamedIds(newEpochTree, epoch)
      nId  $\leftarrow$  getNodeId(newEpochTree, epoch)
      nSeq  $\leftarrow$  getNodeSeq(newEpochTree, epoch)
20:      revertRenameIdpartial  $\leftarrow$  papply(revertRenameId, renamedIds, nId, nSeq)
      ids  $\leftarrow$  map(ids, revertRenameIdpartial)
    end for

    for epoch in epochsToApply do
25:      renamedIds  $\leftarrow$  getRenamedIds(newEpochTree, epoch)
      nId  $\leftarrow$  getNodeId(newEpochTree, epoch)
      nSeq  $\leftarrow$  getNodeSeq(newEpochTree, epoch)
      renameIdpartial  $\leftarrow$  papply(renameId, renamedIds, nId, nSeq)
      ids  $\leftarrow$  map(ids, renameIdpartial)
30:    end for

    nId  $\leftarrow$  getNodeId(seq)
    nSeq  $\leftarrow$  getNodeSeq(seq)
    newIdIntervals  $\leftarrow$  aggregate(ids)
35:    content  $\leftarrow$  getContent(seq)
    blocks  $\leftarrow$  generateBlocks(newIdIntervals, content)
    newSeq  $\leftarrow$  new LogootSplit(nId, nSeq, blocks)

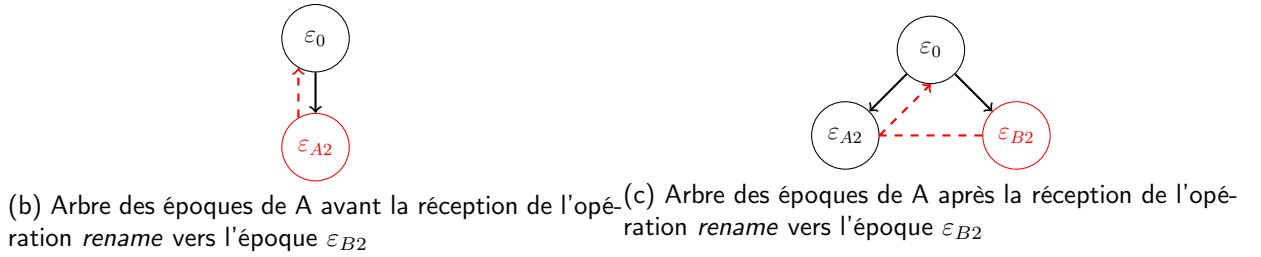
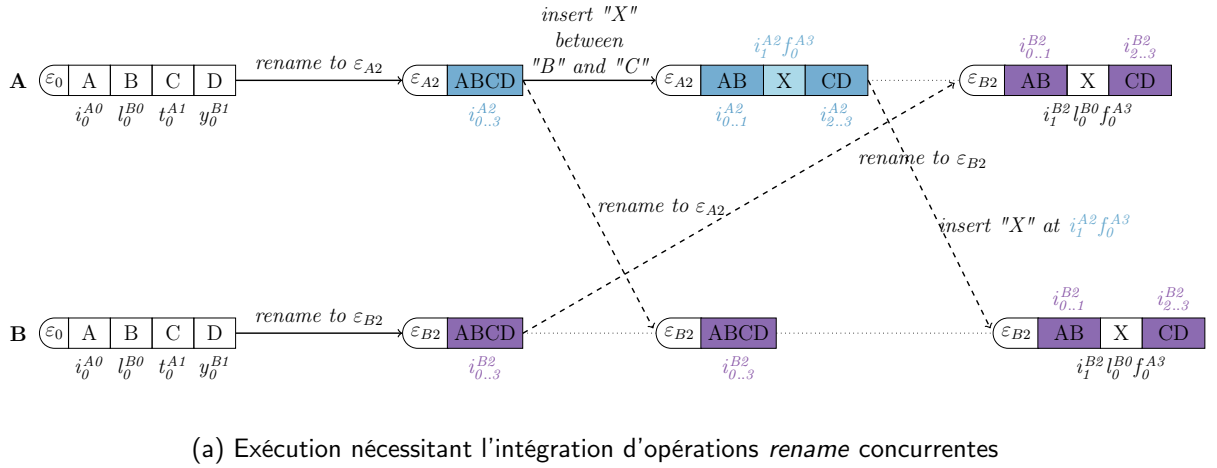
    return  $\langle$ newSeq, newEpochTree, introducedEpoch $\rangle$ 
40:  end if
end function

```

Dans le cas présent, on a $\varepsilon_{A2} <_{\varepsilon} \varepsilon_{B2}$. ε_{B2} devient donc la nouvelle époque courante. Le noeud A procède au renommage de son état vers cette nouvelle époque.

Pour cela, le noeud récupère l'ensemble des identifiants formant son état courant (lignes 11-12). Puis, comme dans INSREMOTE, le noeud récupère le chemin entre son époque courante et l'époque cible à l'aide de GETPATHBETWEENEPOCHS puis renomme chaque identifiant à travers les différents époques (lignes 16-30).

Le noeud obtient alors la liste des identifiants courant, à la nouvelle époque cible. Il ne

2.3. Gestion des opérations *rename* concurrentesFIGURE 2.12 – Intégration d'une opération *rename* distante

lui reste plus qu'à construire une nouvelle séquence à partir de ces identifiants. Pour cela, le noeud regénère des blocs à partir des intervalles d'identifiants obtenus et du contenu de la séquence courante. Le noeud utilise ensuite ces données pour instancier une nouvelle séquence équivalente à l'époque cible (ligne 37). Finalement, `RENREMOTE` renvoie cette nouvelle séquence, la nouvelle époque courante ainsi que le nouvel *arbre des époques*.

2.3.5 Règles de récupération de la mémoire des états précédents

Les noeuds stockent les époques et les *anciens états* correspondant pour transformer les identifiants d'une époque à l'autre. Au fur et à mesure que le système progresse, certaines époques et métadonnées associées deviennent obsolètes puisque plus aucune opération ne peut être émise depuis ces époques. Les noeuds peuvent alors supprimer ces époques. Dans cette section, nous présentons un mécanisme permettant aux noeuds de déterminer les époques obsolètes.

Pour proposer un tel mécanisme, nous nous reposons sur la notion de *stabilité causale des opérations* [10]. Une opération est causalement stable une fois qu'elle a été délivrée à tous les noeuds. Dans le contexte de l'opération *rename*, cela implique que tous les noeuds ont progressé à l'époque introduite par cette opération ou à une époque plus grande d'après la relation *priority*. À partir de ce constat, nous définissons les *potentielles époques courantes* :

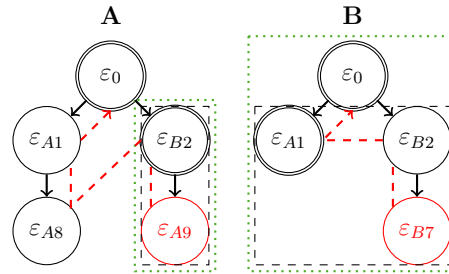
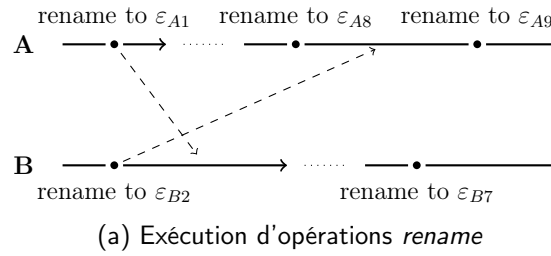
Définition 20 (Potentielles époques courantes) *L'ensemble des époques auxquelles les noeuds peuvent se trouver actuellement et à partir desquelles ils peuvent émettre des*

opérations, du point de vue du noeud courant. Il s'agit d'un sous-ensemble de l'ensemble des époques, composé de l'époque maximale introduite par une opération *rename* causalement stable et de toutes les époques plus grande que cette dernière d'après la relation *priority*.

Pour traiter les prochaines opérations, les noeuds doivent maintenir les chemins entre toutes les époques de l'ensemble des *potentielles époques courantes*. Nous appelons *époques requises* l'ensemble des époques correspondant.

Définition 21 (Époques requises) *L'ensemble des époques qu'un noeud doit conserver pour traiter les potentielles prochaines opérations. Il s'agit de l'ensemble des époques qui forment les chemins entre chaque époque appartenant à l'ensemble des potentielles époques courantes et leur Plus Petit Ancêtre Commun (PPAC).*

Il s'ensuit que toute époque qui n'appartient pas à l'ensemble des *époques requises* peut être retirée par les noeuds. La Figure 2.13 illustre un cas d'utilisation du mécanisme de récupération de mémoire proposé.



(b) Arbres des époques respectifs avec les ensembles *potentielles époques courantes* et *époques requises* illustrés

FIGURE 2.13 – Suppression des époques obsolètes et récupération de la mémoire des anciens états associés

Dans la Figure 2.13a, nous représentons une exécution au cours de laquelle deux noeuds A et B génère respectivement plusieurs opérations *rename*. Dans la Figure 2.13b, nous représentons les *arbre des époques* respectifs de chaque noeud. Les époques introduites par des opérations *rename* causalement stables sont représentées en utilisant des doubles cercles. L'ensemble des *potentielles époques courantes* est montré sous la forme d'un rectangle noir tireté, tandis que l'ensemble des *époques requises* est représenté par un rectangle vert pointillé.

Le noeud A génère tout d'abord une opération *rename* vers ε_{A1} et ensuite une opération *rename* vers ε_{A8} . Il reçoit ensuite une opération *rename* du noeud B qui introduit ε_{B2} .

Puisque ε_{B2} est plus grand que son époque courante actuelle ($\varepsilon_{e0}\varepsilon_{A1}\varepsilon_{A8} < \varepsilon_{e0}\varepsilon_{B2}$), le noeud A la sélectionne comme sa nouvelle époque cible et procède au renommage de son état en conséquence. Finalement, le noeud A génère une troisième opération *rename* vers ε_{A9} .

De manière concurrente, le noeud B génère l'opération *rename* vers ε_{B2} . Il reçoit ensuite l'opération *rename* vers ε_{A1} du noeud A. Cependant, le noeud B conserve ε_{B2} comme époque courante (puisque $\varepsilon_{e0}\varepsilon_{A1} < \varepsilon_{e0}\varepsilon_{B2}$). Après, le noeud B génère une autre opération *rename* vers ε_{B7} .

À la livraison de l'opération *rename* introduisant l'époque ε_{B2} au noeud A, cette opération devient causalement stable. À partir de ce point, le noeud A sait que tous les noeuds ont progressé jusqu'à cette époque ou une plus grande d'après la relation *priority*. Les époques ε_{B2} et ε_{A9} forment donc l'ensemble des *potentielles époques courantes* et les noeuds peuvent seulement émettre des opérations depuis ces époques ou une de leur descendante encore inconnue. Le noeud A procède ensuite au calcul de l'ensemble des *époques requises*. Pour ce faire, il détermine le PPAC des *potentielles époques courantes* : ε_{B2} . Il génère ensuite l'ensemble des *époques requises* en ajoutant toutes les époques formant les chemins entre ε_{B2} et les *potentielles époques courantes*. Les époques ε_{B2} et ε_{A9} forment donc l'ensemble des *époques requises*. Le noeud A déduit que les époques ε_0 , ε_{A1} et ε_{A8} peuvent être supprimées de manière sûre.

À l'inverse, la livraison de l'opération *rename* vers ε_{A1} au noeud B ne lui permet pas de supprimer la moindre métadonnée. À partir de ses connaissances, le noeud B calcule que ε_{A1} , ε_{B2} et ε_{B7} forment l'ensemble des *potentielles époques courantes*. De cette information, le noeud B détermine que ces époques et leur PPAC forment l'ensemble des *époques requises*. Toute époque connue appartient donc à l'ensemble des *époques requises*, empêchant leur suppression.

À terme, une fois que le système devient inactif, les noeuds atteignent la même époque et l'opération *rename* correspondante devient causalement stable. Les noeuds peuvent alors supprimer toutes les autres époques et métadonnées associées, supprimant ainsi le surcoût mémoire introduit par le mécanisme de renommage.

Notons que le mécanisme de récupération de mémoire peut être simplifié dans les systèmes empêchant les opérations *rename* concurrentes. Puisque les époques forment une chaîne dans de tels systèmes, la dernière époque introduite par une opération *rename* causalement stable devient le PPAC des *potentielles époques courantes*. Il s'ensuit que cette époque et ses descendantes forment l'ensemble des *époques requises*. Les noeuds n'ont donc besoin que de suivre les opérations *rename* causalement stables pour déterminer quelles époques peuvent être supprimées dans les systèmes sans opérations *rename* concurrentes.

Pour déterminer qu'une opération *rename* donnée est causalement stable, les noeuds doivent être conscients des autres et de leur avancement. Un protocole de gestion de groupe tel que [19, 17] est donc requis.

La stabilité causale peut prendre un certain temps à être atteinte. En attendant, les noeuds peuvent néanmoins décharger les anciens états sur le disque dur puisqu'ils ne sont seulement nécessaires que pour traiter les opérations concurrentes aux opérations *rename*. Nous approfondissons ce sujet dans la sous-section 2.5.2.

2.4 Validation

2.4.1 Complexité en temps des opérations

Afin d'évaluer `RenamableLogootSplit`, nous analysons tout d'abord la complexité en temps de ses opérations. Ces complexités dépendent de plusieurs paramètres : nombre d'identifiants et de blocs stockés au sein de la structure, taille des identifiants, structures de données utilisées...

Hypothèses

Afin d'établir les valeurs de complexité des différentes opérations, nous prenons les hypothèses suivantes vis-à-vis des paramètres. Nous supposons que le nombre n d'identifiants présents dans la séquence a tendance à croître, c.-à-d. que plus d'insertions sont effectuées que de suppressions. Nous considérons que la taille des identifiants, qui elle croît avec le nombre d'insertions mais qui est réinitialisée à chaque renommage, devient négligeable par rapport au nombre d'identifiants. Nous ne prenons donc pas en considération ce paramètre dans nos complexités et considérons que les manipulations d'identifiants (comparaison, génération) s'effectuent en temps constant. Afin de simplifier les complexités, nous considérons que les *anciens états* associés aux époques contiennent aussi n identifiants. Finalement, nous considérons que nous utilisons comme structures de données un arbre AVL pour représenter l'état interne de la séquence, des tableaux pour les *anciens états* et une table de hachage pour l'*arbre des époques*.

Complexité en temps des opérations *insert* et *remove*

À partir de ces hypothèses, nous établissons les complexités en temps des opérations. Pour chaque opération, nous distinguons deux complexités : une complexité pour l'intégration de l'opération locale, une pour l'intégration de l'opération distante.

La complexité de l'intégration de l'opération *insert* locale est inchangée par rapport à celle obtenue pour `LogootSplit`. Son intégration consiste toujours à déterminer entre quels identifiants se situe les nouveaux éléments insérés, à générer de nouveaux identifiants correspondants à l'ordre souhaité puis à insérer le bloc dans l'arbre AVL. D'après ANDRÉ et al. [6], nous obtenons donc une complexité de $\mathcal{O}(\log b)$ pour cette opération locale, où b représente le nombre de blocs dans la séquence.

La complexité de l'intégration de l'opération *insert* distante, elle, évolue par rapport à celle définie pour `LogootSplit`. Comme indiqué dans la section 2.3.4, plusieurs étapes se rajoutent au processus d'intégration de l'opération notamment dans le cas où celle-ci provient d'une autre époque que l'époque courante.

Tout d'abord, il est nécessaire d'identifier l'époque PPAC entre l'époque de l'opération et l'époque courante. L'algorithme correspondant consiste à déterminer la première intersection entre deux branches de l'*arbre des époques*. Cette étape peut être effectuée en $\mathcal{O}(h)$, où h représente la hauteur de l'*arbre des époques*.

L'obtention de l'époque PPAC entre l'époque de l'opération et l'époque courante permet de déterminer les k renommages dont les effets doivent être retirés de l'opération et les

l renommages dont les effets doivent être intégrés à l'opération. Le noeud intégrant l'opération procède ainsi aux k inversions de renommages successives puis aux l application de renommages, et ce pour tous les s identifiants insérés par l'opération.

Pour retirer les effets des renommages à inverser, le noeud intégrant l'opération utilise REVERTRENAMEID. Cet algorithme retourne pour un identifiant donné un nouvel identifiant correspondant à l'époque précédente. Pour cela, REVERTRENAMEID utilise le prédécesseur et le successeur de l'identifiant donné dans l'*ancien état* renommé. Pour retrouver ces deux identifiants au sein de l'*ancien état*, REVERTRENAMEID utilise l'offset du premier tuple de l'identifiant donné. Par définition, cet élément correspond à l'index du prédécesseur de l'identifiant donné dans l'*ancien état*. Aucun parcours de l'*ancien état* n'est nécessaire. Le reste de REVERTRENAMEID consistant en des comparaisons et manipulations d'identifiants, nous obtenons que REVERTRENAMEID s'effectue en $\mathcal{O}(1)$.

Pour inclure les effets des renommages à appliquer, le noeud utilise ensuite RENAMEID. De manière similaire à REVERTRENAMEID, RENAMEID génère pour un identifiant donné un nouvel identifiant équivalent à l'époque suivante en se basant sur son prédécesseur. Cependant, il est nécessaire ici de faire une recherche pour déterminer le prédécesseur de l'identifiant donné dans l'*ancien état*. L'*ancien état* étant un tableau trié d'identifiants, il est possible de procéder à une recherche dichotomique. Cela permet de trouver le prédécesseur en $\mathcal{O}(\log n)$, où n correspond ici au nombre d'identifiants composant l'*ancien état*. Comme pour REVERTRENAMEID, les instructions restantes consistent en des comparaisons et manipulations d'identifiants. La complexité de RENAMEID est donc de $\mathcal{O}(\log n)$.

Une fois les identifiants introduits par l'opération *insert* renommés pour l'époque courante, il ne reste plus qu'à les insérer dans la séquence. Cette étape se réalise en $\mathcal{O}(\log b)$ pour chaque identifiant, le temps nécessaire pour trouver son emplacement dans l'arbre AVL.

Ainsi, en reprenant l'ensemble des étapes composant l'intégration de l'opération *insert* distante, nous obtenons la complexité suivante : $\mathcal{O}(h + s(k + l \cdot \log n + \log b))$.

Le procédé de l'intégration de l'opération *remove* étant similaire à celui de l'opération *insert*, aussi bien en local qu'en distant, nous obtenons les mêmes complexités en temps.

Complexité en temps de l'opération *rename*

Étudions à présent la complexité en temps de l'opération *rename*.

L'opération *rename* locale se décompose en 2 étapes : (i) La génération de l'*ancien état* à intégrer au message de l'opération (cf. Définition 15) (ii) Le remplacement de la séquence courante par une séquence équivalente, renommée. La première étape consiste à parcourir et à linéariser la séquence actuelle pour en extraire les intervalles d'identifiants. Elle s'effectue donc en $\mathcal{O}(b)$. La seconde consiste à instancier une nouvelle séquence vide, et à y insérer un bloc qui associe le contenu actuel de la séquence à l'intervalle d'identifiants $pos_{0..n-1}^{nodeId\ nodeSeq}$, avec pos la position du premier tuple du premier id de l'état, $nodeId$ et $nodeSeq$ l'identifiant et le numéro de séquence actuel du noeud et n la taille du contenu. Cette seconde étape s'effectue en $\mathcal{O}(1)$. L'opération *rename* locale a donc une complexité de $\mathcal{O}(b)$.

L'intégration de l'opération *rename* se décompose en les étapes suivantes : (i) L'insertion de la nouvelle époque et de l'*ancien état* associé dans l'*arbre des époques* (ii) La

récupération des n identifiants formant l'état courant (iii) Le calcul de l'époque PPAC entre l'époque courante et l'époque cible (iv) L'identification des k opérations *rename* à inverser et des l opérations *rename* à jouer (v) Le renommage de chacun des identifiants à l'aide de REVERTRENAMEID et RENAMEID (vi) L'insertion de chacun des identifiants renommés dans une nouvelle séquence L'*arbre des époques* étant représenté à l'aide d'une table de hachage, la première étape s'effectue en $\mathcal{O}(1)$. La seconde étape nécessite elle de parcourir l'arbre AVL et de convertir chaque intervalle d'identifiants en liste d'identifiants, ce qui nécessite $\mathcal{O}(n)$ instructions.

Les étapes (iii) à (vi) peuvent être effectuées en réutilisant pour chaque identifiant l'algorithme pour l'intégration d'opérations *insert* distantes analysé précédemment. Ces étapes s'effectuent donc en $\mathcal{O}(n(k + l \cdot \log n + \log b))$.

Nous obtenons donc une complexité en temps de $\mathcal{O}(h + n(k + l \cdot \log n + \log b))$ pour l'intégration de l'opération *rename* distante.

Nous pouvons néanmoins améliorer ce premier résultat. Notamment, nous pouvons tirer parti des faits suivants : (i) Le fonctionnement de RENAMEID repose sur l'utilisation de l'identifiant prédecesseur comme préfixe (ii) Les identifiants de l'état courant et de l'*ancien état* forment tous deux des listes triées. Ainsi, plutôt que d'effectuer une recherche dichotomique sur l'*ancien état* pour trouver le prédecesseur de l'identifiant à renommer, nous pouvons parcourir les deux listes en parallèle. Ceci nous permet de renommer l'intégralité des identifiants en un seul parcours de l'état courant et de l'*ancien état*, c.-à-d. en $\mathcal{O}(n)$ instructions. Ensuite, plutôt que d'insérer les identifiants un à un dans la nouvelle séquence, nous pouvons recomposer au préalable les différents blocs en parcourant la liste des identifiants et en les agrégeant au fur et à mesure. Il ne reste plus qu'à constituer la nouvelle séquence à partir des blocs obtenus. Ces actions s'effectuent respectivement en $\mathcal{O}(n)$ et $\mathcal{O}(b)$ instructions.

Ainsi, ces améliorations nous permettent d'obtenir une complexité en temps en $\mathcal{O}(h + n(k + l) + b)$ pour le traitement de l'opération *rename* distante.

Récapitulatif

Nous récapitulons les complexités en temps présentées précédemment dans le Tableau 2.1.

Complexité en temps du mécanisme de récupération de mémoire des époques

Pour compléter notre analyse théorique des performances de RenamableLogootSplit, nous proposons une analyse en complexité en temps du mécanisme présenté en sous-section 2.3.5 qui permet de supprimer les époques devenues obsolètes et de récupérer la mémoire occupée par leur *ancien état* respectif.

L'algorithme du mécanisme de récupération de la mémoire se compose des étapes suivantes. Tout d'abord, il établit le vecteur de version des opérations causalement stables. Pour cela, chaque noeud doit maintenir une matrice des vecteurs de version de tous les noeuds. L'algorithme génère le vecteur de version des opérations causalement stable en récupérant pour chaque noeud la valeur minimale qui y est associée dans la matrice des

TABLE 2.1 – Complexité en temps des différentes opérations

Type d'opération	Complexité en temps	
	Locale	Distante
<i>insert</i>	$\log b$	$h + s(k + l \cdot \log n + \log b)$
<i>remove</i>	$\log b$	$h + s(k + l \cdot \log n + \log b)$
<i>naïve rename</i>	b	$h + n(k + l \cdot \log n + \log b)$
<i>rename</i>	b	$h + n(k + l) + b$

b : nombre de blocs, n : nombre d'éléments de l'état courant et des *anciens états*, h : hauteur de l'*arbre des époques*, k : nombre de renommages à inverser, l : nombre de renommages à appliquer, s : nombre d'éléments insérés/supprimés par l'opération

vecteurs de version. Cette étape correspond à fusionner n vecteurs de version contenant n entrées, elle s'exécute donc en $\mathcal{O}(n^2)$ instructions.

La seconde étape consiste à parcourir l'arbre des époques de manière inverse à l'ordre défini par la relation *priority*. Ce parcours s'effectue jusqu'à trouver l'époque maximale causalement stable, c.-à-d. la première époque pour laquelle l'opération *rename* associée est causalement stable. Pour chaque époque parcourue, le mécanisme de récupération de mémoire calcule et stocke son chemin jusqu'à la racine. Cette étape s'exécute donc en $\mathcal{O}(e \cdot h)$, avec e le nombre d'époques composant l'arbre des époques et h la hauteur de l'arbre.

À partir de ces chemins, le mécanisme calcule l'époque PPAC. Pour ce faire, l'algorithme calcule de manière successive la dernière intersection entre le chemin de la racine jusqu'à l'époque PPAC courante et les chemins précédemment calculés. L'époque PPAC est la dernière époque du chemin résultant. Cette étape s'exécute aussi en $\mathcal{O}(e \cdot h)$.

L'algorithme peut alors calculer l'ensemble des *époques requises*. Pour cela, il parcourt les chemins calculés au cours de la seconde étape. Pour chaque chemin, il ajoute les époques se trouvant après l'époque PPAC à l'ensemble des *époques requises*. De nouveau, cette étape s'exécute en $\mathcal{O}(e \cdot h)$.

Après avoir déterminé l'ensemble des *époques requises*, le mécanisme peut supprimer les époques obsolètes. Il parcourt l'arbre des époques et supprime toute époque qui n'appartient pas à cet ensemble. Cette étape finale s'exécute en $\mathcal{O}(e)$.

Ainsi, nous obtenons que la complexité en temps du mécanisme de récupération de mémoire des époques est en $\mathcal{O}(n^2 + e \cdot h)$. Nous récapitulons ce résultat dans Tableau 2.2.

Malgré sa complexité en temps, le mécanisme de récupération de mémoire des époques devrait avoir un impact limité sur les performances de l'application. En effet, ce mécanisme n'appartient pas au chemin critique de l'application, c.-à-d. l'intégration des modifications. Il peut être déclenché occasionnellement, en tâche de fond. Nous pouvons même viser des fenêtres spécifiques pour le déclencher, e.g. pendant les périodes d'inactivité. Ainsi, nous avons pas étudié plus en détails cette partie de RenamableLogootSplit dans le cadre de cette thèse. Des améliorations de ce mécanisme doivent donc être possibles.

TABLE 2.2 – Complexité en temps du mécanisme de récupération de mémoire des époques

Étape	Temps
<i>calculer le vecteur de version des opérations causalement stables</i>	n^2
<i>calculer les chemins de la racine aux</i> potentielles époques courantes	$e \cdot h$
<i>identifier le PPAC</i>	$e \cdot h$
<i>calculer l'ensemble des époques requises</i>	$e \cdot h$
<i>supprimer les époques obsolètes</i>	e
<i>total</i>	$n^2 + e \cdot h$

n : nombre de noeuds du système, e : nombre d'époques dans l'arbre des époques, h : hauteur de l'arbre des époques

2.4.2 Expérimentations

Afin de valider l'approche que nous proposons, nous avons procédé à une évaluation expérimentale. Les objectifs de cette évaluation étaient de mesurer (i) le surcoût mémoire de la séquence répliquée (ii) le surcoût en calculs ajouté aux opérations *insert* et *remove* par le mécanisme de renommage (iii) le coût d'intégration des opérations *rename*.

Par le biais de simulations, nous avons généré le jeu de données utilisé par nos benchmarks. Ces simulations suivent le scénario suivant.

Scénario d'expérimentation

Le scénario reproduit la rédaction d'un article par plusieurs pairs de manière collaborative, en temps réel. La collaboration ainsi décrite se décompose en 2 phases.

Dans un premier temps, les pairs spécifient principalement le contenu de l'article. Quelques opérations *remove* sont tout même générées pour simuler des fautes de frappes. Une fois que le document atteint une taille critique (définie de manière arbitraire), les pairs passent à la seconde phase de la collaboration. Lors de cette seconde phase, les pairs arrêtent d'ajouter du nouveau contenu mais se concentrent à la place sur la reformulation et l'amélioration du contenu existant. Ceci est simulé en équilibrant le ratio entre les opérations *insert* et *remove*.

Chaque pair doit émettre un nombre donné d'opérations *insert* et *remove*. La simulation prend fin une fois que tous les pairs ont reçu toutes les opérations. Pour suivre l'évolution de l'état des pairs, nous prenons des instantanés de leur état à plusieurs points donnés de la simulation.

Implémentation des simulations

Nous avons effectué nos simulations avec les paramètres expérimentaux suivants : nous avons déployé 10 bots à l'aide de conteneurs Docker sur une même machine. Chaque conteneur correspond à un processus Node.js mono-threadé et permet de simuler un pair. Les bots sont connectés entre eux par le biais d'un réseau P2P maillé entièrement connecté. Enfin, ils partagent et éditent le document de manière collaborative en utilisant soit LogootSplit soit RenamableLogootSplit en fonction des paramètres de la session.

Toutes les 200 ± 50 ms, chaque bot génère localement une opération *insert* ou *remove* et la diffuse immédiatement aux autres noeuds. Au cours de la première phase, la probabilité d'émettre une opération *insert* (resp. *remove*) est de 80% (resp. 20%). Une fois que leur copie locale du document atteint 60k caractères (environ 15 pages), les bots basculent à la seconde phase et redéfinissent chaque probabilité à 50%. De plus, tout au long de la collaboration, les bots ont une probabilité de 5% de déplacer leur curseur à une position aléatoire dans le document après chaque opération locale.

Chaque bot doit générer 15k opérations *insert* ou *remove*, et s'arrête donc une fois qu'il a observé les 150k opérations. Pour chaque bot, nous enregistrons un instantané de son état toutes les 10k opérations observées. Nous enregistrons aussi son log des opérations à l'issue de la simulation.

De plus, dans le cas de RenamableLogootSplit, 1 à 4 bots sont désignés de façon arbitraire comme des *renaming bots* en fonction de la session. Les *renaming bots* génèrent des opérations *rename* toutes les 7.5k ou toutes les 30k opérations qu'ils observent, en fonction des paramètres de la simulation. Ces opérations *rename* sont générées de manière à assurer qu'elles soient concurrentes.

Dans un but de reproductibilité, nous avons mis à disposition notre code, nos benchmarks et les résultats à l'adresse suivante : <https://github.com/coast-team/mute-bot-random/>.

2.4.3 Résultats

En utilisant les instantanés et les logs d'opérations générés, nous avons effectué plusieurs benchmarks. Ces benchmarks évaluent les performances de RenamableLogootSplit et les comparent à celles de LogootSplit. Sauf mention contraire, les benchmarks utilisent les données issues des simulations au cours desquelles les opérations *rename* étaient générées toutes les 30k opérations. Les résultats sont présentés et analysés ci-dessous.

Convergence

Nous avons tout d'abord vérifié la convergence de l'état des noeuds à l'issue des simulations. Pour chaque simulation, nous avons comparé l'état final de chaque noeud à l'aide de leur instantanés respectifs. Nous avons pu confirmer que les noeuds convergaient sans aucune autre forme de communication que les opérations, satisfaisant donc le modèle de la SEC.

Ce résultat établit un premier jalon dans la validation de la correction de RenamableLogootSplit. Il n'est cependant qu'empirique. Des travaux supplémentaires pour prouver formellement sa correction doivent être entrepris.

Consommation mémoire

Nous avons ensuite procédé à l'évaluation de l'évolution de la consommation mémoire du document au cours des simulations, en fonction du CRDT utilisé et du nombre de *renaming bots*. Nous présentons les résultats obtenus dans la Figure 2.14.

Pour chaque graphique dans la Figure 2.14, nous représentons 4 données différentes. La ligne tiretée bleue correspond à la taille du contenu du document, c.-à-d. du texte,

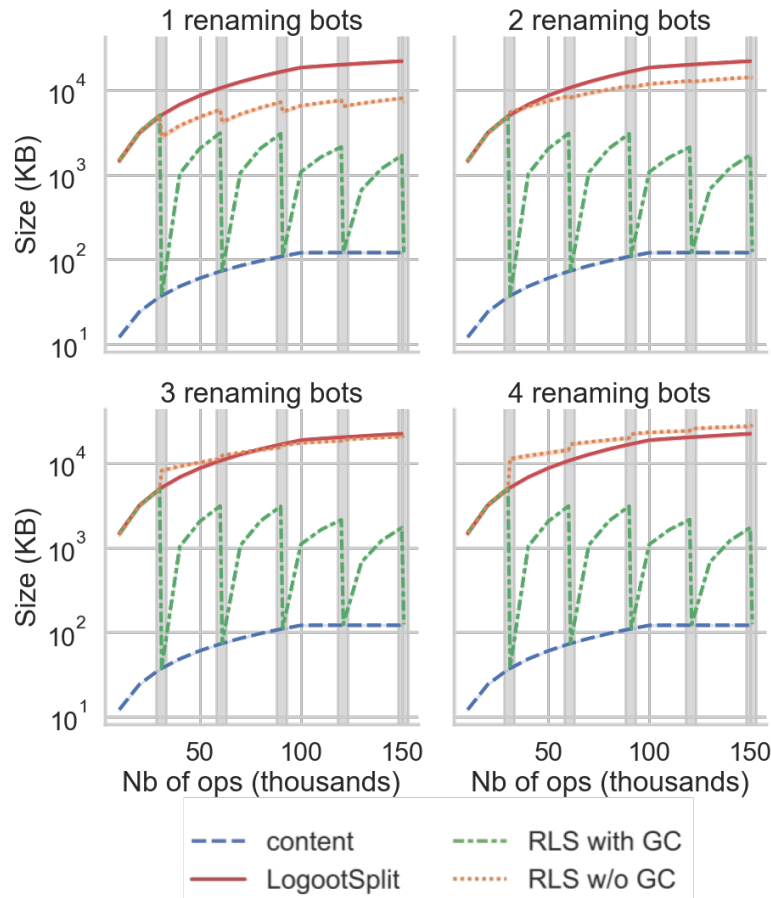


FIGURE 2.14 – Évolution de la taille du document en fonction du CRDT utilisé et du nombre de *renaming bots* dans la collaboration

tandis que la ligne continue rouge représente la taille complète du document LogootSplit.

La ligne verte pointillée-tirée représente la taille du document RenamableLogootSplit dans son meilleur cas. Dans ce scénario, les noeuds considèrent que les opérations *rename* sont causalement stables dès qu'ils les reçoivent. Les noeuds peuvent alors bénéficier des effets du mécanisme de renommage tout en supprimant les métadonnées qu'il introduit : les *anciens états* et époques. Ce faisant, les noeuds peuvent minimiser de manière périodique le surcoût en métadonnées de la structure de données, indépendamment du nombre de *renaming bots* et d'opérations *rename* concurrentes générées.

La ligne pointillée orange représente la taille du document RenamableLogootSplit dans son pire cas. Dans ce scénario, les noeuds considèrent que les opérations *rename* ne deviennent jamais causalement stables. Les noeuds doivent alors conserver de façon permanente les métadonnées introduites par le mécanisme de renommage. Les performances de RenamableLogootSplit diminuent donc au fur et à mesure que le nombre de *renaming bots* et d'opérations *rename* générées augmente. Néanmoins, même dans ces conditions, nous observons que RenamableLogootSplit offre de meilleures performances que LogootSplit tant que le nombre de *renaming bots* reste faible (1 ou 2). Ce résultat s'explique par le fait que le mécanisme de renommage permet aux noeuds de supprimer les métadonnées

de la structure de données utilisée en interne pour représenter la séquence (c.-à-d. l'arbre AVL).

Pour récapituler les résultats présentés, le mécanisme de renommage introduit un surcoût temporaire en métadonnées qui augmente avec chaque opération *rename*. Mais ce surcoût se résorbe à terme une fois que le système devient quiescent et que les opérations *rename* deviennent causalement stables. Dans la sous-section 2.5.2, nous détaillerons l'idée que les *anciens états* peuvent être déchargés sur le disque en attendant que la stabilité causale soit atteinte pour atténuer l'impact du surcoût temporaire en métadonnées.

Temps d'intégration des opérations standards

Nous avons ensuite comparé l'évolution du temps d'intégration des opérations standards, c.-à-d. les opérations *insert* et *remove*, sur des documents LogootSplit et RenamableLogootSplit. Puisque les deux types d'opérations partagent la même complexité en temps, nous avons seulement utilisé des opérations *insert* dans nos benchmarks. Nous faisons par contre la différence entre les mises à jours *locales* et *distantes*. Conceptuellement, les modifications locales peuvent être décomposées comme présenté dans [8] en les deux étapes suivantes : (i) la génération de l'opération correspondante (ii) l'application de l'opération correspondante sur l'état local. Cependant, pour des raisons de performances, nous avons fusionné ces deux étapes dans notre implémentation. Nous distinguons donc les résultats des modifications *locales* et des modifications *distantes* dans nos benchmarks. La Figure 2.15 présente les résultats obtenus.

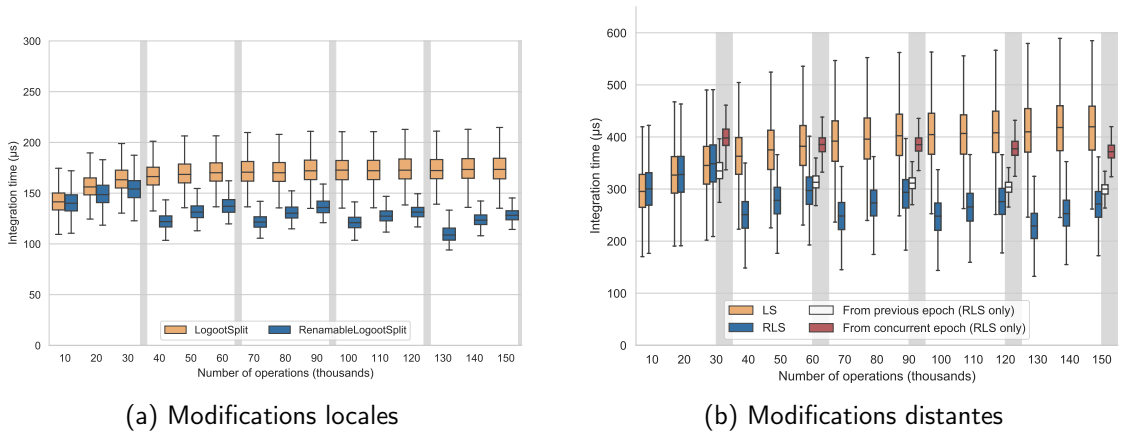


FIGURE 2.15 – Temps d'intégration des opérations standards

Dans ces figures, les boxplots oranges correspondent aux temps d'intégration sur des documents LogootSplit, les boxplots bleues sur des documents RenamableLogootSplit. Bien que les temps d'intégration soient initialement équivalents, les temps d'intégration sur des documents RenamableLogootSplit sont ensuite réduits par rapport à ceux de LogootSplit une fois que des opérations *rename* ont été intégrées. Cette amélioration s'explique par le fait que l'opération *rename* optimise la représentation interne de la séquence (c.-à-d. elle réduit le nombre de blocs stockés dans l'arbre AVL).

Dans le cadre des opérations distantes, nous avons mesuré des temps d'intégration spécifiques à `RenamableLogootSplit` : le temps d'intégration d'opérations distantes provenant d'époques *parentes* et d'époques *soeurs*, respectivement affiché sous la forme de boxplots blanches et rouges dans la Figure 2.15b.

Les opérations distantes provenant d'époques *parentes* sont des opérations générées de manière concurrente à l'opération *rename* mais appliquées après cette dernière. Puisque l'opération doit être transformée au préalable en utilisant `RENAMEID`, nous observons un surcoût computationnel par rapport aux autres opérations. Mais ce surcoût est compensé par l'optimisation de la représentation interne de la séquence effectuée par l'opération *rename*.

Concernant les opérations provenant d'époques *soeurs*, nous observons de nouveau un surcoût puisque les noeuds doivent tout d'abord annuler les effets de l'opération *rename* concurrente en utilisant `REVERTRENAMEID`. À cause de cette étape supplémentaire, les performances de `RenamableLogootSplit` pour ces opérations sont comparables à celles de `LogootSplit`.

Pour récapituler, les fonctions de transformation ajoutent un surcoût aux temps d'intégration des opérations concurrentes aux opérations *rename*. Malgré ce surcoût, `RenamableLogootSplit` offre de meilleures performances que `LogootSplit` pour intégrer ces opérations grâce aux réductions de la taille de l'état effectuées par les opérations *rename*. Cependant, cette affirmation n'est vraie que tant que la distance entre l'époque de génération de l'opération et l'époque courante du noeud reste limitée, puisque les performances de `RenamableLogootSplit` dépendent linéairement de cette dernière (cf. Tableau 2.1). Néanmoins, ce surcoût ne concerne que les opérations concurrentes aux opérations *rename*. Il ne concerne pas la majorité des opérations, c.-à-d. les opérations générées entre deux séries d'opérations *rename*. Ces opérations, elles, ne souffrent d'aucun surcoût tout en bénéficiant des réductions de taille de l'état.

Temps d'intégration de l'opération de renommage

Finalement, nous avons mesuré l'évolution du temps d'intégration de l'opération *rename* en fonction du nombre d'opérations émises précédemment, c.-à-d. en fonction de la taille de l'état. Comme précédemment, nous distinguons les performances des modifications *locales* et *distantes*.

Nous rappelons que le traitement d'une opération *rename* dépend de l'ordre défini par la relation *priority* entre l'époque qu'elle introduit et l'époque courante du noeud qui intègre l'opération. Le cas des opérations *rename* distantes se décompose donc en trois catégories. Les opérations *distantes directes* désignent les opérations *rename* distantes qui introduisent une nouvelle époque *enfant* de l'époque courante du noeud. Les opérations *concurrentes introduisant une plus grande* (resp. *petite*) *époque* désignent les opérations *rename* qui introduisent une époque *soeur* de l'époque courante du noeud. D'après la relation *priority*, l'époque introduite est plus grande (resp. petite) que l'époque courante du noeud. Les résultats obtenus sont présentés dans le Tableau 2.3.

Le principal résultat de ces mesures est que les opérations *rename* sont particulièrement coûteuses quand comparées aux autres types d'opérations. Les opérations *rename* locales s'intègrent en centaines de millisecondes tandis que les opérations *distantes di-*

TABLE 2.3 – Temps d’intégration de l’opération *rename*

Paramètres		Temps d’intégration (ms)					
Type	Nb Ops (k)	Moyenne	Médiane	IQR	1 ^{er} Percent.	99 ^{ème} Percent.	
Locale	30	41.8	38.7	5.66	37.3	71.7	
	60	78.3	78.2	1.58	76.2	81.4	
	90	119	119	2.17	116	124	
	120	144	144	3.24	139	149	
	150	158	158	3.71	153	164	
Distante directe	30	481	477	15.2	454	537	
	60	982	978	28.9	926	1073	
	90	1491	1482	58.8	1396	1658	
	120	1670	1664	41	1568	1814	
	150	1694	1676	60.6	1591	1853	
Cc. int. plus grande époque	30	644	644	16.6	620	683	
	60	1318	1316	26.5	1263	1400	
	90	1998	1994	46.6	1906	2112	
	120	2240	2233	54	2144	2368	
	150	2242	2234	63.5	2139	2351	
Cc. int. plus petite époque	30	1.36	1.3	0.038	1.22	3.53	
	60	2.82	2.69	0.476	2.43	4.85	
	90	4.45	4.23	1.1	3.69	5.81	
	120	5.33	5.1	1.34	4.42	8.78	
	150	5.53	5.26	1.05	4.84	8.7	

rectes et *concurrentes* introduisant une *plus grande époque* s’intègrent en secondes lorsque la taille du document dépasse 40k éléments. Ces résultats s’expliquent facilement par la complexité en temps de l’opération *rename* qui dépend supra-linéairement du nombre de blocs et d’éléments stockés dans l’état (cf. Tableau 2.1). Il est donc nécessaire de prendre en compte ce résultat et de (i) concevoir des stratégies de génération des opérations *rename* pour éviter d’impacter négativement l’expérience utilisateur (ii) proposer des versions améliorées des algorithmes `RENAMEID` et `REVERTRENAMEID` pour réduire ces temps d’intégration :

- Au lieu d’utiliser `RENAMEID`, qui renomme l’état identifiant par identifiant, nous pourrions définir et utiliser `RENAMEBLOCK`. Cette fonction permettrait de renommer l’état bloc par bloc, offrant ainsi une meilleur complexité en temps. De plus, puisque les opérations *rename* fusionnent les blocs existants en un unique bloc, `RENAMEBLOCK` permettrait de mettre en place un cercle vertueux où chaque opération *rename* réduirait le temps d’exécution de la suivante.
- Puisque chaque appel à `REVERTRENAMEID` et `REVERTRENAMEID` est indépendant des autres, ces fonctions sont adaptées à la programmation parallèle. Au lieu de renommer les identifiants (ou blocs) de manière séquentielle, nous pourrions diviser la séquence en plusieurs parties et les renommer en parallèle.

Un autre résultat intéressant de ces benchmarks est que les opérations *concurrentes* introduisant une *plus petite époque* sont rapides à intégrer. Puisque ces opérations introduisent une époque qui n’est pas sélectionnée comme nouvelle époque cible, les noeuds ne procèdent pas au renommage de leur état. L’intégration des opérations *concurrentes* introduisant une *plus petite époque* consiste simplement à ajouter l’époque introduite et l’ancien état correspondant à l’*arbre des époques*. Les noeuds peuvent donc réduire de ma-

nière significative le coût d'intégration d'un ensemble d'opérations *rename* concurrentes en les appliquant dans l'ordre le plus adapté en fonction du contexte. Nous développons ce sujet dans la sous-section 2.5.5.

Temps pour rejouer le log d'opérations

Afin de comparer les performances de RenamableLogootSplit et de LogootSplit de manière globale, nous avons mesuré le temps nécessaire pour un nouveau noeud pour rejouer l'entière du log d'opérations d'une session de collaboration, en fonction du nombre de *renaming bots* de la session. Nous présentons les résultats obtenus dans la Figure 2.16.

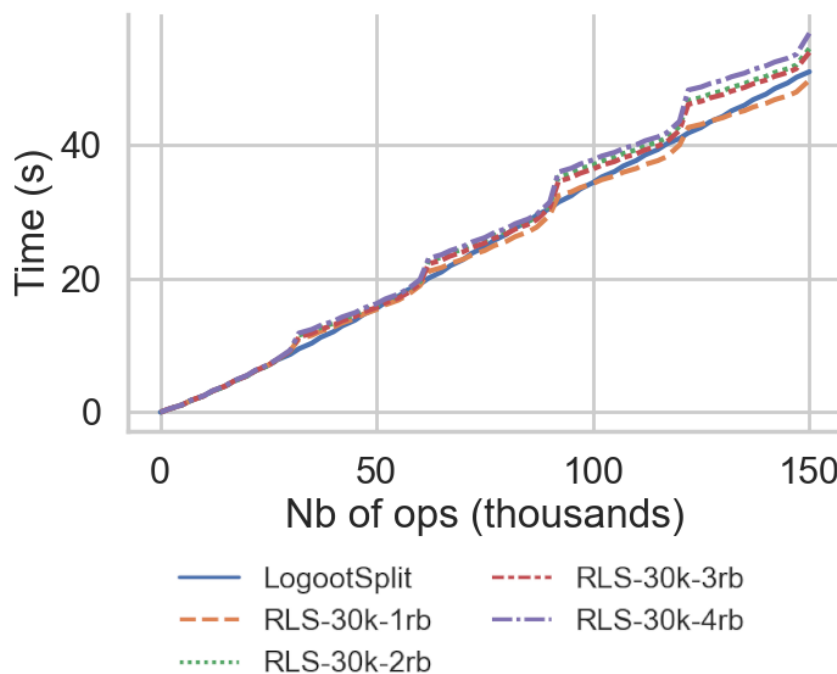


FIGURE 2.16 – Progression du nombre d'opérations du log rejouées en fonction du temps

Nous observons que le gain sur le temps d'intégration des opérations *insert* et *remove* permet initialement de contrebalancer le surcoût des opérations *rename*. Mais au fur et à mesure que la collaboration progresse, le temps nécessaire pour intégrer les opérations *rename* augmente car plus d'éléments sont impliqués. Cette tendance est accentuée dans les scénarios avec des opérations *rename* concurrentes.

Dans un cas réel d'utilisation, ce scénario (c.-à-d. rejouer l'entière du log) ne correspond pas au scénario principal et peut être mitigé, par exemple en utilisant un mécanisme de compression du log d'opérations. Dans la sous-section 2.5.6, nous présentons comment mettre en place un tel mécanisme en se basant justement sur les possibilités offertes par l'opération *rename*.

Impact de la fréquence de l'opération *rename* sur les performances

Pour évaluer l'impact de la fréquence de l'opération *rename* sur les performances, nous avons réalisé un benchmark supplémentaire. Ce benchmark consiste à rejouer les logs d'opérations des simulations en utilisant divers CRDTs et configurations : LogootSplit, RenamableLogootSplit effectuant des opérations *rename* toutes les 30k opérations, RenamableLogootSplit effectuant des opérations *rename* toutes les 7.5k opérations. Au fur et à mesure que le benchmark rejoue le log des opérations, il mesure le temps d'intégration des opérations ainsi que leur taille. Les résultats de ce benchmark sont présentés dans le Tableau 2.4.

Paramètres		Temps d'intégration (μ s)						Taille (o)					
Type	CRDT	Moyenne	Médiane	IQR	1 ^{er} Percent.	99 ^{ème} Percent.		Moyenne	Médiane	IQR	1 ^{er} Percent.	99 ^{ème} Percent.	
insert	LS	471	460	130	224	768		593	584	184	216	1136	
	RLS - 30k	397	323	66.7	171	587		442	378	92	314	958	
	RLS - 7.5k	393	265	54.5	133	381		389	378	0	314	590	
remove	LS	280	270	71.4	140	435		632	618	184	250	1170	
	RLS - 30k	247	181	39	97.9	308		434	412	0	320	900	
	RLS - 7.5k	296	151	34.8	74.9	214		401	412	0	320	596	

Paramètres		Temps d'intégration (ms)						Taille (Ko)					
Type	CRDT	Moyenne	Médiane	IQR	1 ^{er} Percent.	99 ^{ème} Percent.		Moyenne	Médiane	IQR	1 ^{er} Percent.	99 ^{ème} Percent.	
rename	RLS - 30k	1022	1188	425	540	1276		1366	1258	514	635	3373	
	RLS - 7.5k	861	974	669	123	1445		273	302	132	159	542	

TABLE 2.4 – Temps d'intégration et taille des opérations par type et par fréquence d'opérations *rename*

Concernant les temps d'intégration, nous observons des opérations *rename* plus fréquentes permettent d'améliorer les temps d'intégration des opérations *insert* et *remove*. Cela confirme les résultats attendus puisque l'opération *rename* réduit la taille des identifiants de la structure ainsi que le nombre de blocs composant la séquence.

Nous remarquons aussi que la fréquence n'a aucun impact significatif sur le temps d'intégration des opérations *rename*. Il s'agit là aussi d'un résultat attendu puisque la complexité en temps de l'implémentation de l'opération *rename* dépend du nombre d'éléments dans la séquence, un facteur qui n'est pas impacté par les opérations *rename*.

Concernant la taille des opérations, nous observons que les opérations *insert* et *remove* de RenamableLogootSplit sont initialement plus lourdes que les opérations correspondantes de LogootSplit, notamment car elles intègrent leur époque de génération comme donnée additionnelle. Mais alors que la taille des opérations de LogootSplit augmentent indéfiniment, celle des opérations de RenamableLogootSplit est bornée. La valeur de cette borne est définie par la fréquence de l'opération *rename*. Cela permet à RenamableLogootSplit d'atteindre un coût moindre par opération.

D'un autre côté, le coût des opérations *rename* est bien plus important (1000x) que celui des autres types d'opérations. Ceci s'explique par le fait que l'opération *rename* intègre l'*ancien état*, c.-à-d. la liste de tous les blocs composant l'état de la séquence au moment de la génération de l'opération. Cependant, nous observons le même phénomène pour les opérations *rename* que pour les autres opérations : la fréquence des opérations *rename* permet d'établir une borne pour la taille des opérations *rename*. Nous pouvons

donc choisir d'émettre fréquemment des opérations *rename* pour limiter leur taille respective. Ceci implique néanmoins un surcoût en computations pour chaque opération *rename* dans l'implémentation actuelle. Nous présentons une autre approche possible pour limiter la taille des opérations *rename* dans la sous-section 2.5.3. Cette approche consiste à implémenter un mécanisme de compression pour les opérations *rename* pour ne transmettre que les composants nécessaires à l'identifiant de chaque bloc de l'*ancien état*.

2.5 Discussion

Matthieu: TODO : Ajouter une partie sur la discussion qu'on a pu avoir avec les reviewers sur la présence de pierres tombales dans RenamableLogootSplit, et comment ces pierres tombales diffèrent de celles présentes dans WOOT et RGA.

2.5.1 Stratégie de génération des opérations *rename*

Comme indiqué dans la sous-section 2.1.2, les opérations *rename* sont des opérations systèmes. C'est donc aux concepteurs de systèmes qu'incombe la responsabilité de déterminer quand les noeuds devraient générer des opérations *rename* et de définir une stratégie correspondante. Il n'existe cependant pas de solution universelle, chaque système ayant ses particularités et contraintes.

Plusieurs aspects doivent être pris en compte lors de la définition de la stratégie de génération des opérations *rename*. Le premier porte sur la taille de la structure de données. Comme illustré dans la Figure 2.14, les métadonnées augmentent de manière progressive jusqu'à représenter 99% de la structure de données. En utilisant les opérations *rename*, les noeuds peuvent supprimer les métadonnées et ainsi réduire la taille de la structure à un niveau acceptable. Pour déterminer quand générer des opérations *rename*, les noeuds peuvent donc monitorer le nombre d'opérations effectuées depuis la dernière opération *rename*, le nombre de blocs qui composent la séquence ou encore la taille des identifiants.

Un second aspect à prendre en compte est le temps d'intégration des opérations *rename*. Comme indiqué dans le Tableau 2.3, l'intégration des opérations *rename* distantes peut nécessiter des secondes si elles sont retardées trop longtemps. Bien que les opérations *rename* travaillent en coulisses, elles peuvent néanmoins impacter négativement l'expérience utilisateur. Notamment, les noeuds ne peuvent pas intégrer d'autres opérations *distantes* tant qu'ils sont en train de traiter des opérations *rename*. Du point de vue des utilisateurs, les opérations *rename* peuvent alors être perçues comme des pics de latence. Dans le domaine de l'édition collaborative temps réel, IGNAT et al. [35, 34] ont montré que le délai dégradait la qualité des collaborations. Il est donc important de générer fréquemment des opérations *rename* pour conserver leur temps d'intégration sous une limite perceptible.

Finalement, le dernier aspect à considérer est le nombre d'opérations *rename* concurrentes. La Figure 2.14 montre que les opérations *rename* concurrentes accroissent la taille de la structure de données tandis que la Figure 2.16 illustre qu'elles augmentent le temps nécessaire pour rejouer le log d'opérations. La stratégie proposée doit donc viser à minimiser le nombre d'opérations *rename* concurrentes générées. Cependant, elle doit éviter

d'utiliser des coordinations synchrones entre les noeuds pour cela (e.g. algorithmes de consensus), pour des raisons de performances. Pour réduire la probabilité de générer des opérations *rename* concurrentes, plusieurs méthodes peuvent être proposées. Par exemple, les noeuds peuvent monitorer à quels autres noeuds ils sont connectés actuellement et déléguer au noeud ayant le plus grand *identifiant de noeud* la responsabilité de générer les opérations *rename*.

Pour récapituler, nous pouvons proposer plusieurs stratégies de génération des opérations *rename*, pour minimiser de manière individuelle chacun des paramètres présentés. Mais bien que certaines de ces stratégies convergent (minimiser la taille de la structure de données et minimiser le temps d'intégration des opérations *rename*), d'autres entrent en conflit (générer une opération *rename* dès qu'un seuil est atteint vs. minimiser le nombre d'opérations *rename* concurrentes générées). Les concepteurs de systèmes doivent proposer un compromis entre les différents paramètres en fonction des contraintes du système concerné (application temps réel ou asynchrone, limitations matérielles des noeuds...). Il est donc nécessaire d'analyser le système pour évaluer ses performances sur chaque aspect, ses usages et trouver le bon compromis entre tous les paramètres de la stratégie de renommage. Par exemple, dans le contexte des systèmes d'édition collaborative temps réel, [35] a montré que le délai diminue la qualité de la collaboration. Dans de tels systèmes, nous viserions donc à conserver le temps d'intégration des opérations (en incluant les opérations *rename*) en dessous du temps limite correspondant à leur perception par les utilisateurs.

2.5.2 Stockage des états précédents sur disque

Les noeuds doivent conserver les *anciens états* associés aux opérations *rename* pour transformer les opérations issues d'époques précédentes ou concurrentes. Les noeuds peuvent recevoir de telles opérations dans deux cas précis : (i) des noeuds ont émis récemment des opérations *rename* (ii) des noeuds se sont récemment reconnectés. Entre deux de ces événements spécifiques, les *anciens états* ne sont pas nécessaires pour traiter les opérations.

Nous pouvons donc proposer l'optimisation suivante : décharger les *anciens états* sur le disque jusqu'à leur prochaine utilisation ou jusqu'à ce qu'ils puissent être supprimés de manière sûre. Décharger les *anciens états* sur le disque permet de mitiger le surcoût en mémoire introduit par le mécanisme de renommage. En échange, cela augmente le temps d'intégration des opérations nécessitant un *ancien état* qui a été déchargé précédemment.

Les noeuds peuvent adopter différentes stratégies, en fonction de leurs contraintes, pour déterminer les *anciens états* comme déchargeables et pour les récupérer de manière préemptive. La conception de ces stratégies peut reposer sur différentes heuristiques : les époques des noeuds actuellement connectés, le nombre de noeuds pouvant toujours émettre des opérations concurrentes, le temps écoulé depuis la dernière utilisation de l'*ancien état*...

2.5.3 Compression et limitation de la taille de l'opération *rename*

Pour limiter la consommation en bande passante des opérations *rename*, nous proposons la technique de compression suivante. Au lieu de diffuser les identifiants complets

formant l'*ancien état*, les noeuds peuvent diffuser seulement les éléments nécessaires pour identifier de manière unique les intervalles d'identifiants. En effet, un identifiant peut être caractérisé de manière unique par le triplet composé de l'*identifiant de noeud*, du *numéro de séquence* et de l'*offset* de son dernier tuple. Par conséquent, un intervalle d'identifiants peut être identifié de manière unique à partir du triplet signature de son identifiant de début et de sa longueur, c.-à-d. du quadruplet $\langle nodeId, nodeSeq, offsetBegin, offsetEnd \rangle$. Cette méthode nous permet de réduire les données à diffuser dans le cadre de l'opération *rename* à un montant fixe par intervalle.

Pour décompresser l'opération reçue, les noeuds doivent reformer les intervalles d'identifiants correspondant aux quadruplets reçus. Pour cela, ils parcourent leur état. Lorsqu'ils rencontrent un identifiant partageant le même couple $\langle nodeId, nodeSeq \rangle$ qu'un des intervalles de l'opération *rename*, les noeuds disposent de l'ensemble des informations requises pour le reconstruire. Cependant, certains couples $\langle nodeId, nodeSeq \rangle$ peuvent avoir été supprimés en concurrence et ne plus être présents dans la séquence. Dans ce cas, il est nécessaire de parcourir le log des opérations *remove* concurrentes pour retrouver les identifiants correspondants et reconstruire l'opération *rename* originale.

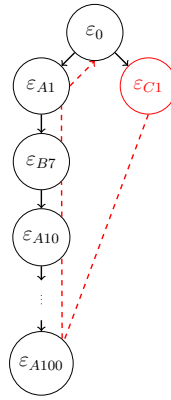
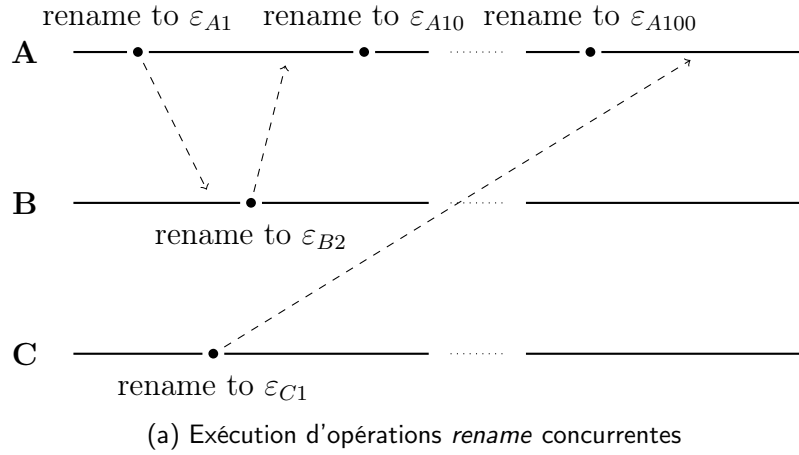
Grâce à cette méthode de compression, nous pouvons instaurer une taille maximale à l'opération *rename*. En effet, les noeuds peuvent émettre une opération *rename* dès que leur état courant atteint un nombre donné d'intervalles d'identifiants, bornant ainsi la taille du message à diffuser.

2.5.4 Définition de relations de priorité pour minimiser les traitements

Bien que la relation *priority* proposée dans la sous-section 2.3.2 est simple et garantit que tous les noeuds désignent la même époque comme époque cible, elle introduit un surcoût computationnel significatif dans certains cas. La Figure 2.17 présente un tel cas.

Dans cet exemple, les noeuds A et B éditent en collaboration un document. Au fur et à mesure de leur collaboration, ils effectuent plusieurs opérations *rename*. Cependant, après un nombre conséquent de modifications de leur part, un autre noeud C se reconnecte. Celui-ci leur transmet sa propre opération *rename*, concurrente à toutes leurs opérations. D'après la relation *priority*, nous avons $\varepsilon_0 <_{\varepsilon} \varepsilon_{A1} <_{\varepsilon} \dots <_{\varepsilon} \varepsilon_{A100} <_{\varepsilon} \varepsilon_{C1}$. La nouvelle époque cible étant ε_{C1} , les noeuds A et B doivent pour l'atteindre annuler successivement l'ensemble des opérations *rename* composant leur branche de l'*arbre des époques*. Ainsi, un noeud isolé peut forcer l'ensemble des noeuds à effectuer un lourd calcul. Il serait plus efficace que, dans cette situation, ce soit seulement le noeud isolé qui doive se mettre à jour.

La relation *priority* devrait donc être conçue pour garantir la convergence des noeuds, mais aussi pour minimiser les calculs effectués globalement par les noeuds du système. Pour concevoir une relation *priority* efficace, nous pourrions incorporer dans les opérations *rename* des métriques qui représentent l'état du système et le travail accumulé sur le document (nombre de noeuds actuellement à l'époque *parente*, nombre d'opérations générées depuis l'époque *parente*, taille du document...). De cette manière, nous pourrions favoriser la branche de l'*arbre des époques* regroupant les collaborateurs les plus actifs et

FIGURE 2.17 – Livraison d'une opération *rename* d'un noeud

empêcher les noeuds isolés d'imposer leurs opérations *rename*.

Afin d'offrir une plus grande flexibilité dans la conception de la relation *priority*, il est nécessaire de retirer la contrainte interdisant aux noeuds de rejouer une opération *rename*. Pour cela, un couple de fonctions réciproques doit être proposée pour `RENAMEID` et `REVERTRENAMEID`. Une solution alternative est de proposer une implémentation du mécanisme de renommage qui repose sur les identifiants originaux plutôt que sur ceux transformés, par exemple en utilisant le log des opérations.

2.5.5 Report de la transition vers la nouvelle époque cible

Comme illustré par le Tableau 2.3, intégrer des opérations *rename* distantes est généralement coûteux. Ce traitement peut générer un surcoût computationnel significatif en cas de multiples opérations *rename* concurrentes. En particulier, un noeud peut recevoir et intégrer les opérations *rename* concurrentes dans l'ordre inverse défini par la relation *priority* sur leur époque. Dans ce scénario, le noeud considérerait chaque nouvelle époque introduite comme la nouvelle époque cible et renommerait son état en conséquence à chaque fois.

En cas d'un grand nombre d'opérations *rename* concurrentes, nous proposons que

les noeuds délaient le renommage de leur état vers l'époque cible jusqu'à ce qu'ils aient obtenu un niveau de confiance donné en l'époque cible. Ce délai réduit la probabilité que les noeuds effectuent des traitements inutiles. Plusieurs stratégies peuvent être proposées pour calculer le niveau de confiance en l'époque cible. Ces stratégies peuvent reposer sur une variété de métriques pour produire le niveau de confiance, tel que le temps écoulé depuis que le noeud a reçu une opération *rename* concurrente et le nombre de noeuds en ligne qui n'ont pas encore reçu l'opération *rename*.

Durant cette période d'incertitude introduite par le report, les noeuds peuvent recevoir des opérations provenant d'époques différentes, notamment de l'époque cible. Néanmoins, les noeuds peuvent toujours intégrer les opérations *insert* et *remove* en utilisant `RENAMEID` et `REVERTRENAMEID` au prix d'un surcoût computationnel pour chaque identifiant. Cependant, ce coût est négligeable (plusieurs centaines de microsecondes par identifiant d'après la Figure 2.15b) comparé au coût de renommer, de manière inutile, complètement l'état (plusieurs centaines de millisecondes à des secondes complètes d'après le Tableau 2.3).

Notons que ce mécanisme nécessite que `RENAMEID` et `REVERTRENAMEID` soient des fonctions réciproques. En effet, au cours de la période d'incertitude, un noeud peut avoir à utiliser `REVERTRENAMEID` pour intégrer les identifiants d'opérations *insert* distantes provenant de l'époque cible. Ensuite, le noeud peut devoir renommer son état vers l'époque cible une fois que celle-ci a obtenu le niveau de confiance requis. Il s'ensuit que `RENAMEID` doit restaurer les identifiants précédemment transformés par `REVERTRENAMEID` à leur valeur initiale pour garantir la convergence.

2.5.6 Utilisation de l'opération de renommage comme mécanisme de compression du log d'opérations

Lorsqu'un nouveau pair rejoint la collaboration, il doit tout d'abord récupérer l'état courant du document avant de pouvoir participer. Le nouveau pair utilise un mécanisme d'anti-entropie [58] pour récupérer l'ensemble des opérations via un autre pair. Puis il reconstruit l'état courant en appliquant successivement chacune des opérations. Ce processus peut néanmoins s'avérer coûteux pour les documents comprenant des milliers d'opérations.

Pour pallier ce problème, des mécanismes de compression du log ont été proposés dans la littérature. Les approches présentées dans [69, 33] consistent à remplacer un sous-ensemble des opérations du log par une opération équivalente, par exemple en agrégeant les opérations *insert* adjacentes. Une autre approche, présentée dans [9], définit une relation *obsolete* sur les opérations. La relation *obsolete* permet de spécifier qu'une nouvelle opération rend obsolètes des opérations précédentes et permet de les retirer du log. Pour donner un exemple, une opération d'ajout d'un élément donné dans un OR-Set CRDT rend obsolètes toutes les opérations précédentes d'ajout et de suppression de cet élément.

Dans notre contexte, il est intéressant de noter que l'opération *rename* peut endosser un rôle comparable à ces mécanismes de compression du log. En effet, l'opération *rename* prend un état donné, somme des opérations passées, et génère en retour un nouvel état équivalent et compacté. Une opération *rename* rend donc obsolète l'ensemble des opérations dont elle dépend causalement, et peut être utilisée pour les remplacer. En partant

de cette observation, nous proposons le mécanisme de compression du log suivant.

Le mécanisme consiste à réduire le nombre d'opérations transmises à un nouveau pair rejoignant la collaboration grâce à l'opération *rename* de l'époque courante. L'opération *rename* ayant introduite l'époque courante fournit un état initial au nouveau pair. À partir de cet état initial, le nouveau pair peut obtenir l'état courant en intégrant les opérations *insert* et *remove* qui ont été générées de manière concurrente ou causale par rapport à l'opération *rename*. En réponse à une demande de synchronisation d'un nouveau pair, un pair peut donc simplement lui envoyer un sous-ensemble de son log composé de : (i) l'opération *rename* ayant introduite son époque courante (ii) les opérations *insert* et *remove* dont l'opération *rename* courante ne dépend pas causalement.

Notons que les données contenues dans l'opération *rename* telle que nous l'avons définie précédemment (cf. Définition 15) sont insuffisantes pour cette utilisation. En effet, les données incluses (*ancien état* au moment du renommage, identifiant du noeud auteur de l'opération *rename* et son numéro de séquence au moment de la génération) nous permettent seulement de recréer la structure de la séquence après le renommage. Mais le contenu de la séquence est omis, celui-ci n'étant jusqu'ici d'aucune utilité pour l'opération *rename*. Afin de pouvoir utiliser l'opération *rename* comme état initial, il est nécessaire d'y inclure cette information.

De plus, des informations de causalité doivent être intégrées à l'opération *rename*. Ces informations doivent permettre aux noeuds d'identifier les opérations supplémentaires nécessaires pour obtenir l'état courant, c.-à-d. toutes les opérations desquelles l'opération *rename* ne dépend pas causalement. L'ajout à l'opération *rename* d'un *vecteur de version*, structure représentant l'ensemble des opérations observées par l'auteur de l'opération *rename* au moment de sa génération, permettrait cela.

Nous définissons donc de la manière suivante l'opération *rename* enrichie compatible avec ce mécanisme de compression du log :

Définition 22 (rename enrichie) Une opération *rename* enrichie est un quintuplet $\langle nodeId, nodeSeq, formerState, versionVector, content \rangle$ où

- *nodeId* est l'identifiant du noeud qui a générée l'opération *rename*.
- *nodeSeq* est le numéro de séquence du noeud au moment de la génération de l'opération *rename*.
- *formerState* est l'ancien état du noeud au moment du renommage.
- *versionVector* est le vecteur de version représentant l'ancien état du noeud au moment du renommage.
- *content* est le contenu du document au moment du renommage.

Ce mécanisme de compression du log introduit néanmoins le problème suivant. Un nouveau pair synchronisé de cette manière ne possède qu'un sous-ensemble du log des opérations. Si ce pair reçoit ensuite une demande de synchronisation d'un second pair, il est possible qu'il ne puisse répondre à la requête. Par exemple, le pair ne peut pas fournir des opérations faisant partie des dépendances causales de l'opération *rename* qui lui a servi d'état initial.

Une solution possible dans ce cas de figure est de rediriger le second pair vers un troisième pour qu’il se synchronise avec lui. Cependant, cette solution pose des problèmes de latence/temps de réponse si le troisième pair s’avère indisponible à ce moment. Une autre approche possible est de généraliser le processus de synchronisation que nous avons présenté ici (opération *rename* comme état initial puis application des autres opérations) à l’ensemble des pairs, et non plus seulement aux nouveaux pairs. Nous présentons les avantages et inconvénients de cette approche dans la sous-section suivante.

2.5.7 Implémentation alternative de l’intégration de l’opération *rename* basée sur le log d’opérations

Nous avons décrit précédemment dans la section 2.3.4, et plus précisément dans l’Algorithme 4, le processus d’intégration de l’opération *rename* évaluée dans ce manuscrit. Pour rappel, le processus consiste à (i) identifier le chemin entre l’époque courante et l’époque cible (ii) appliquer les fonctions de transformations REVERTRENAMEID et RENAMEID à l’ensemble des identifiants composant l’état courant (iii) re-crée une séquence à partir des nouveaux identifiants calculés et du contenu courant.

Dans cette section, nous abordons une implémentation alternative de l’intégration de l’opération *rename*. Cette implémentation repose sur le log des opérations.

Cette implémentation se base sur les observations suivantes : (i) L’état courant est obtenu en intégrant successivement l’ensemble des opérations. (ii) L’opération *rename* est une opération subsumant les opérations passées : elle prend un état donné (l’*ancien état*), somme des opérations précédentes, et génère un nouvel état équivalent compacté. (iii) L’ordre d’intégration des opérations concurrentes n’a pas d’importance sur l’état final obtenu.

Ainsi, pour intégrer une opération *rename* distante, un noeud peut (i) générer l’état correspondant au renommage de l’*ancien état* (ii) identifier le chemin entre l’époque courante et l’époque cible (iii) identifier les opérations concurrentes à l’opération *rename* présentes dans son log (iv) transformer et intégrer successivement les opérations concurrentes à l’opération *rename* à ce nouvel état

Cet algorithme est équivalent à ré-ordonner le log des opérations de façon à intégrer les opérations précédant l’opération *rename*, puis à intégrer l’opération *rename* elle-même, puis à intégrer les opérations concurrentes à cette dernière.

Cette approche présente plusieurs avantages par rapport à l’implémentation décrite dans la section 2.3.4. Tout d’abord, elle modifie le facteur du nombre de transformations à effectuer. La version décrite dans la section 2.3.4 transforme de l’époque courante vers l’époque cible chaque identifiant (ou chaque bloc si on dispose de RENAMEBLOCK) de l’état courant. La version présentée ici effectue une transformation pour chaque opération du log concurrente à l’opération *rename* à intégrer. Le nombre de transformation peut donc être réduit de plusieurs ordres de grandeur avec cette approche, notamment si les opérations sont propagées aux pairs du réseau rapidement.

Un autre avantage de cette approche est qu’elle permet de récupérer et de réutiliser les identifiants originaux des opérations. Lorsqu’une suite de transformations est appliquée sur les identifiants d’une opération, elle est appliquée sur les identifiants originaux et

non plus sur leur équivalents présents dans l'état courant. Ceci permet de réinitialiser les transformations appliquées à un identifiant et d'éviter le cas de figure mentionné dans la sous-section 2.3.3 : le cas où `REVERTRENAMEID` est utilisé pour retirer l'effet d'une opération *rename* sur un identifiant, avant d'utiliser `RENAMEID` pour ré-intégrer l'effet de la même opération *rename*. Cette implémentation supprime donc la contrainte de définir un couple de fonctions réciproques `RENAMEID` et `REVERTRENAMEID`, ce qui nous offre une plus grande flexibilité dans le choix de la relation $<_\varepsilon$ et du couple de fonctions `RENAMEID` et `REVERTRENAMEID`.

Cette implémentation dispose néanmoins de plusieurs limites. Tout d'abord, elle nécessite que chaque noeud maintienne localement le log des opérations. Les métadonnées accumulées par la structure de données répliquées vont alors croître avec le nombre d'opérations effectuées. Cependant, ce défaut est à nuancer. En effet, les noeuds doivent déjà maintenir le log des opérations pour le mécanisme d'anti-entropie, afin de renvoyer une opération passée à un noeud l'ayant manquée. Plus globalement, les noeuds doivent aussi conserver le log des opérations pour permettre à un nouveau noeud de rejoindre la collaboration et de calculer l'état courant en rejouant l'ensemble des opérations. Il s'agit donc d'une contrainte déjà imposée aux noeuds pour d'autres fonctionnalités du système.

Un autre défaut de cette implémentation est qu'elle nécessite de détecter les opérations concurrentes à l'opération *rename* à intégrer. Cela implique d'ajouter des informations de causalité à l'opération *rename*, tel qu'un vecteur de version. Cependant, la taille des vecteurs de version croît de façon monotone avec le nombre de noeuds qui participent à la collaboration. Diffuser cette information à l'ensemble des noeuds peut donc représenter un coût significatif dans les collaborations à large échelle. Néanmoins, il faut rappeler que les noeuds échangent déjà régulièrement des vecteurs de version dans le cadre du fonctionnement du mécanisme d'anti-entropie. Les opérations *rename* étant rares en comparaison, ce surcoût nous paraît acceptable.

Finalement, cette approche implique aussi de parcourir le log des opérations à la recherche d'opérations concurrentes. Comme dit précédemment, la taille du log croît de façon monotone au fur et à mesure que les noeuds émettent des opérations. Cette étape du nouvel algorithme d'intégration de l'opération *rename* devient donc de plus en plus coûteuse. Des méthodes permettent néanmoins de réduire son coût computationnel. Notamment, chaque noeud traquent les informations de progression des autres noeuds afin de supprimer les métadonnées du mécanisme de renommage (cf. sous-section 2.3.5). Ces informations permettent de déterminer la stabilité causale des opérations et donc d'identifier les opérations qui ne peuvent plus être concurrentes à une nouvelle opération *rename*. Les noeuds peuvent ainsi maintenir, en plus du log complet des opérations, un log composé uniquement des opérations non stables causalement. Lors du traitement d'une nouvelle opération *rename*, les noeuds peuvent alors parcourir ce log réduit à la recherche des opérations concurrentes.

2.6 Comparaison avec les approches existantes

2.6.1 Core-Nebula

L'approche *core-nebula* [42, 81] a été proposée pour réduire la taille des identifiants dans Treedoc [62]. Dans ces travaux, les auteurs définissent l'opération *rebalance* qui permet aux noeuds de réassigner des identifiants plus courts aux éléments du document. Cependant, cette opération *rebalance* n'est ni commutative avec les opérations *insert* et *remove*, ni avec elle-même. Pour assurer la convergence à terme [71], l'approche *core-nebula* empêche la génération d'opérations *rebalance* concurrentes. Pour ce faire, l'approche requiert un consensus entre les noeuds pour générer les opérations *rebalance*. Des opérations *insert* et *remove* sont elles toujours générées sans coordination entre les noeuds et peuvent donc être concurrentes aux opérations *rebalance*. Pour gérer les opérations concurrentes aux opérations *rebalance*, les auteurs proposent de transformer les opérations concernées par rapport aux effets des opérations *rebalance*, à l'aide un mécanisme de *catch-up*, avant de les appliquer.

Cependant, les protocoles de consensus ne passent pas à l'échelle et ne sont pas adaptés aux systèmes distribués à large échelle. Pour pallier ce problème, l'approche *core-nebula* propose de répartir les noeuds dans deux groupes : le *core* et la *nebula*. Le *core* est un ensemble, de taille réduite, de noeuds stables et hautement connectés tandis que la *nebula* est un ensemble, de taille non-bornée, de noeuds. Seuls les noeuds du *core* participent à l'exécution du protocole de consensus. Les noeuds de la *nebula* contribuent toujours au document par le biais des opérations *insert* et *remove*.

Notre travail peut être vu comme une extension de celui présenté dans *core-nebula*. Avec RenamableLogootSplit, nous adaptons l'opération *rebalance* et le mécanisme de *catch-up* à LogootSplit pour tirer partie de la fonctionnalité offerte par les blocs. De plus, nous proposons un mécanisme pour supporter les opérations *rename* concurrentes, ce qui supprime la nécessité de l'utilisation d'un protocole de consensus. Notre contribution est donc une approche plus générique puisque RenamableLogootSplit est utilisable dans des systèmes composés d'un *core* et d'une *nebula*, ainsi que dans les systèmes ne disposant pas de noeuds stables pour former un *core*.

Dans les systèmes disposant d'un *core*, nous pouvons donc combiner RenamableLogootSplit avec un protocole de consensus pour éviter la génération d'opérations *rename* concurrentes. Cette approche offre plusieurs avantages. Elle permet de se passer de tout ce qui à attiré au support d'opérations *rename* concurrentes, c.-à-d. la définition d'une relation *priority* et l'implémentation de REVERTRENAMEID. Elle permet aussi de simplifier l'implémentation du mécanisme de récupération de mémoire des époques et *anciens états* pour reposer seulement sur la stabilité causale des opérations. Concernant ses performances, cette approche se comporte de manière similaire à RenamableLogootSplit avec un seul *renaming bot* (cf. sous-section 2.4.3), mais avec un surcoût correspondant au coût du protocole de consensus sélectionné.

2.6.2 LSEQ

L'approche LSEQ [48, 46] est une approche visant à réduire la croissance des identifiants dans les Séquences CRDTs à identifiants densément ordonnés. Au lieu de réduire périodiquement la taille des métadonnées liées aux identifiants à l'aide d'un mécanisme de renommage coûteux, les auteurs définissent de nouvelles stratégies d'allocation des identifiants pour réduire leur vitesse de croissance. Dans ces travaux, les auteurs notent que la stratégie d'allocation des identifiants proposée dans Logoot [72] n'est adaptée qu'à un seul comportement d'édition : de gauche à droite, de haut en bas. Si les insertions sont effectuées en suivant d'autres comportements, les identifiants générés saturent rapidement l'espace des identifiants pour une taille donnée. Les insertions suivantes déclenchent alors une augmentation de la taille des identifiants. En conséquent, la taille des identifiants dans Logoot augmente de façon linéaire au nombre d'insertions, au lieu de suivre la progression logarithmique attendue.

LSEQ définit donc plusieurs stratégies d'allocation d'identifiants adaptées à différents comportements d'édition. Les noeuds choisissent aléatoirement une de ces stratégies pour chaque taille d'identifiants. De plus, LSEQ adopte une structure d'arbre exponentiel pour allouer les identifiants : l'intervalle des identifiants possibles double à chaque fois que la taille des identifiants augmente. Cela permet à LSEQ de choisir avec soin la taille des identifiants et la stratégie d'allocation en fonction des besoins. En combinant les différentes stratégies d'allocation avec la structure d'arbre exponentiel, LSEQ offre une croissance polylogarithmique de la taille des identifiants en fonction du nombre d'insertions.

Bien que l'approche LSEQ réduit la vitesse de croissance des identifiants dans les Séquences CRDTs à identifiants densément ordonnés, le surcoût de la séquence reste proportionnel à son nombre d'éléments. À l'inverse, le mécanisme de renommage de RenamableLogootSplit permet de réduire les métadonnées à une quantité fixe, indépendamment du nombre d'éléments.

Ces deux approches sont néanmoins orthogonales et peuvent, comme avec l'approche précédente, être combinées. Le système résultant réinitialiserait périodiquement les métadonnées de la séquence répliquée à l'aide de l'opération *rename* tandis que les stratégies d'allocation d'identifiants de LSEQ réduiraient leur croissance entretemps. Cela permettrait aussi de réduire la fréquence de l'opération *rename*, réduisant ainsi les calculs effectués par le système de manière globale.

2.7 Conclusion

Dans ce chapitre, nous avons présenté un nouvel Sequence CRDT : RenamableLogootSplit. Ce nouveau type de données répliquées associe à LogootSplit un mécanisme de renommage optimiste permettant de réduire périodiquement les métadonnées stockées et d'optimiser l'état interne de la structure de données.

Ce mécanisme prend la forme d'une nouvelle opération, l'opération *rename*, qui peut être émise à tout moment par n'importe quel noeud. Cette opération génère une nouvelle séquence LogootSplit, équivalente à l'état précédent, avec une empreinte minimale en métadonnées. L'opération *rename* transporte aussi suffisamment d'informations pour que

les noeuds puissent intégrer les opérations concurrentes à l'opération *rename* dans le nouvel état.

En cas d'opérations *rename* concurrentes, la relation d'ordre strict total $<_{\epsilon}$ permet aux noeuds de décider quelle opération *rename* utiliser, sans coordination. Les autres opérations *rename* sont quant à elles ignorées. Seules leurs informations sont stockées par `RenamableLogootSplit`, afin de gérer les opérations concurrentes potentielles.

Une fois qu'une opération *rename* a été propagée à l'ensemble des noeuds, elle devient causalement stable. À partir de ce point, il n'est plus possible qu'un noeud émette une opération concurrente à cette dernière. Les informations incluses dans l'opération *rename* pour intégrer les opérations concurrentes potentielles peuvent donc être supprimées par l'ensemble des noeuds.

Ainsi, le mécanisme de renommage permet à `RenamableLogootSplit` d'offrir de meilleures performances que `LogootSplit`. La génération du nouvel état minimal et la suppression à terme des métadonnées du mécanisme de renommage divisent par 100 la taille de la structure de données répliquée. L'optimisation de l'état interne représentant la séquence réduit aussi le coût d'intégration des opérations suivantes, amortissant ainsi le coût de transformation et d'intégration des opérations concurrentes à l'opération *rename*.

`RenamableLogootSplit` souffre néanmoins de plusieurs limitations. La première d'entre elles est le besoin d'observer la stabilité causale des opérations *rename* pour supprimer de manière définitive les métadonnées associées. Il s'agit d'une contrainte forte, notamment dans les systèmes dynamiques à grande échelle dans lesquels nous n'avons aucune garantie et aucun contrôle sur les noeuds. Il est donc possible qu'un noeud déconnecté ne se reconnecte jamais, bloquant ainsi la progression de la stabilité causale pour l'ensemble des opérations. Il s'agit toutefois d'une limite partagée avec les autres mécanismes de réduction des métadonnées pour Sequence CRDTs proposés dans la littérature [65, 81], à l'exception de l'approche LSEQ [46]. En pratique, il serait intéressant d'étudier la mise en place d'un mécanisme d'éviction des noeuds inactifs pour répondre à ce problème.

La seconde limitation de `RenamableLogootSplit` concerne la génération d'opérations *rename* concurrentes. Chaque opération *rename* est coûteuse, aussi bien en terme de métadonnées à stocker et diffuser qu'en terme de traitements à effectuer. Il est donc important de chercher à minimiser le nombre d'opérations *rename* concurrentes émises par les noeuds. Une approche possible est d'adopter une architecture du type *core-nebula* [81]. Mais pour les systèmes incompatibles avec ce type d'architecture système, il serait intéressant de proposer d'autres approches ne nécessitant aucune coordination entre les noeuds. Mais par définition, ces approches ne pourraient offrir de garanties fortes sur le nombre d'opérations concurrentes possibles.

Chapitre 3

MUTE, un éditeur de texte web collaboratif P2P temps réel chiffré de bout en bout

Sommaire

3.1	Couche interface utilisateur	78
3.2	Couche réplication	80
3.2.1	Modèle de données du document texte	80
3.2.2	Module de livraison des opérations	80
3.2.3	Collaborateurs	88
3.2.4	Curseurs	92
3.3	Couche réseau	92
3.3.1	Établissement d'un réseau P2P entre navigateurs	92
3.3.2	Topologie réseau	94
3.4	Couche sécurité	94
3.5	Conclusion	96

Les systèmes collaboratifs temps réels permettent à plusieurs utilisateurs de réaliser une tâche de manière coopérative. Ils permettent aux utilisateurs de consulter le contenu actuel, de le modifier et d'observer en direct les modifications effectuées par les autres collaborateurs. L'observation en temps réel des modifications des autres favorise une réflexion de groupe et permet une répartition efficace des tâches. L'utilisation des systèmes collaboratifs se traduit alors par une augmentation de la qualité du résultat produit[56, 27].

Plusieurs outils d'édition collaborative centralisés basés sur l'approche OT ont permis de populariser l'édition collaborative temps réel de texte[29, 25]. Ces approches souffrent néanmoins de leur architecture centralisée. Notamment, ces solutions rencontrent des difficultés à passer à l'échelle[34, 18] et posent des problèmes de confidentialité[26, 30].

L'approche CRDT offre une meilleure capacité de passage à l'échelle et est compatible avec une architecture P2P[2]. Ainsi, de nombreux travaux[45, 60, 59] ont été entrepris pour proposer une alternative distribuée répondant aux limites des éditeurs collaboratifs

centralisés. De manière plus globale, ces travaux s’inscrivent dans le nouveau paradigme d’application des *Local-First Softwares*[39, 32]. Ce paradigme vise le développement d’applications collaboratives, P2P, pérennes et rendant la souveraineté de leurs données aux utilisateurs.

De manière semblable, l’équipe Coast conçoit depuis plusieurs années des applications avec ces mêmes objectifs et étudient les problématiques de recherche liées. Elle développe Multi User Text Editor (MUTE)[55]^{6 7}, un éditeur collaboratif P2P temps réel chiffré de bout en bout. MUTE sert de plateforme d’expérimentation et de démonstration pour les travaux de l’équipe. Nous avons donc intégré dans MUTE les travaux de cette thèse.

La Figure 3.1 représente l’architecture système d’une collaboration utilisant MUTE. MUTE étant une application web, chaque noeud représenté ici correspond à un navigateur.

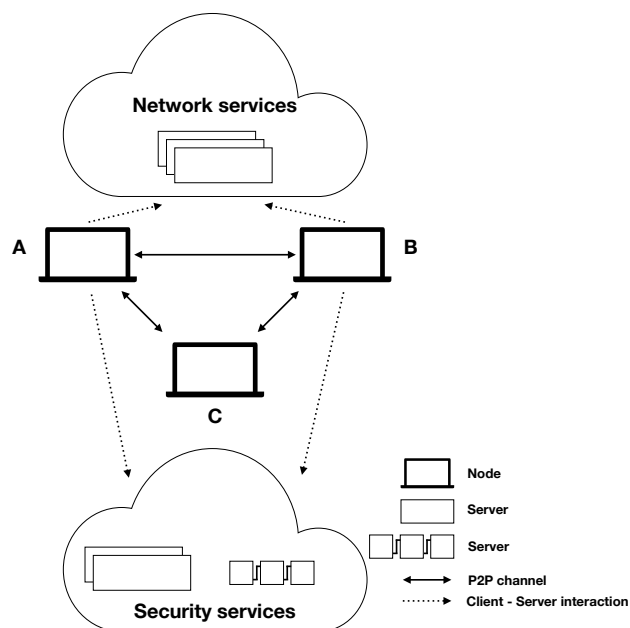


FIGURE 3.1 – Architecture système de l’application MUTE

Nous décrivons l’architecture logicielle d’un noeud dans la Figure 3.2. Dans ce chapitre, nous présentons les différentes couches logicielles de MUTE. Notamment, nous détaillons les couches correspondantes à l’implémentation des travaux présentés dans le chapitre 2. De la même façon, nous précisons dans ce chapitre les différents composants de l’architecture système de MUTE autre que les pairs eux-mêmes.

3.1 Couche interface utilisateur

La Figure 3.3 illustre l’interface utilisateur de l’éditeur de document de MUTE.

L’interface se compose d’un éditeur de texte supportant le langage de balisage Markdown. Ainsi, l’éditeur permet d’inclure plusieurs éléments légers de style. Les balises du

6. Disponible à l’adresse : <https://mutehost.loria.fr>

7. Code source disponible à l’adresse suivante : <https://github.com/coast-team/mute>

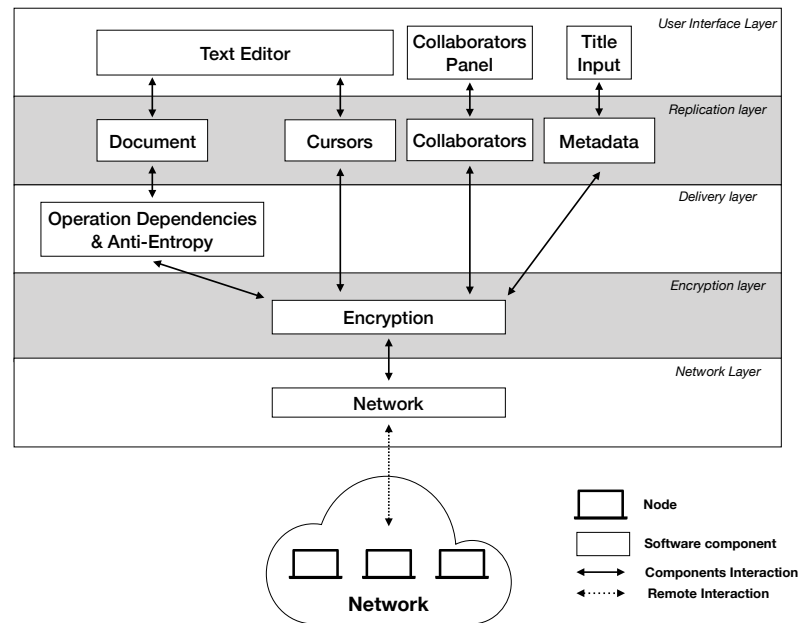


FIGURE 3.2 – Architecture logicielle de l'application MUTE

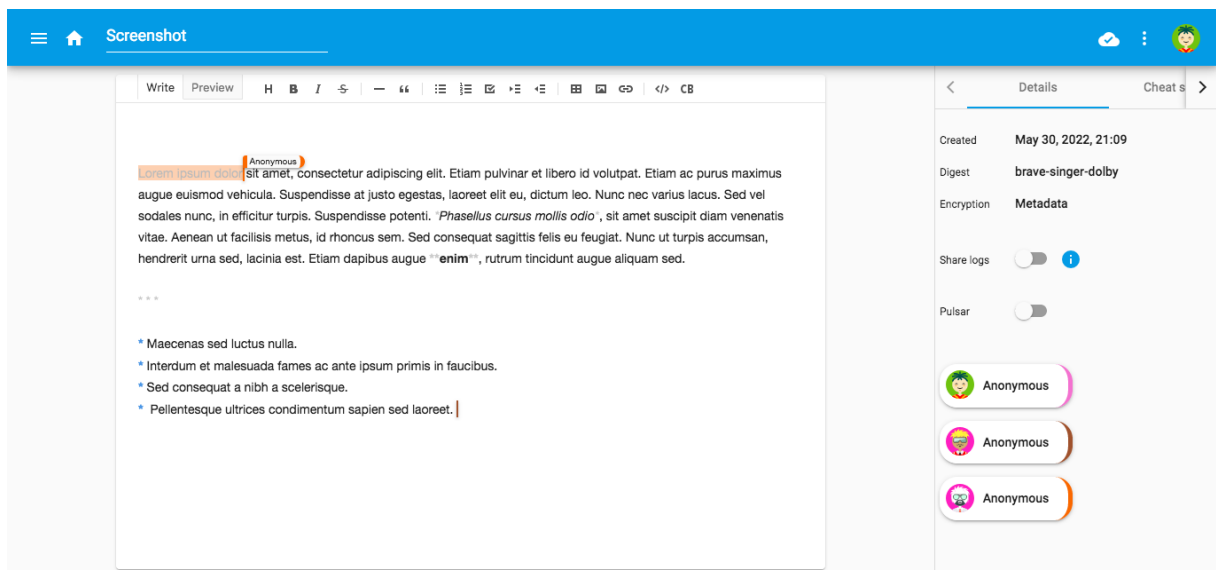


FIGURE 3.3 – Capture d'écran d'une session d'édition collaborative avec MUTE

langage Markdown étant du texte, elles sont répliquées nativement par la structure de données utilisée en interne par MUTE.

L'éditeur est agrémenté de plusieurs mécanismes permettant d'établir une conscience de groupe entre les collaborateurs. L'indicateur en haut à droite de la fenêtre représente le statut de connexion de l'utilisateur. Ceci lui indique s'il est actuellement connecté au réseau P2P, en cours de connexion, ou si un incident réseau a lieu.

MUTE affiche la liste des collaborateurs actuellement connectés sur la droite de l'éditeur. De plus, un curseur ou une sélection distante est associée à chaque collaborateur de

la liste. Elle permet d'indiquer à l'utilisateur courant dans quelles sections du document ses collaborateurs sont en train de travailler. Ainsi, ils peuvent se répartir la rédaction du document de manière implicite ou suivre facilement les modifications d'un collaborateur.

Les documents de l'utilisateur étant sauvegardés dans le navigateur, les documents sont aussi bien disponibles en étant en ligne que hors ligne. Une seconde page, listant les documents sauvegardés, permet à l'utilisateur de parcourir ses différents documents.

3.2 Couche réplication

3.2.1 Modèle de données du document texte

MUTE propose plusieurs alternatives pour représenter le document texte. MUTE permet de soit utiliser une implémentation de LogootSplit⁸, soit de RenamableLogootSplit⁸ ou soit de Dotted LogootSplit⁹. Ce choix est effectué via une valeur de configuration de l'application choisie au moment de son déploiement.

Le modèle de données utilisé interagit avec l'éditeur de texte par l'intermédiaire d'opérations textes. Lorsque l'utilisateur effectue des modifications locales, celles-ci sont détectées et mises sous la forme d'opérations textes. Elles sont transmises au modèle de données, qui les intègre alors à la structure de données répliquées. Le CRDT retourne en résultat l'opération distante à propager aux autres noeuds.

De manière complémentaire, lorsqu'une opération distante est délivrée au modèle de données, elle est intégrée par le CRDT pour actualiser son état. Le CRDT génère les opérations textes correspondantes et les transmet à l'éditeur de texte pour mettre à jour la vue.

En plus du texte, MUTE maintient un ensemble de métadonnées par document. Par exemple, les utilisateurs peuvent donner un titre au document. Pour représenter cette donnée additionnelle, nous associons un Last-Writer-Wins Register CRDT synchronisé par opérations[68] au document. De façon similaire, nous utilisons un First-Writer-Wins Register CRDT synchronisé par opérations pour représenter la date de création du document.

3.2.2 Module de livraison des opérations

Dans le cadre de LogootSplit et de RenamableLogootSplit, le modèle de données utilisé pour représenter le document texte est couplé au composant **Sync**. Le rôle de ce composant est d'assurer le respect du modèle de livraison des opérations au CRDT. Pour cela, le module **Sync** doit implémenter les contraintes présentées dans la Définition 13 et dans la Définition 17.

8. Les deux implémentations proviennent de la librairie `mute-structs` : <https://github.com/coast-team/mute-structs>

9. Implémentation fournie par la librairie suivante : <https://github.com/coast-team/dotted-logootsplit>

Livraison des opérations en exactement un exemplaire

Afin de respecter la contrainte de *exactly-once delivery*, il est nécessaire d'identifier de manière unique chaque opération. Pour cela, le module **Sync** ajoute un *Dot*[5] à chaque opération :

Définition 23 (Dot) *Un Dot est une paire $\langle nodeId, nodeSyncSeq \rangle$ où*

- *$nodeId$ est l'identifiant unique du noeud qui a généré l'opération.*
- *$nodeSyncSeq$ est le numéro de séquence courant du noeud à la génération de l'opération.*

Il est à noter que *nodeSyncSeq* est différent du *nodeSeq* utilisé dans LogootSplit et RenamableLogootSplit (cf. Définition 10). En effet, *nodeSyncSeq* se doit d'augmenter à chaque opération tandis que *nodeSeq* n'augmente qu'à la création d'un nouveau bloc. Les contraintes étant différentes, il est nécessaire de distinguer ces deux données.

Chaque noeud maintient une structure de données représentant l'ensemble des opérations reçues par le pair. Elle permet de vérifier à la réception d'une opération si le dot de cette dernière est déjà connu. S'il s'agit d'un nouveau dot, le module **Sync** peut délivrer l'opération au CRDT et ajouter son dot à la structure. Le cas échéant, cela indique que l'opération a déjà été délivrée précédemment et doit être ignorée cette fois-ci.

Plusieurs structures de données sont adaptées pour maintenir l'ensemble des opérations reçues. Dans le cadre de MUTE, nous avons choisi d'utiliser un vecteur de versions. Cette structure nous permet de réduire à un dot par noeud le surcoût en métadonnées du module **Sync**, puisqu'il ne nécessite que de stocker le dot le plus récent par noeud. Cette structure permet aussi de vérifier en temps constant si une opération est déjà connue. La Figure 3.4 illustre son fonctionnement.

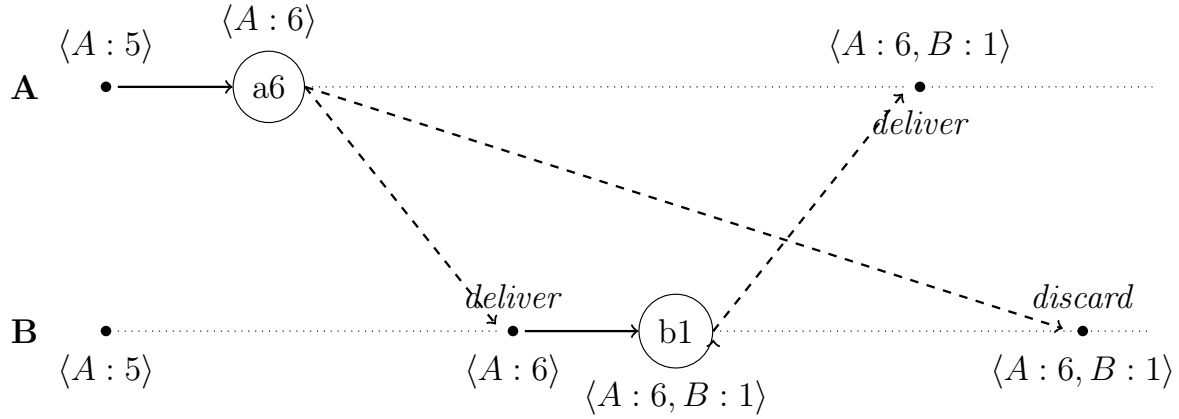
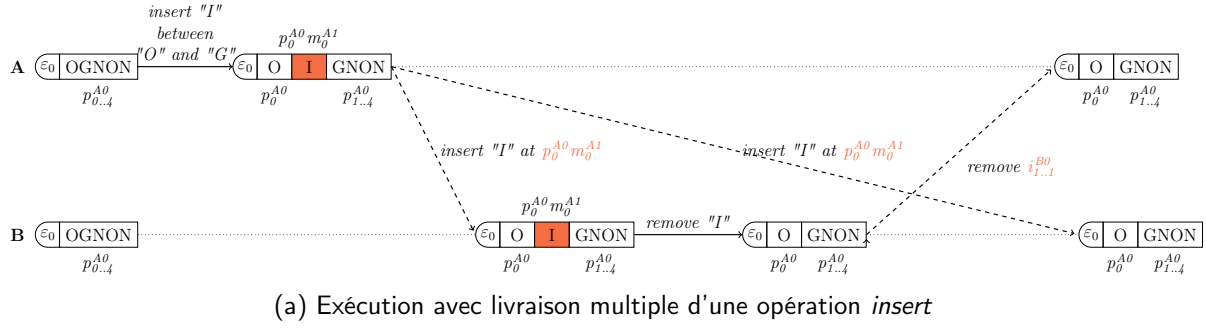
Dans cet exemple, qui reprend celui de la Figure 1.12, deux noeuds A et B répliquent une séquence. Initialement, celle-ci contient les éléments "OGNON". Ces éléments ont été insérés un par un par le noeud A, donc par le biais des opérations *a1* à *a5*. Le module **Sync** de chaque noeud maintient donc initialement le vecteur de version $\langle A : 5 \rangle$.

Le noeud A insère l'élément "I" entre les éléments "O" et "G". Cette modification est alors labellisée *a6* par son module **Sync** et est envoyée au noeud B. À la réception de cette opération, le module **Sync** de B compare son dot avec son vecteur de version local. L'opération *a6* étant la prochaine opération attendue de A, celle-ci est acceptée : elle est alors délivrée au CRDT et le vecteur de version est mis à jour.

Le noeud B supprime ensuite l'élément nouvellement inséré. S'agissant de la première modification de B, cette modification *b1* ajoute l'entrée correspondante dans le vecteur de version $\langle A : 6, B : 1 \rangle$. L'opération est envoyée au noeud A. Cette opération étant la prochaine opération attendue de B, elle est acceptée et délivrée.

Finalement, le noeud B reçoit de nouveau l'opération *a6*. Son module **Sync** détermine alors qu'il s'agit d'un doublon : l'opération apparaît déjà dans le vecteur de version $\langle A : 6, B : 1 \rangle$. L'opération est donc ignorée, et la résurgence de l'élément "I" illustrée dans la Figure 1.12 est évitée.

Il est à noter que dans le cas où un noeud reçoit une opération avec un dot plus élevé que celui attendu (e.g. le noeud A reçoit une opération *b3* à la fin de l'exemple), cette

FIGURE 3.4 – Gestion de la livraison *exactly-once* des opérations

opération est mise en attente. En effet, délivrer cette opération nécessiterait de mettre à jour le vecteur de version à $\langle A : 6, B : 3 \rangle$ et masquerait le fait que l'opération *b2* n'a jamais été reçue. L'opération *b3* serait donc mise en attente jusqu'à la livraison de l'opération *b2*.

Ainsi, l'implémentation de livraison *exactly-once* avec un vecteur de version comme structure de données force une livraison First In, First Out (FIFO) des opérations par noeuds. Il s'agit d'une contrainte non-nécessaire et qui peut introduire des délais dans la collaboration, notamment si une opération d'un noeud est perdue par le réseau. Nous jugeons cependant acceptable ce compromis entre le surcoût du mécanisme de livraison *exactly-once* et son impact sur l'expérience utilisateur.

Pour retirer cette contrainte superflue, il est possible de remplacer cette structure de données par un *Interval Version Vector* [44]. Au lieu d'enregistrer seulement le dernier dot observé par noeud, cette structure de données enregistre les intervalles de dots observés. Ceci permet une livraison *out of order* des opérations tout en garantissant une livraison *exactly-once* et en compactant efficacement les données stockées par le module **Sync** à terme.

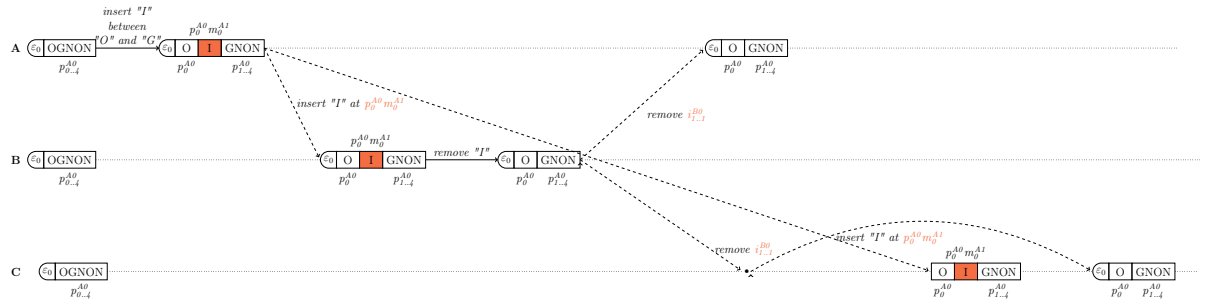
Livraison de l'opération *remove* après l'opération *insert*

La seconde contrainte que le modèle de livraison doit respecter spécifie qu'une opération *remove* doit être délivrée après les opérations *insert* insérant les éléments concernés.

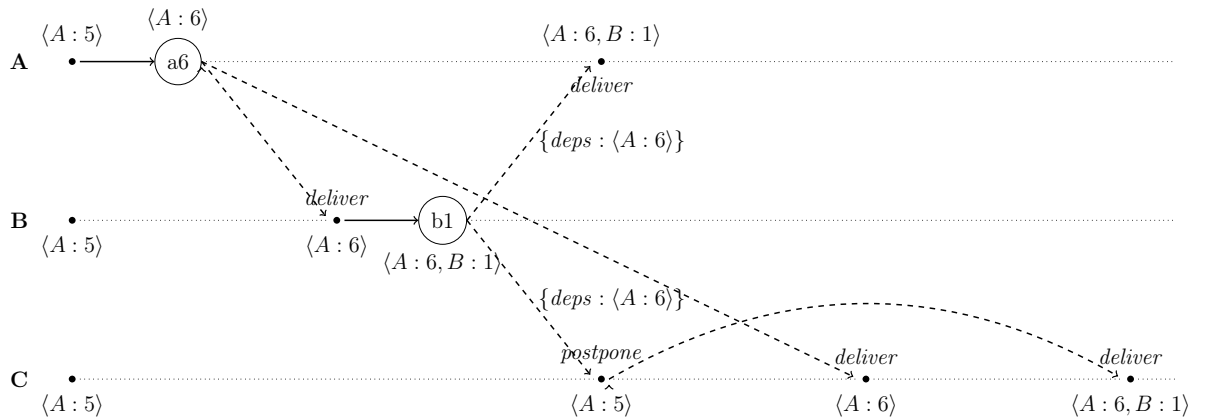
Pour cela, le module **Sync** ajoute un ensemble *Deps* à chaque opération *remove* avant de la diffuser :

Définition 24 (Deps) *Deps* est un ensemble d'opérations. Il représente l'ensemble des opérations dont dépend l'opération *remove* et qui doivent donc être livrées au préalable.

Plusieurs structures de données sont adaptées pour représenter les dépendances de l'opération *remove*. Dans le cadre de MUTE, nous avons choisi d'utiliser un ensemble de dots : pour chaque élément supprimé par l'opération *remove*, nous identifions le noeud l'ayant inséré et nous ajoutons le dot correspondant à l'opération la plus récente de ce noeud à l'ensemble des dépendances. Cette approche nous permet de limiter à un dot par élément supprimé le surcoût en métadonnées des dépendances et de les calculer en un temps linéaire par rapport au nombre d'éléments supprimés. Nous illustrons le calcul et l'utilisation des dépendances de l'opération *remove* à l'aide de la Figure 3.5.



(a) Exécution avec livraison dans le désordre d'une insertion et de sa suppression



(b) État et comportement de la couche Sync au cours de l'exécution décrite en Figure 3.5a

FIGURE 3.5 – Gestion de la livraison *causale-remove* des opérations

Cet exemple reprend et complète celui de la Figure 3.5. Trois noeuds A, B et C répliquent et éditent collaborativement une séquence. Les trois noeuds partagent le même

état initial : une séquence contenant les éléments "OGNON" et un vecteur de version $\langle A : 5 \rangle$.

Le noeud A insère l'élément "I" entre les éléments "O" et "G". Cet élément se voit attribué l'identifiant $p_o^{A0} m_o^{A1}$. L'opération correspondante $a6$ est diffusée aux autres noeuds.

À la réception de cette dernière, le noeud B supprime l'élément "I" nouvellement inséré et génère l'opération $b1$ correspondante. Comme indiqué précédemment, l'opération $b1$ étant une opération *remove*, le module **Sync** calcule ses dépendances avant de la diffuser. Pour chaque élément supprimé ("I"), **Sync** récupère l'identifiant de l'élément ($p_o^{A0} m_o^{A1}$) et en extrait l'identifiant du noeud qui l'a inséré (A). **Sync** ajoute alors le dot de l'opération la plus récente reçue de ce noeud ($\langle A : 6 \rangle$) à l'ensemble des dépendances de l'opération. L'opération est ensuite diffusée.

À la réception de l'opération $b1$, le noeud A vérifie s'il possède l'ensemble des dépendances de l'opération. Le noeud A ayant déjà observé l'opération $a6$, le module **Sync** délivre l'opération $b1$ au CRDT.

À l'inverse, lorsque le noeud C reçoit l'opération $b1$, il n'a pas encore reçu l'opération $a6$. L'opération $b1$ est alors mise en attente. À la réception de l'opération $a6$, celle-ci est délivrée. Le module **Sync** ré-évalue alors le cas de l'opération $b1$ et détermine qu'elle peut à présent être délivrée.

Il est à noter que notre approche pour générer l'ensemble des dépendances est une approximation. En effet, nous ajoutons les dots des opérations les plus récentes des auteurs des éléments supprimés. Nous n'ajoutons pas les dots des opérations qui ont spécifiquement insérés les éléments supprimés. Pour cela, il serait nécessaire de parcourir le log des opérations à la recherche des opérations *insert* correspondante. Cette méthode serait plus coûteuse, sa complexité dépendant du nombre d'opérations dans le log d'opérations, et incompatible avec un mécanisme tronquant le log des opérations en utilisant la stabilité causale. Notre approche introduit un potentiel délai dans la livraison d'une opération *remove* par rapport à une livraison utilisant ses dépendances exactes, puisqu'elle va reposer sur des opérations plus récentes et potentiellement encore inconnues par le noeud. Mais il s'agit là aussi d'un compromis que nous jugeons acceptable entre le surcoût du mécanisme de livraison et l'expérience utilisateur.

Livraison des opérations après l'opération *rename* introduisant leur époque

La troisième contrainte spécifiée par le modèle de livraison est qu'une opération doit être délivrée après l'opération *rename* qui a introduite son époque de génération.

Pour cela, le module **Sync** doit donc récupérer l'époque courante de la séquence répliquée, récupérer le dot de l'opération *rename* l'ayant introduite et l'ajouter en tant que dépendance de chaque opération. Cependant, dans notre implémentation, le module **Sync** et le module représentant la séquence répliquée sont découplés et ne peuvent interagir directement l'un avec l'autre.

Pour remédier à ce problème, le module **Sync** maintient une structure supplémentaire : un vecteur des dots des opérations *rename* connues. À la réception d'une opération *rename* distante, l'entrée correspondante de son auteur est mise à jour avec le dot de la nouvelle époque introduite. À la génération d'une opération locale, l'opération est examinée pour récupérer son époque de génération. **Sync** conserve alors seulement l'entrée correspondante

dans le vecteur des dots des opérations *rename*. À ce stade, le contenu du vecteur est ajouté en tant que dépendance de l'opération. Ensuite, si l'opération locale s'avère être une opération *rename*, le vecteur est modifié pour ne conserver que le dot de l'époque introduite par l'opération. La Figure 3.6 illustre ce fonctionnement.

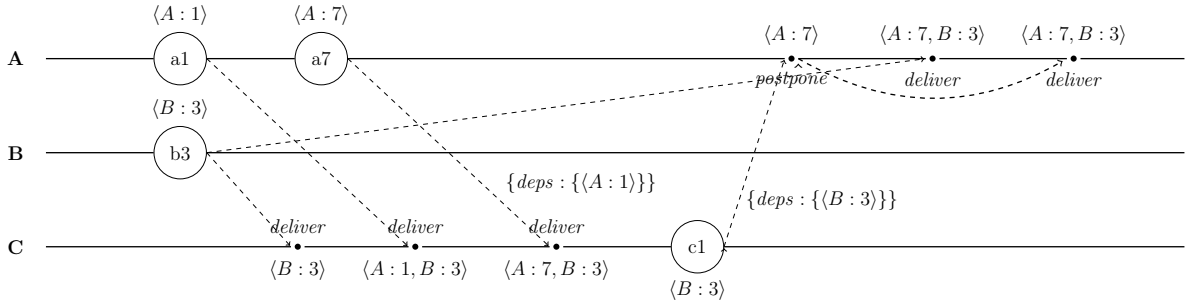
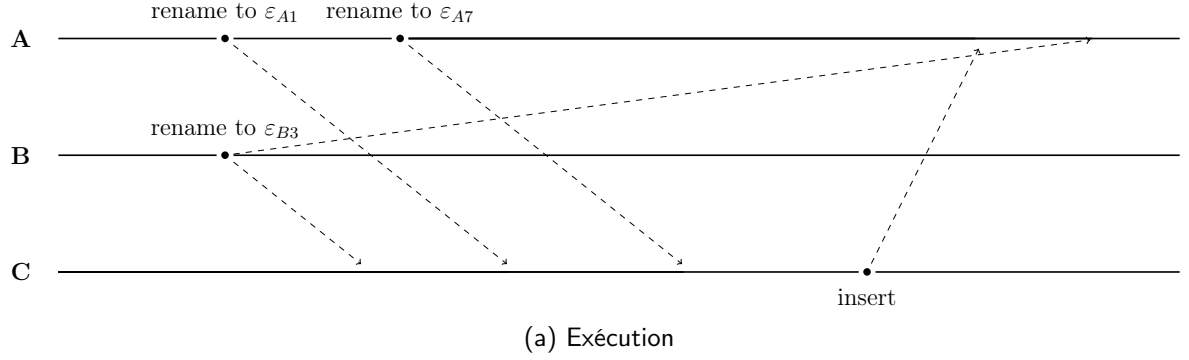


FIGURE 3.6 – Gestion de la livraison *epoch based* des opérations

Dans la Figure 3.6a, nous décrivons une exécution suivante en ne faisant apparaître que les opérations importantes : les opérations *rename* et une opération *insert* finale. Dans cette exécution, trois noeuds A, B et C répliquent et éditent collaborativement une séquence. Initialement, aucune opération *rename* n'a encore eu lieu. Le noeud A effectue une première opération *rename* (*a1*) puis une seconde opération *rename* (*a7*), et les diffuse. En concurrence, le noeud B génère et propage sa propre opération *rename* (*b3*). De son côté, le noeud C reçoit les opérations *b3*, puis *a1* et *a7*. Il émet ensuite une opération *insert* (*c1*). Le noeud A reçoit cette opération avant de finalement recevoir l'opération *b3*.

Dans la Figure 3.6b, nous faisons apparaître l'état de **Sync** et les décisions prises par ce dernier au cours de l'exécution. Initialement, le vecteur des dots des opérations *rename* connues est vide. Ainsi, lorsque A génère l'opération *a1*, celle-ci ne se voit ajouter aucune dépendance (nous ne représentons pas les dépendances des opérations qui correspondent à l'ensemble vide). A met ensuite à jour son vecteur des dots des opérations *rename* avec le dot $\langle A : 1 \rangle$. B procède de manière similaire avec l'opération *b3*.

Quand A génère l'opération *a7*, le dot $\langle A : 1 \rangle$ est ajouté en tant que dépendance. Le dot $\langle A : 7 \rangle$ remplace ensuite ce dernier dans le vecteur des dots des opérations *rename*.

À la réception de l'opération *b3*, le module **Sync** de C peut la délivrer au CRDT,

l'ensemble de ses dépendances étant vérifié. Le noeud C ajoute alors à son vecteur des dots des opérations *rename* le dot $\langle B : 3 \rangle$. Il procède de même pour l'opération *a1* : il la délivre et ajoute le dot $\langle A : 1 \rangle$. Le module **Sync** ne connaissant pas l'époque courante de la séquence répliquée, il maintient les deux dots localement.

Lorsque le noeud C reçoit l'opération *a7*, l'ensemble de ses contraintes est vérifié : l'opération *a1* a été délivrée précédemment. L'opération est donc délivrée et le vecteur de dots des opérations *rename* mis à jour avec $\langle A : 7 \rangle$.

Quand le noeud C effectue l'opération locale *c1*, le module **Sync** obtient l'information de l'époque courante de la séquence : ε_{b3} . C met à jour son vecteur de dots des opérations *rename* pour ne conserver que l'entrée du noeud B : $\langle B : 3 \rangle$. Ce dot est ajouté en tant que dépendance de l'opération *c1* avant sa diffusion.

À la réception de l'opération *c1* par le noeud A, cette opération est mise en attente par **Sync**, l'opération *b3* n'ayant pas encore été délivrée. Le noeud reçoit ensuite l'opération *b3*. Son vecteur des dots des opérations *rename* est mis à jour et l'opération délivrée. Les conditions pour l'opération *c1* étant désormais remplies, l'opération est alors délivrée.

Cette implémentation de la contrainte de la livraison *epoch-based* dispose de plusieurs avantages : sa complexité spatiale dépend linéairement du nombre de noeuds et les opérations de mise à jour du vecteur des dots des opérations *rename* s'effectuent en temps constant. De plus, seul un dot est ajouté en tant que dépendance des opérations, la taille du vecteur des dots étant ramené à 1 au préalable. Finalement, cette implémentation ne contraint pas une livraison causale des opérations *rename* et permet donc de les appliquer dès que possible.

Livraison des opérations à terme

La dernière contrainte du modèle de livraison précise que toutes les opérations doivent être délivrées à tous les noeuds à terme. Cependant, le réseau étant non-fiable, des messages peuvent être perdus au cours de l'exécution. Il est donc nécessaire que les noeuds rediffusent les messages perdus pour assurer leur livraison à terme.

Pour cela, nous implémentons un mécanisme d'anti-entropie basé sur [58]. Ce mécanisme permet à un noeud source de se synchroniser avec un autre noeud cible. Il est exécuté par l'ensemble des noeuds de manière indépendante. Nous décrivons ci-dessous son fonctionnement.

De manière périodique, le noeud choisit un autre noeud cible de manière aléatoire. Le noeud source lui envoie alors une représentation de son état courant, c.-à-d. son vecteur de version.

À la réception de ce message, le noeud cible compare le vecteur de version reçu par rapport à son propre vecteur de version. À partir de ces données, il identifie les dots des opérations de sa connaissance qui sont inconnues au noeud source. Grâce à leur dot, le noeud cible retrouve ces opérations depuis son log des opérations. Il envoie alors une réponse composée de ces opérations au noeud source.

À la réception de la réponse, le noeud source intègre normalement les opérations reçues. La Figure 3.7 illustre ce mécanisme.

Dans cette figure, nous représentons une exécution à laquelle participent trois noeuds : A, B et C. Initialement, les trois noeuds sont synchronisés. Leur vecteurs de version sont

3.2. Couche réplication

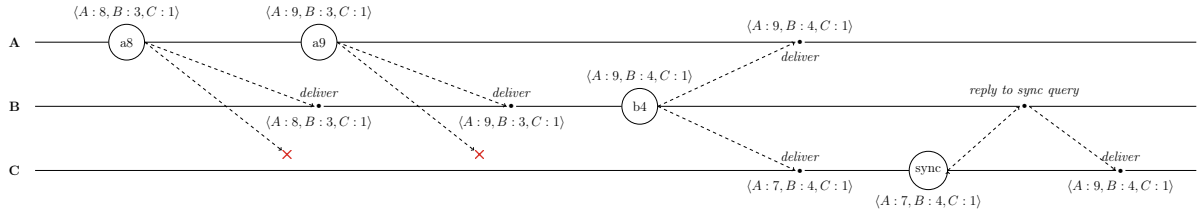


FIGURE 3.7 – Utilisation du mécanisme d’anti-entropie par le noeud C pour se synchroniser avec le noeud B

identiques et ont pour valeur $\langle A : 7, B : 3, C : 1 \rangle$.

Le noeud A effectue les opérations $a8$ puis $a9$ et les diffusent sur le réseau. Le noeud B reçoit ces opérations et les délivre à son CRDT. Il effectue ensuite et propage l’opération $b4$, qui est reçue et délivrée par A. Ils atteignent tous deux la version représenté par le vecteur $\langle A : 9, B : 4, C : 1 \rangle$

De son côté, le noeud C ne reçoit pas les opérations $a8$ et $a9$ à cause d’une défaillance réseau. Néanmoins, cela ne l’empêche pas de délivrer l’opération $b4$ à sa réception et d’obtenir la version $\langle A : 7, B : 4, C : 1 \rangle$.

Le noeud C déclenche ensuite son mécanisme d’anti-entropie. Il choisit aléatoirement le noeud B comme noeud cible. Il lui envoie un message de synchronisation avec pour contenu le vecteur de version $\langle A : 7, B : 8, C : 1 \rangle$.

À la réception de ce message, le noeud B compare ce vecteur avec le sien. Il détermine que le noeud C n’a pas reçu les opérations $a8$ et $a9$. B les récupère depuis son log des opérations et les envoie à C par le biais d’un nouveau message.

À la réception de la réponse de B, le noeud C délivre les opérations $a8$ et $a9$. Il atteint alors le même état que A et B, représenté par le vecteur de version $\langle A : 9, B : 4, C : 1 \rangle$.

Ce mécanisme d’anti-entropie nous permet ainsi de garantir la livraison à terme de toutes les opérations et de compenser les défaillances du réseau. Il nous sert aussi de mécanisme de synchronisation : à la connexion d’un pair, celui-ci utilise ce mécanisme pour récupérer les opérations effectuées depuis sa dernière connexion. Dans le cas où il s’agit de la première connexion du pair, il lui suffit d’envoyer un vecteur de version vide pour récupérer l’intégralité des opérations.

Ce mécanisme propose plusieurs avantages. Son exécution n’implique que le noeud source et le noeud cible, ce qui limite les coûts de coordination. De plus, si une défaillance a lieu lors de l’exécution du mécanisme (perte d’un des messages, panne du noeud cible...), cette défaillance n’est pas critique : le noeud source se synchronisera à la prochaine exécution du mécanisme. Ensuite, ce mécanisme réutilise le vecteur de version déjà nécessaire pour la livraison *exactly-once*, comme présenté en section 3.2.2. Il ne nécessite donc pas de stocker une nouvelle structure de données pour détecter les différences entre noeuds.

En contrepartie, la principale limite de ce mécanisme d’anti-entropie est qu’il nécessite de maintenir et de parcourir périodiquement le log des opérations pour répondre aux requêtes de synchronisation. La complexité spatiale et en temps du mécanisme dépend donc linéairement du nombre d’opérations. Qui plus est, nous sommes dans l’incapacité de tronquer le log des opérations en se basant sur la stabilité causale des opérations puisque nous utilisons ce mécanisme pour mettre à niveau les nouveaux pairs. À moins de mettre

en place un mécanisme de compression du log comme évoqué en sous-section 2.5.6, ce log des opérations croît de manière monotone. Néanmoins, une alternative possible est de mettre en place un système de chargement différé des opérations pour ne pas surcharger la mémoire.

3.2.3 Collaborateurs

Pour assurer la qualité de la collaboration même à distance, il est important d’offrir des fonctionnalités de conscience de groupe aux utilisateurs. Une de ces fonctionnalités est de fournir la liste des collaborateurs actuellement connectés. Les protocoles d’appartenance au réseau sont une catégorie de protocoles spécifiquement dédiée à cet effet. Ainsi, nous devons en implémenter un dans MUTE.

MUTE présente cependant plusieurs contraintes liées à notre modèle du système que le protocole sélectionné doit respecter. Tout d’abord, le protocole doit être compatible avec un environnement P2P, où les noeuds partagent les mêmes droits et responsabilités. De plus, le protocole doit présenter une capacité de passage à l’échelle pour être adapté aux collaborations à large échelle.

En raison de ces contraintes, notre choix s’est porté sur le protocole SWIM[19]. Proposé par DAS, GUPTA et MOTIVALA, ce protocole d’appartenance au réseau offre les propriétés intéressantes suivantes. Tout d’abord, le nombre de messages diffusés sur le réseau est proportionnel de façon linéaire au nombre de pairs. Pour être plus précis, le nombre de messages envoyés par un pair par période du protocole est constant. De plus, il fournit à chaque noeud une vue de la liste des collaborateurs cohérente à terme, même en cas de réception désordonnée des messages du protocole. Finalement, il intègre un mécanisme permettant de réduire le taux de faux positifs, c.-à-d. le taux de pairs déclarés injustement comme défaillants.

Pour cela, SWIM découple les deux composants d’un protocole d’appartenance au réseau : le mécanisme de *détection des défaillances des pairs* et le mécanisme de *dissémination des mises à jour du groupe*.

Mécanisme de détection des défaillances des pairs

Le mécanisme de détection des défaillances des pairs est exécuté de manière périodique, toutes les T unités de temps, par chacun des noeuds du système de manière non-coordonnée. Son fonctionnement est illustré par la Figure 3.8.

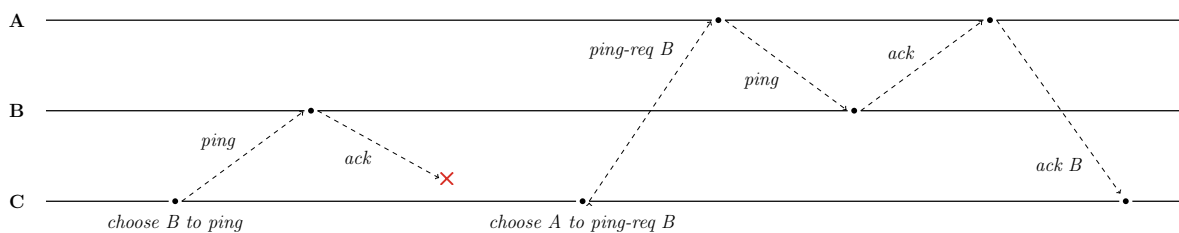


FIGURE 3.8 – Exécution du mécanisme de détection des défaillances par le noeud C pour tester le noeud B

Dans cet exemple, le réseau est composé des trois noeuds A, B et C. Le noeud C démarre l'exécution du mécanisme de détection des défaillances.

Tout d'abord, le noeud C sélectionne un noeud cible de manière aléatoire, ici B, et lui envoie un message *ping*. À la réception de ce message, le noeud B lui signifie qu'il est toujours opérationnel en lui répondant avec un message *ack*. À la réception de ce message par C, cette exécution du mécanisme de détection des défaillances prendrait fin. Mais dans l'exemple présenté ici, ce message est perdu par le réseau.

En l'absence de réponse de la part de B au bout d'un temps spécifié au préalable, le noeud C passe à l'étape suivante du mécanisme. Le noeud C sélectionne un autre noeud, ici A, et lui demande de vérifier via le message *ping-req B* si B a eu une défaillance. À la réception de la requête de ping, le noeud A envoie un message *ping* à B. Comme précédemment, B répond au *ping* par le biais d'un *ack* à A. A informe alors C du bon fonctionnement du B via le message *ack B*. Le mécanisme prend alors fin, jusqu'à sa prochaine exécution.

Si C n'avait pas reçu de réponse suite à sa *ping-req B* envoyée à A, C aurait supposé que B a eu une défaillance. Afin de réduire le taux de faux positifs, SWIM ne considère pas directement les noeuds n'ayant pas répondu comme en panne : ils sont tout d'abord *suspectés* d'être en panne. Après un certain temps sans signe de vie d'un noeud suspecté d'être en panne, le noeud est *confirmé* comme défaillant.

L'information qu'un noeud est suspecté d'être en panne est propagé dans le réseau via le mécanisme de dissémination des mises à jour du groupe décrit ci-dessous. Si un noeud apprend qu'il est suspecté d'une panne, il dissémine à son tour l'information qu'il est toujours opérationnel pour éviter d'être confirmé comme défaillant.

Pour éviter qu'un message antérieur n'invalide une suspicion d'une défaillance et retarde ainsi sa détection, SWIM introduit un numéro d'*incarnation*. Chaque noeud maintient un numéro d'incarnation. Lorsqu'un noeud apprend qu'il est suspecté d'une panne, il incrémente son numéro d'incarnation avant de propager l'information contradictoire.

Ainsi, afin de représenter la liste des collaborateurs, le protocole SWIM utilise la structure de données présentée par la Définition 25 :

Définition 25 (Liste des collaborateurs) *La liste des collaborateurs est un ensemble de triplets $\langle nodeId, nodeStatus, nodeIncarn \rangle$ où*

- *$nodeId$ correspond à l'identifiant du noeud correspondant à ce tuple.*
- *$nodeStatus$ correspond au statut courant du noeud correspondant à ce tuple, c.-à-d. *Alive* s'il est considéré comme opérationnel, *Suspect* s'il est suspecté d'une défaillance, *Confirm* s'il est considéré comme défaillant.*
- *$nodeIncarn$ correspond au numéro d'incarnation maximal, c.-à-d. le plus récent, connu pour le noeud correspondant à ce tuple.*

Chaque noeud réplique cette liste et la fait évoluer au cours de l'exécution du mécanisme présenté jusqu'ici. Lorsqu'une mise à jour est effectuée, celle-ci est diffusée de la manière présentée ci-dessous.

Mécanisme de dissémination des mises à jour du groupe

Quand l'exécution du mécanisme de détection des défaillances par un noeud met en lumière une évolution de la liste des collaborateurs, cette mise à jour doit être propagée au reste des noeuds.

Or, diffuser cette mise à jour à l'ensemble du réseau serait coûteux pour un seul noeud. Afin de propager cette information de manière efficace, SWIM propose d'utiliser un protocole de diffusion épidémique : le noeud transmet la mise à jour qu'à un nombre réduit λ^{10} de pairs, qui se chargeront de la transmettre à leur tour. Le mécanisme de dissémination des mises à jour de SWIM fonctionne donc de la manière suivante.

Chaque mise à jour du groupe est stockée dans une liste et se voit attribuer un compteur, initialisé avec $\lambda \log n$ avec n le nombre de noeuds. À chaque génération d'un message pour le mécanisme de détection des défaillances, un nombre arbitraire de mises à jour sont sélectionnées dans la liste et attachées au message. Leur compteurs respectifs sont décrémentés. Une fois que le compteur d'une mise à jour atteint 0, celle-ci est retirée de la liste.

À la réception d'un message, le noeud le traite comme définit précédemment en section 3.2.3. De manière additionnelle, il intègre dans sa liste des collaborateurs les mises à jour attachées au message en utilisant la règle suivante :

$$\forall i, j, k \cdot i \leq j \cdot \langle Alive, i \rangle < \langle Suspect, j \rangle < \langle Confirm, k \rangle$$

Ainsi, le mécanisme de dissémination des mises à jour du groupe réutilise les messages du mécanisme de détection des défaillances pour diffuser les modifications. Cela permet de propager les évolutions de la liste des collaborateurs sans ajouter de message supplémentaire. De plus, les règles de précedence sur l'état d'un collaborateur permettent aux noeuds de converger même si les mises à jour sont reçues dans un ordre distinct.

Modifications apportées

Nous avons ensuite apporté plusieurs modifications à la version du protocole SWIM présentée dans [19]. Notre première modification porte sur l'ordre de priorité entre les états d'un pair.

Modification de l'ordre de précedence Dans la version originale, un pair désigné comme défaillant l'est de manière irrévocable. Ce comportement est dû à la règle de précedence suivante :

$$\forall i, j \in \mathbb{N}, \forall s \in \{Alive, Suspect\} \cdot \langle s, i \rangle < \langle Confirm, j \rangle$$

pour un noeud donné. Ainsi, un noeud déclaré comme défaillant par un autre noeud doit changer d'identité pour rejoindre de nouveau le groupe.

10. [19] montre que choisir une valeur constante faible comme λ suffit néanmoins à garantir la dissémination des mises à jour à l'ensemble du réseau.

Ce choix n'est cependant pas anodin : il implique que la taille de la liste des collaborateurs croît de manière linéaire avec le nombre de connexions. S'agissant du paramètre avec le plus grand ordre de grandeur de l'application, nous avons cherché à le diminuer.

Nous avons donc modifié les règles de précedence de la manière suivante :

$$\forall i, j \in \mathbb{N}, i < j, \forall s, t \in \{Alive, Suspect, Confirm\} \cdot \langle i, s \rangle < \langle j, t \rangle$$

et

$$\forall i \in \mathbb{N} \cdot \langle i, Alive \rangle < \langle i, Suspect \rangle < \langle i, Confirm \rangle$$

Ces modifications permettent de donner la précedence au numéro d'incarnation, et d'utiliser le statut du collaborateur pour trancher seulement en cas d'égalité par rapport au numéro d'incarnation actuel. Ceci permet à un noeud auparavant déclaré comme défaillant de revenir dans le groupe en incrémentant son numéro d'incarnation. La taille de la liste des collaborateurs devient dès lors linéaire par rapport au nombre de noeuds.

Ces modifications n'ont pas d'impact sur la convergence des listes des collaborateurs des différents noeuds. Une étude approfondie reste néanmoins à effectuer pour déterminer si ces modifications ont un impact sur la vitesse à laquelle un noeud défaillant est déterminé comme tel par l'ensemble des noeuds.

Ajout d'un mécanisme de synchronisation La seconde modification que nous avons effectué concerne l'ajout d'un mécanisme de synchronisation entre pairs. En effet, le papier ne précise pas de procédure particulière lorsqu'un nouveau pair rejoint le réseau. Pour obtenir la liste des collaborateurs, ce dernier doit donc la demander à un autre pair.

Nous avons donc implémenté pour la liste des collaborateurs un mécanisme similaire à celui présenté en section 3.2.2 : à sa connexion, puis de manière périodique, un noeud envoie une requête de synchronisation à un noeud cible choisi de manière aléatoire. Ce message sert aussi à transmettre l'état courant du noeud source au noeud cible. En réponse, le noeud cible lui envoie l'état courant de sa liste. À la réception de cette dernière, le noeud source fusionne la liste reçue avec sa propre liste. Cette fusion conserve l'entrée la plus récente pour chaque noeud.

Pour récapituler, les mises à jour du groupe sont diffusées de manière atomique de façon épidémique, en utilisant les messages du mécanisme de détection des défaillances des noeuds. De manière additionnelle, un mécanisme d'anti-entropie permet à deux noeuds de synchroniser leur état. Ce mécanisme nous permet de pallier aux défaillances éventuelles du réseau. Ainsi, nous avons dans les faits mis en place un CRDT synchronisé par différences pour la liste des collaborateurs.

Synthèse

Pour générer et maintenir la liste des collaborateurs, nous avons implémenté le protocole distribué d'appartenance au réseau SWIM[19]. Par rapport à la version originale, nous avons procédé à plusieurs modifications, notamment pour gérer plus efficacement les reconnections successives d'un même noeud.

Ainsi, nous avons implémenté un mécanisme dont la complexité spatiale dépend linéairement du nombre de noeuds. Sa complexité en temps et sa complexité en communication, elles, sont indépendantes de ce paramètre. Elles dépendent en effet de paramètres dont nous choisissons les valeurs : la fréquence de déclenchement du mécanisme de détection de défaillance et le nombre de mises à jour du groupe propagées par message.

Des améliorations au protocole SWIM furent proposées dans [17]. Ces modifications visent notamment à réduire le délai de détection d'un noeud défaillant, ainsi que réduire le taux de faux positifs. Ainsi, une perspective est d'implémenter ces améliorations dans MUTE.

3.2.4 Curseurs

Toujours dans le but d'offrir des fonctionnalités de conscience de groupe aux utilisateurs pour leur permettre de se coordonner aisément, nous avons implémenté dans MUTE l'affichage des curseurs distants.

Pour représenter fidèlement la position des curseurs des collaborateurs distants, nous nous reposons sur les identifiants du CRDT choisi pour représenter la séquence. Le fonctionnement est similaire à la gestion des modifications du document : lorsque l'éditeur indique que l'utilisateur a déplacé son curseur, nous récupérons son nouvel index. Nous recherchons ensuite l'identifiant correspondant à cet index dans la séquence répliquée et le diffusons aux collaborateurs.

À la réception de la position d'un curseur distant, nous récupérons l'index correspondant à cet identifiant dans la séquence répliquée et représentons un curseur à cet index. Il est intéressant de noter que si l'identifiant a été supprimé en concurrence, nous pouvons à la place récupérer l'index de l'élément précédent et ainsi indiquer à l'utilisateur où son collaborateur est actuellement en train de travailler.

De façon similaire, nous gérons les sélections de texte à l'aide de deux curseurs : un curseur de début et un curseur de fin de sélection.

3.3 Couche réseau

Pour permettre aux différents noeuds de communiquer, MUTE repose sur la librairie Netflux¹¹. Développée au sein de l'équipe Coast, cette librairie permet de construire un réseau P2P entre des navigateurs, mais aussi des bots.

3.3.1 Établissement d'un réseau P2P entre navigateurs

Pour créer un réseau P2P entre navigateurs, Netflux utilise la technologie Web Real-Time Communication (WebRTC). WebRTC est une API¹² de navigateur spécifiée en 2011, et en cours d'implémentation dans les différents navigateurs depuis 2013. Elle permet de créer une connexion directe entre deux navigateurs pour échanger des médias audio et/ou vidéo, ou simplement des données.

11. <https://github.com/coast-team/netflux>

12. Application Programming Interface (API) : Interface de Programmation

Cette API utilise pour cela un ensemble de protocoles. Ces protocoles réintroduisent des serveurs dans l'architecture système de MUTE. Dans la Figure 3.9, nous représentons un réseau P2P créé avec WebRTC et les différents serveurs impliqués.

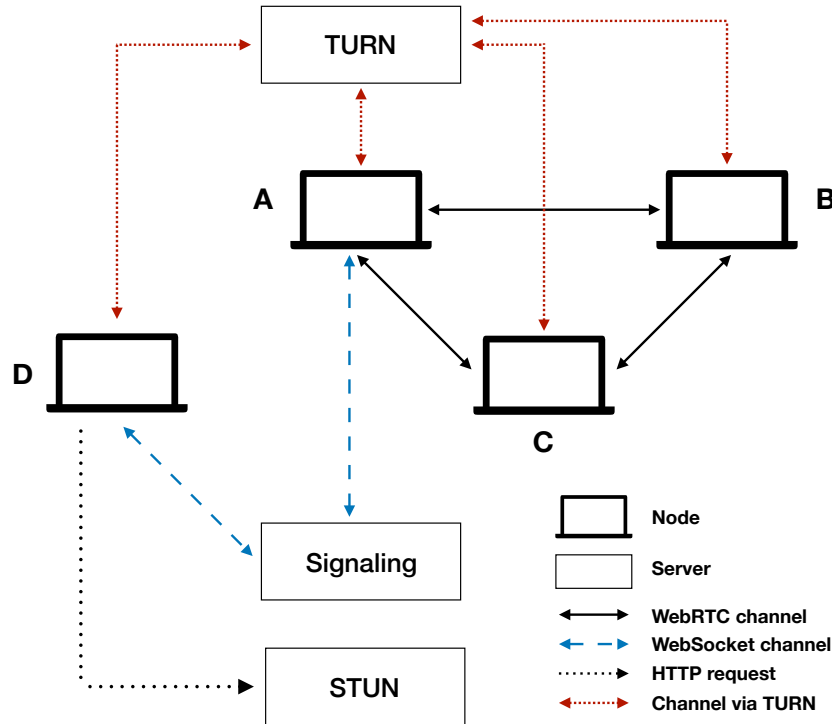


FIGURE 3.9 – Architecture système pour la couche réseau de MUTE

Nous décrivons ci-dessous leur rôle respectif dans la collaboration.

Serveur de signalisation

Pour rejoindre un réseau P2P déjà établi, un nouveau nœud a besoin de découvrir les nœuds déjà connectés et de pouvoir communiquer avec eux. Le serveur de signalisation offre ces fonctionnalités.

Au moins un nœud du réseau P2P doit maintenir une connexion avec le serveur de signalisation. À sa connexion, un nouveau nœud contacte le serveur de signalisation. Il est mis en relation avec un nœud du réseau P2P par son intermédiaire et échange les différents messages de WebRTC nécessaires à l'établissement d'une connexion P2P entre eux.

Une fois cette première connexion P2P établie, le nouveau nœud contacte et communique avec les autres nœuds par l'intermédiaire du premier nœud. Il peut alors terminer sa connexion avec le serveur de signalisation.

Serveur STUN

Pour se connecter, les nœuds doivent s'échanger plusieurs informations logicielles et matérielles, notamment leur adresse IP publique respective. Cependant, un nœud n'a pas

accès à cette donnée lorsque son routeur utilise le protocole NAT. Le noeud doit alors la récupérer.

Pour permettre aux noeuds de découvrir leur adresse IP publique, WebRTC repose sur le protocole STUN. Ce protocole consiste simplement à contacter un serveur tiers dédié à cet effet. Ce serveur retourne en réponse au noeud qui le contacte son adresse IP publique.

Serveur TURN

Il est possible que des noeuds provenant de réseaux différents ne puissent établir une connection P2P directe entre eux, par exemple à cause de restrictions imposées par leur pare-feux respectifs. Pour contourner ce cas de figure, WebRTC utilise le protocole TURN.

Ce protocole consiste à utiliser un serveur tiers comme relais entre les noeuds. Ainsi, les noeuds peuvent communiquer par son intermédiaire tout au long de la collaboration. Les échanges sont chiffrés, afin que le serveur TURN ne représente pas une faille de sécurité.

Rôle des serveurs

Ainsi, WebRTC implique l'utilisation de plusieurs serveurs.

Les serveurs de signalisation et STUN sont nécessaires pour permettre à de nouveaux noeuds de rejoindre la collaboration. Autrement dit, leur rôle est ponctuel : une fois le réseau P2P établi, les noeuds n'ont plus besoin d'eux. Ces serveurs peuvent alors être coupés sans impacter la collaboration.

À l'inverse, les serveurs TURN jouent un rôle plus prédominant dans la collaboration. Ils sont nécessaires dès lors que des noeuds proviennent de réseaux différents et sont alors requis tout au long de la collaboration. Une panne de ces derniers entraverait la collaboration puisqu'elle résulterait en une partition des noeuds. Il est donc primordial de s'assurer de la disponibilité et fiabilité de ces serveurs.

3.3.2 Topologie réseau

Netflux établit un réseau P2P par document. Chaque réseau P2P est un réseau entièrement maillé : chaque noeud se connecte à l'ensemble des autres noeuds.

Cette topologie simple est adaptée à des groupes de petite taille, mais ne passe pas à l'échelle. D'autres topologies limitant le nombre de connexions par noeuds, telle que celle décrite par [47], pourraient être implémentées pour adresser cette limite.

3.4 Couche sécurité

La couche sécurité a pour but de garantir l'authenticité et la confidentialité des messages échangés par les noeuds. Pour cela, elle implémente un mécanisme de chiffrement de bout en bout.

Pour chiffrer les messages, MUTE utilise un mécanisme de chiffrement à base de clé de groupe. Le protocole choisi est le protocole Burmester-Desmedt[14]. Il nécessite que

chaque noeud possède une paire de clés de chiffrement et enregistre sa clé publique auprès d'un PKI¹³.

Afin d'éviter qu'un PKI malicieux n'effectue une attaque de l'homme au milieu sur la collaboration, les noeuds doivent vérifier le bon comportement des PKI de manière non-coordonnée. À cet effet, MUTE implémente le mécanisme d'audit de PKI Trusternity[49, 50]. Son fonctionnement nécessite l'utilisation d'un registre publique sécurisé *append-only*, c.-à-d. une blockchain.

L'architecture système nécessaire pour la couche sécurité est présentée dans la Figure 3.10.

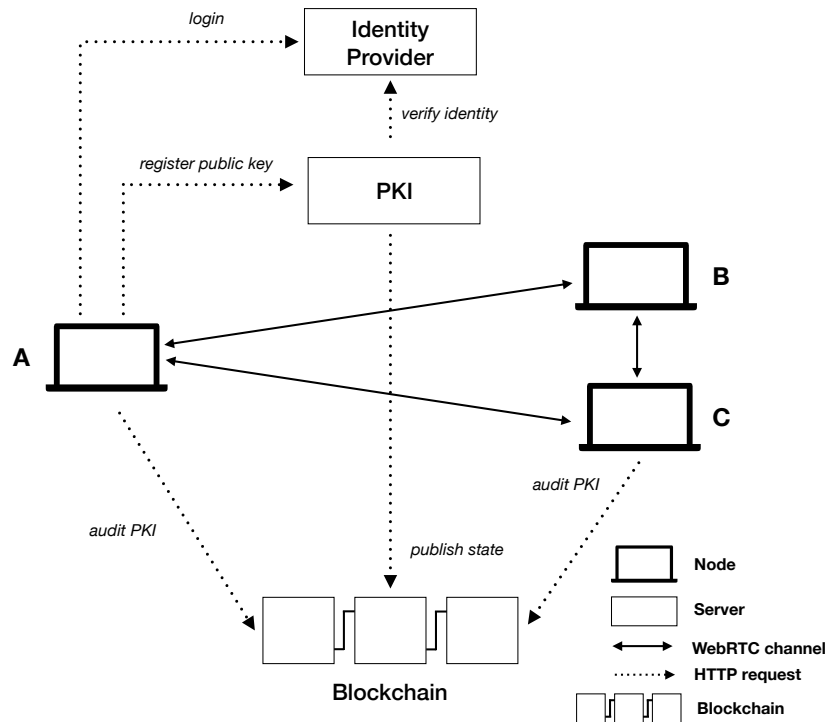


FIGURE 3.10 – Architecture système pour la couche sécurité de MUTE

Cette couche sécurité s'ajoute au mécanisme de chiffrement des messages inhérent à WebRTC. Cela nous offre de nouvelles possibilités : plutôt que de créer un réseau P2P par document, nous pouvons désormais mettre en place un réseau P2P global. Les messages étant chiffrés de bout en bout, les noeuds peuvent communiquer en toute sécurité et confidentialité par l'intermédiaire de noeuds tiers, c.-à-d. des noeuds extérieurs à la collaboration.

Une limite de l'approche actuelle est que la clé de groupe change avec l'évolution des noeuds connectés : à chaque connexion ou déconnexion d'un noeud, une nouvelle clé est recalculée avec les collaborateurs présents. Cette évolution fréquente de la clé de chiffrement, nécessaire pour garantir la *backward secrecy* et *forward secrecy*, nous empêche par exemple de stocker les opérations de manière chiffrée chez des noeuds tiers. Cette fonctionnalité serait cependant bien pratique pour permettre à un noeud de récupérer la

13. Public Key Infrastructure (PKI) : Infrastructure de gestion de clés

dernière version de ses documents, même en l'absence de ses collaborateurs. Une autre clé de chiffrement, dédiée au stockage, devrait être mise en place, ainsi qu'un mécanisme de découverte des noeuds tiers stockant les données de la collaboration.

3.5 Conclusion

Dans ce chapitre, nous avons présenté Multi User Text Editor (MUTE), notre éditeur collaboratif temps réel P2P chiffré de bout en bout.

MUTE permet d'éditer de manière collaborative des documents texte. Pour représenter les documents, MUTE implémente les structures de données répliquées décrites dans la section 1.5 et le chapitre 2. Ces CRDTs offrent de nouvelles méthodes de collaborer, notamment en permettant de collaborer de manière synchrone ou asynchrone de manière transparente.

Pour permettre aux noeuds de communiquer, MUTE utilise WebRTC. Cette technologie permet de construire un réseau P2P entre navigateurs. Plusieurs serveurs sont néanmoins requis, notamment pour la découverte des pairs et pour la communication entre des noeuds dont les pare-feux respectifs empêche l'établissement d'une connexion directe.

Finalement, MUTE implémente un mécanisme de chiffrement de bout en bout garantissant l'authenticité et la confidentialité des échanges entre les noeuds. Ce mécanisme reposant sur d'autres serveurs, les PKIs, MUTE intègre un mécanisme d'audit permettant de détecter leurs éventuels comportements malicieux.

Chapitre 4

Conclusions et perspectives

Sommaire

4.1	Résumé des contributions	98
4.2	Perspectives	98
4.2.1	Définition de relations de priorité pour minimiser les traitements	98
4.2.2	Redéfinition de la sémantique du renommage en déplacement d'éléments	98
4.2.3	Définition de types de données répliquées sans conflits plus complexes	98
4.2.4	Étude comparative des différentes familles de CRDTs	98
4.2.5	Définition d'opérations supplémentaires pour fonctionnalités liées à l'édition collaborative	99
4.2.6	Conduction d'expériences utilisateurs d'édition collaborative . .	99
4.2.7	Comparaison des mécanismes de synchronisation	100
4.2.8	Distance entre versions d'un document	100
4.2.9	Contrôle d'accès	100
4.2.10	Détection et éviction de pairs malhonnêtes	100
4.2.11	Vecteur <i>epoch-based</i>	101
4.2.12	Fusion de versions distantes d'un document collaboratif	102
4.2.13	Rôles et places des bots dans systèmes collaboratifs	102

4.1 Résumé des contributions

4.2 Perspectives

4.2.1 Définition de relations de priorité pour minimiser les traitements

4.2.2 Redéfinition de la sémantique du renommage en déplacement d'éléments

4.2.3 Définition de types de données répliquées sans conflits plus complexes

4.2.4 Étude comparative des différentes familles de CRDTs

- La spécification récente des Delta-based CRDTs . Ce nouveau type de CRDTs se base sur celui des State-based CRDTs. Partage donc les mêmes pré-requis :
 - États du type de données répliqué forment un sup-demi-treillis
 - Modifications locales entraînent une inflation de l'état
 - Possède une fonction de **merge**, permettant de fusionner deux états S et S' , et qui
 - Est associative, commutative et idempotente
 - Retourne S'' , la LUB de S et S' (c.-à-d. $\nexists S''' \cdot merge(S, S') < S''' < S''$)

Et bénéficie de son principal avantage : synchronisation possible entre deux pairs en fusionnant leur états, peu importe le nombre de modifications les séparant.

- Spécificité des Delta-based CRDTs est de proposer une synchronisation par différence d'états. Plutôt que de diffuser l'entièreté de l'état pour permettre aux autres pairs de se mettre à jour, idée est de seulement transmettre la partie de l'état ayant été mise à jour. Correspond à un élément irréductible du sup-demi-treillis. Permet ainsi de mettre en place une synchronisation en temps réel de manière efficace. Et d'utiliser la synchronisation par fusion d'états complets pour compenser les défaillances du réseau
- Ainsi, ce nouveau type de CRDTs semble allier le meilleur des deux mondes :
 - Absence de contrainte sur le réseau autre que la livraison à terme
 - Propagation possible en temps réel des modifications

Semble donc être une solution universelle :

- Utilisable peu importe la fiabilité réseau à disposition
- Empreinte réseau du même ordre de grandeur qu'un Op-based CRDT
- Utilisable peu importe la fréquence de synchronisation désirée

Pose la question de l'intérêt des autres types de CRDTs.

- Delta-based CRDT est un State-based CRDT dont on a identifié les éléments irréductibles et qui utilise ces derniers pour la propagation des modifications plutôt que l'état complet. Famille des State-based CRDTs semble donc rendue obsolète par celle des Delta-based CRDTs. À confirmer.
- Les Op-based CRDTs proposent une spécification différente du type répliqué de leur équivalent Delta-based, généralement plus simple. À première vue, famille des Op-based CRDTs semble donc avoir la simplicité comme avantage par rapport à celle des Delta-based CRDTs. S'agit d'un paramètre difficilement mesurable et auquel on peut objecter si on considère qu'un Op-based CRDT s'accompagne d'une couche livraison de messages, qui cache sa part de complexité. Intéressant d'étudier si la spécification différente des Op-based CRDTs présente d'autres avantages par rapport aux Delta-based CRDTs : performances (temps d'intégration des modifications, délai de convergence...), fonctionnalités spécifiques (composition, undo...)
- But serait de fournir des guidelines sur la famille de CRDT à adopter en fonction du cas d'utilisation.

4.2.5 Définition d'opérations supplémentaires pour fonctionnalités liées à l'édition collaborative

- Commentaires
- Suggestions

4.2.6 Conduction d'expériences utilisateurs d'édition collaborative

- Absence d'un dataset réel et réutilisable sur les sessions d'édition collaborative
- Généralement, expériences utilisent données d'articles de Wikipédia *Matthieu: TODO : Revoir références, mais me semble que c'est celui utilisé pour Logoot, LogootSplit et RGASplit entre autres.* Mais ces données correspondent à une exécution séquentielle, c.-à-d. aucune édition concurrente ne peut être réalisée avec le système de résolution de conflits de Wikipédia. *Matthieu: TODO : Me semble que Kleppmann a aussi utilisé et mis à disposition ses traces correspondant à la rédaction d'un de ses articles. Mais que cet article n'était rédigé que par lui. Peu de chances de présence d'édérations concurrentes. À retrouver et vérifier.*
- Inspiré par expériences de Claudia, pourrait mener des sessions d'édition collaborative sur des outils orchestrés pour produire ce dataset
- Devrait rendre ce dataset agnostique de l'approche choisie pour la résolution automatique de conflits
- Absence de retours sur les collaborations à grande échelle
- Comment on collabore lorsque plusieurs centaines d'utilisateur-rices ?

4.2.7 Comparaison des mécanismes de synchronisation

Serait intéressant de comparer à d'autres méthodes de synchronisation : mécanisme d'anti-entropie basé sur un Merkle Tree[21, 40, 28], synchronisation par états (state/delta-based CRDTs). Dans le cadre des Delta-based CRDTs, pourrait évaluer un protocole de diffusion épidémique des deltas comme celui proposé par SWIM[19].

4.2.8 Distance entre versions d'un document

- Est-ce que ça a vraiment du sens d'intégrer automatiquement des modifications ayant été généré sur une version du document distante de l'état actuel du document (voir distance de Hamming, Levenstein, String-to-string correction problem (Tichy et al))
- Jusqu'à quelle distance est-ce que la fusion automatique a encore du sens ? *Matthieu: NOTE : Peut connecter ça à la nécessité de conserver un chemin d'une époque à l'autre : si les opérations émises depuis cette époque ont probablement plus d'intérêt pour l'état actuel, couper l'arbre ?*

4.2.9 Contrôle d'accès

- Pour le moment, n'importe quel utilisateur ayant l'URL du document peut y accéder dans MUTE
- Pour des raisons de confidentialité, peut vouloir contrôler quels utilisateurs ont accès à un document
- Nécessite l'implémentation de liste de contrôle d'accès
- Mais s'agit d'une tâche complexe dans le cadre d'un système distribué
- Peut s'inspirer des travaux réalisés au sein de la communauté CRDTs [76, 63] pour cela

4.2.10 Détection et éviction de pairs malhonnêtes

- À l'heure actuelle, MUTE suppose qu'ensemble des collaborateurs honnêtes
- Vulnérable à plusieurs types d'attaques par des adversaires byzantins, tel que l'équivoque
- Ce type d'attaque peut provoquer des divergences durables et faire échouer des collaborations
- Dans [23, 22], ELVINGER propose un mécanisme permettant de maintenir des logs authentifiés dans un système distribué
- Les logs authentifiés permettent de mettre en lumière les comportements malveillants des adversaires et de borner le nombre d'actions malveillantes qu'ils peuvent effectuer avant d'être évincé
- Implémenter ce mécanisme permettrait de rendre compatible MUTE avec des environnements avec adversaires byzantins

- Nécessiterait tout de même de faire évoluer le CRDT pour résoudre les équivoques détectés

4.2.11 Vecteur *epoch-based*

- Comme présenté précédemment, nous utilisons plusieurs vecteurs pour représenter des données dans l'application MUTE
- Notamment pour le vecteur de version, utilisé pour respecter le modèle de livraison requis par le CRDT
- Et pour la liste des collaborateurs, utilisé pour offrir des informations nécessaires à la conscience de groupe aux utilisateurs
- Ces vecteurs sont maintenus localement par chacun des noeuds et sont échangés de manière périodique
- Cependant, la taille de ces vecteurs croît de manière linéaire au nombre de noeuds impliqués dans la collaboration
- Les systèmes P2P à large échelle sont sujets au *churn*
- Dans le cadre d'un tel système, ces structures croissent de manière non-bornée
- Ceci pose un problème de performances, notamment d'un point de vue consommation en bande-passante
- Cependant, même si on observe un grand nombre de pairs différents dans le cadre d'une collaboration à large échelle
- Intuition est qu'une collaboration repose en fait sur un petit noyau de collaborateurs principaux
- Et que majorité des collaborateurs se connectent de manière éphémère
- Serait intéressant de pouvoir réduire la taille des vecteurs en oubliant les collaborateurs éphémères
- Dynamo[21] tronque le vecteur de version lorsqu'il dépasse une taille seuil
- Conduit alors à une perte d'informations
- Pour la liste des collaborateurs, approche peut être adoptée (pas forcément gênant de limiter à 100 la taille de la liste)
- Mais pour vecteur de version, conduirait à une relivraison d'opérations déjà observées
- Approche donc pas applicable pour cette partie
- Autre approche possible est de réutiliser le système d'époque
- Idée serait de ACK un vecteur avec un changement d'époque
- Et de ne diffuser à partir de là que les différences
- Un mécanisme de transformation (une simple soustraction) permettrait d'obtenir le dot dans la nouvelle époque d'une opération concurrente au renommage
- Peut facilement mettre en place un mécanisme d'inversion du renommage (une simple addition) pour revenir à une époque précédente

- Et ainsi pouvoir circuler librement dans l'arbre des époques et gérer les opérations *rename* concurrentes
- Serait intéressant d'étudier si on peut aller plus loin dans le cadre de cette structure de données et notamment rendre commutatives les opérations de renommage concurrentes

4.2.12 Fusion de versions distantes d'un document collaboratif

4.2.13 Rôles et places des bots dans systèmes collaboratifs

- Stockage du document pour améliorer sa disponibilité
- Overleaf en P2P ?
- Comment réinsérer des bots dans la collaboration sans en faire des éléments centraux, sans créer des failles de confidentialité, et tout en rendant ces fonctionnalités accessibles ?

Annexe A

Algorithmes RENAMEID

Algorithme 5 Remaining functions to rename an identifier

```
function RENIDLESTHANFIRSTID(id, newFirstId)
  if id < newFirstId then
    return id
  else
    pos ← position(newFirstId)
    nId ← nodeId(newFirstId)
    nSeq ← nodeSeq(newFirstId)
    predNewFirstId ← new Id(pos, nId, nSeq, -1)

    return concat(predNewFirstId, id)
  end if
end function

function RENIDGREATERTHANLASTID(id, newLastId)
  if id < newLastId then
    return concat(newLastId, id)
  else
    return id
  end if
end function
```

Annexe B

Algorithmes REVERTRENAMEID

Algorithme 6 Remaining functions to revert an identifier renaming

```
function REVRENIDLESSTHANNEWFIRSTID(id, firstId, newFirstId)
  predNewFirstId  $\leftarrow$  createIdFromBase(newFirstId, -1)
  if isPrefix(predNewFirstId, id) then
    tail  $\leftarrow$  getTail(id, 1)
    if tail < firstId then
      return tail
    else
       $\triangleright$  id has been inserted causally after the rename op
      offset  $\leftarrow$  getLastOffset(firstId)
      predFirstId  $\leftarrow$  createIdFromBase(firstId, offset)
      return concat(predFirstId, MAX_TUPLE, tail)
    end if
  else
    return id
  end if
end function

function REVRENIDGREATERTHANNEWLASTID(id, lastId)
  if id < lastId then
     $\triangleright$  id has been inserted causally after the rename op
    return concat(lastId, MIN_TUPLE, id)
  else if isPrefix(newLastId, id) then
    tail  $\leftarrow$  getTail(id, 1)
    if tail < lastId then
       $\triangleright$  id has been inserted causally after the rename op
      return concat(lastId, MIN_TUPLE, tail)
    else if tail < newLastId then
      return tail
    else
       $\triangleright$  id has been inserted causally after the rename op
      return id
    end if
  else
    return id
  end if
end function
```

Index

Voici un index

FiXme :

Notes :

- 10 : Matthieu : TODO : Ajouter mention que OT a été abandonné à cause de cette contrainte même., 14
- 11 : Matthieu : TODO : Ajouter refs Scuttlebutt si applicable à Op-based, 15
- 12 : Matthieu : TODO : Ajouter refs, 16
- 13 : Matthieu : TODO : Vérifier que c'est bien le cas dans [11], 16
- 14 : Matthieu : TODO : Ajouter IPA, 18
- 15 : Matthieu : TODO : Revoir et ajouter Melda (PaPoC'22) si fitting, 18
- 16 : Matthieu : TODO : Ajouter Redis et Akka, 19
- 17 : Matthieu : TODO : Vérifier et ajouter l'article avec Digital Garden (PaPoC'22 ?) si fitting, 19
- 18 : Matthieu : NOTE : Ajouter LogootSplit de manière sommaire aussi à cet endroit ?, 19
- 19 : Matthieu : TODO : Autres Sequence CRDTs à considérer : Stringwise CRDT [78], Chronofold [31], 19
- 1 : Matthieu : TODO : Voir si angle écologique/réduction consommation d'énergie peut être pertinent., 1
- 20 : Matthieu : TODO : Ajouter une relation d'ordre sur les tuples, 20
- 21 : Matthieu : TODO : Définir la notion de base (et autres fonctions utiles sur les identifiants ? genre isPrefix, concat, getTail...), 21
- 22 : Matthieu : TODO : Montrer que cet ensemble d'identifiants est un ensemble dense, 21
- 23 : Matthieu : TODO : indiquer que le couple `hAnodeId`, `nodeSeqB` permet d'identifier de manière unique la base d'un bloc ou d'un identifiant, 22
- 24 : Matthieu : NOTE : Pourrait définir dans cette sous-section la notion de séquence bien-formée, 24
- 25 : Matthieu : QUESTION : Ajouter quelques lignes ici sur comment faire ça en pratique (Ajout d'un dot aux opérations, maintien d'un dot store au niveau de la couche livraison, vérification que dot pas encore présent dans dot store avant de passer opération à la structure de données) ? Ou je garde ça pour le chapitre sur MUTE ?, 25
- 26 : Matthieu : QUESTION : Même que pour la exactly-once delivery, est-ce que j'explique ici comment assurer cette contrainte plus en détails (Ajout des dots des opérations *insert* en dépendances de l'opération *remove*, vérification que dots présents dans dot store avant de passer l'opération *remove* à la structure de données) ou je garde ça pour le chapitre sur MUTE ?, 26
- 27 : Matthieu : TODO : Ajouter une phrase pour expliquer que la crois-

- sance des identifiants impacte aussi le temps d'intégration des modifications, 27
- 28 : Matthieu : TODO : Trouver référence sur la stabilité causale dans systèmes dynamiques, 28
- 29 : Matthieu : TODO : Ajouter une partie sur la discussion qu'on a pu avoir avec les reviewers sur la présence de pierres tombales dans RenamableLogootSplit, et comment ces pierres tombales diffèrent de celles présentent dans WOOT et RGA. , 66
- 2 : Matthieu : TODO : Mentionner TP1 et TP2, 4
- 30 : Matthieu : TODO : Revoir références, mais me semble que c'est celui utilisé pour Logoot, LogootSplit et RGASplit entre autres, 99
- 31 : Matthieu : TODO : Me semble que Kleppmann a aussi utilisé et mis à disposition ses traces correspondant à la rédaction d'un de ses articles. Mais que cet article n'était rédigé que par lui. Peu de chances de présence d'éditions concurrentes. À retrouver et vérifier. , 99
- 32 : Matthieu : NOTE : Peut connecter ça à la nécessité de conserver un chemin d'une époque à l'autre : si les opérations émises depuis cette époque ont probablement plus d'intérêt pour l'état actuel, couper l'arbre ?, 100
- 3 : Matthieu : TODO : Spécification faible et forte des séquences répliquées, 4
- 4 : Matthieu : TODO : insérer ref sticky model, 5
- 5 : Matthieu : TODO : Ajouter ref, 6
- 6 : Matthieu : TODO : Faire le lien avec les travaux de Burckhardt [13] et les MRDTs [37], 7
- 7 : Matthieu : TODO : Ajouter refs des horloges logiques plus intelligentes (Interval Tree Clock, Hybrid Clock...), 8
- 8 : Matthieu : TODO : Ajouter ref, 13
- 9 : Matthieu : TODO : Insérer ref, 14
- FiXme (Matthieu) :
- Notes :
- 10 : TODO : Ajouter mention que OT a été abandonné à cause de cette contrainte même., 14
- 11 : TODO : Ajouter refs Scuttlebutt si applicable à Op-based, 15
- 12 : TODO : Ajouter refs, 16
- 13 : TODO : Vérifier que c'est bien le cas dans [11], 16
- 14 : TODO : Ajouter IPA, 18
- 15 : TODO : Revoir et ajouter Melda (PaPoC'22) si fitting, 18
- 16 : TODO : Ajouter Redis et Akka, 19
- 17 : TODO : Vérifier et ajouter l'article avec Digital Garden (PaPoC'22 ?) si fitting, 19
- 18 : NOTE : Ajouter LogootSplit de manière sommaire aussi à cet endroit ?, 19
- 19 : TODO : Autres Sequence CRDTs à considérer : String-wise CRDT [78], Chronofold [31], 19
- 1 : TODO : Voir si angle écologique/réduction consommation d'énergie peut être pertinent., 1
- 20 : TODO : Ajouter une relation d'ordre sur les tuples, 20
- 21 : TODO : Définir la notion de base (et autres fonctions utiles sur les identifiants ? genre isPrefix, concat, getTail...), 21
- 22 : TODO : Montrer que cet ensemble d'identifiants est un ensemble dense, 21
- 23 : TODO : indiquer que le couple hAnodeId, nodeSeqB permet d'identifier de manière unique la base d'un bloc ou d'un identifiant, 22

- 24 : NOTE : Pourrait définir dans cette sous-section la notion de séquence bien-formée, 24
- 25 : QUESTION : Ajouter quelques lignes ici sur comment faire ça en pratique (Ajout d'un dot aux opérations, maintien d'un dot store au niveau de la couche livraison, vérification que dot pas encore présent dans dot store avant de passer opération à la structure de données) ? Ou je garde ça pour le chapitre sur MUTE ?, 25
- 26 : QUESTION : Même que pour la exactly-once delivery, est-ce que j'explique ici comment assurer cette contrainte plus en détails (Ajout des dots des opérations *insert* en dépendances de l'opération *remove*, vérification que dots présents dans dot store avant de passer l'opération *remove* à la structure de données) ou je garde ça pour le chapitre sur MUTE ?, 26
- 27 : TODO : Ajouter une phrase pour expliquer que la croissance des identifiants impacte aussi le temps d'intégration des modifications, 27
- 28 : TODO : Trouver référence sur la stabilité causale dans systèmes dynamiques, 28
- 29 : TODO : Ajouter une partie sur la discussion qu'on a pu avoir avec les reviewers sur la présence de pierres tombales dans RenamableLogootSplit, et comment ces pierres tombales diffèrent de celles présentent dans WOOT et RGA. , 66
- 2 : TODO : Mentionner TP1 et TP2, 4
- 30 : TODO : Revoir références, mais me semble que c'est celui utilisé pour Logoot, LogootSplit et RGASplit entre autres, 99
- 31 : TODO : Me semble que Kleppmann a aussi utilisé et mis à disposition ses traces correspondant à la rédaction d'un de ses articles. Mais que cet article n'était rédigé que par lui. Peu de chances de présence d'éditions concurrentes. À retrouver et vérifier. , 99
- 32 : NOTE : Peut connecter ça à la nécessité de conserver un chemin d'une époque à l'autre : si les opérations émises depuis cette époque ont probablement plus d'intérêt pour l'état actuel, couper l'arbre ?, 100
- 3 : TODO : Spécification faible et forte des séquences répliquées, 4
- 4 : TODO : insérer ref sticky model, 5
- 5 : TODO : Ajouter ref, 6
- 6 : TODO : Faire le lien avec les travaux de Burckhardt [13] et les MRDTs [37], 7
- 7 : TODO : Ajouter refs des horloges logiques plus intelligentes (Interval Tree Clock, Hybrid Clock...), 8
- 8 : TODO : Ajouter ref, 13
- 9 : TODO : Insérer ref, 14

Bibliographie

- [1] D. ABADI. « Consistency Tradeoffs in Modern Distributed Database System Design : CAP is Only Part of the Story ». In : *Computer* 45.2 (2012), p. 37–42. DOI : 10.1109/MC.2012.33.
- [2] Mehdi AHMED-NACER et al. « Evaluating CRDTs for Real-time Document Editing ». In : *11th ACM Symposium on Document Engineering*. Sous la dir. d'ACM. Mountain View, California, United States, sept. 2011, p. 103–112. DOI : 10.1145/2034691.2034717. URL : <https://hal.inria.fr/inria-00629503>.
- [3] Paulo Sérgio ALMEIDA, Ali SHOKER et Carlos BAQUERO. « Delta state replicated data types ». In : *Journal of Parallel and Distributed Computing* 111 (jan. 2018), p. 162–173. ISSN : 0743-7315. DOI : 10.1016/j.jpdc.2017.08.003. URL : <http://dx.doi.org/10.1016/j.jpdc.2017.08.003>.
- [4] Paulo Sérgio ALMEIDA, Ali SHOKER et Carlos BAQUERO. « Efficient State-Based CRDTs by Delta-Mutation ». In : *Networked Systems*. Sous la dir. d'Ahmed BOUAJJANI et Hugues FAUCONNIER. Cham : Springer International Publishing, 2015, p. 62–76. ISBN : 978-3-319-26850-7.
- [5] Paulo Sérgio ALMEIDA et al. « Scalable and Accurate Causality Tracking for Eventually Consistent Stores ». In : *Distributed Applications and Interoperable Systems*. Sous la dir. de Kostas MAGOUTIS et Peter PIETZUCH. Berlin, Heidelberg : Springer Berlin Heidelberg, 2014, p. 67–81. ISBN : 978-3-662-43352-2.
- [6] Luc ANDRÉ et al. « Supporting Adaptable Granularity of Changes for Massive-Scale Collaborative Editing ». In : *International Conference on Collaborative Computing : Networking, Applications and Worksharing - CollaborateCom 2013*. Austin, TX, USA : IEEE Computer Society, oct. 2013, p. 50–59. DOI : 10.4108/icst.collaboratecom.2013.254123.
- [7] AUTOMERGE. *Automerge : data structures for building collaborative applications in Javascript*. URL : <https://github.com/automerge/automerge>.
- [8] Carlos BAQUERO, Paulo Sergio ALMEIDA et Ali SHOKER. *Pure Operation-Based Replicated Data Types*. 2017. arXiv : 1710.04469 [cs.DC].
- [9] Carlos BAQUERO, Paulo Sérgio ALMEIDA et Ali SHOKER. « Making Operation-Based CRDTs Operation-Based ». In : *Proceedings of the First Workshop on Principles and Practice of Eventual Consistency*. PaPEC '14. Amsterdam, The Netherlands : Association for Computing Machinery, 2014. ISBN : 9781450327169. DOI : 10.1145/2596631.2596632. URL : <https://doi.org/10.1145/2596631.2596632>.

- [10] Carlos BAQUERO, Paulo Sérgio ALMEIDA et Ali SHOKER. « Making Operation-Based CRDTs Operation-Based ». In : *Distributed Applications and Interoperable Systems*. Sous la dir. de Kostas MAGOUTIS et Peter PIETZUCH. Berlin, Heidelberg : Springer Berlin Heidelberg, 2014, p. 126–140.
- [11] Jim BAUWENS et Elisa Gonzalez BOIX. « Improving the Reactivity of Pure Operation-Based CRDTs ». In : *Proceedings of the 8th Workshop on Principles and Practice of Consistency for Distributed Data*. PaPoC '21. Online, United Kingdom : Association for Computing Machinery, 2021. ISBN : 9781450383387. DOI : 10.1145/3447865.3457968. URL : <https://doi.org/10.1145/3447865.3457968>.
- [12] Loïck BRIOT, Pascal URSO et Marc SHAPIRO. « High Responsiveness for Group Editing CRDTs ». In : *ACM International Conference on Supporting Group Work*. Sanibel Island, FL, United States, nov. 2016. DOI : 10.1145/2957276.2957300. URL : <https://hal.inria.fr/hal-01343941>.
- [13] Sebastian BURCKHARDT et al. « Replicated Data Types : Specification, Verification, Optimality ». In : *Proceedings of the 41st ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*. POPL '14. San Diego, California, USA : Association for Computing Machinery, 2014, p. 271–284. ISBN : 9781450325448. DOI : 10.1145/2535838.2535848. URL : <https://doi.org/10.1145/2535838.2535848>.
- [14] Mike BURMESTER et Yvo DESMEDT. « A secure and efficient conference key distribution system ». In : *Advances in Cryptology — EUROCRYPT'94*. Sous la dir. d'Alfredo DE SANTIS. Berlin, Heidelberg : Springer Berlin Heidelberg, 1995, p. 275–286. ISBN : 978-3-540-44717-7.
- [15] CONCORDANT. *Concordant*. URL : <http://www.concordant.io/>.
- [16] The SyncFree CONSORTIUM. *AntidoteDB : A planet scale, highly available, transactional database*. URL : <http://antidoteDB.eu/>.
- [17] Armon DADGAR, James PHILLIPS et Jon CURREY. « Lifeguard : Local health awareness for more accurate failure detection ». In : *2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)*. IEEE. 2018, p. 22–25.
- [18] Quang-Vinh DANG et Claudia-Lavinia IGNAT. « Performance of real-time collaborative editors at large scale : User perspective ». In : *Internet of People Workshop, 2016 IFIP Networking Conference*. Proceedings of 2016 IFIP Networking Conference, Networking 2016 and Workshops. Vienna, Austria, mai 2016, p. 548–553. DOI : 10.1109/IFIPNetworking.2016.7497258. URL : <https://hal.inria.fr/hal-01351229>.
- [19] A. DAS, I. GUPTA et A. MOTIVALA. « SWIM : scalable weakly-consistent infection-style process group membership protocol ». In : *Proceedings International Conference on Dependable Systems and Networks*. 2002, p. 303–312. DOI : 10.1109/DSN.2002.1028914.

-
- [20] Kevin DE PORRE et al. « CScript : A distributed programming language for building mixed-consistency applications ». In : *Journal of Parallel and Distributed Computing volume 144* (oct. 2020), p. 109–123. ISSN : 0743-7315. DOI : 10.1016/j.jpdc.2020.05.010.
 - [21] Giuseppe DECANDIA et al. « Dynamo : Amazon’s highly available key-value store ». In : *ACM SIGOPS operating systems review* 41.6 (2007), p. 205–220.
 - [22] Victorien ELVINGER. « Réplication sécurisée dans les infrastructures pair-à-pair de collaboration ». Theses. Université de Lorraine, juin 2021. URL : <https://hal.univ-lorraine.fr/tel-03284806>.
 - [23] Victorien ELVINGER, G  rald OSTER et Fran  ois CHAROY. « Prunable Authenticated Log and Authenticable Snapshot in Distributed Collaborative Systems ». In : *2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC)*. IEEE. 2018, p. 156–165.
 - [24] V. ENES et al. « Efficient Synchronization of State-Based CRDTs ». In : *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. 2019, p. 148–159. DOI : 10.1109/ICDE.2019.00022.
 - [25] ETHERPAD. *Etherpad*. URL : <https://etherpad.org/>.
 - [26] Barton GELLMAN et Laura POITRAS. *U.S., British intelligence mining data from nine U.S. Internet companies in broad secret program*. URL : https://www.washingtonpost.com/investigations/us-intelligence-mining-data-from-nine-us-internet-companies-in-broad-secret-program/2013/06/06/3a0c0da8-cebf-11e2-8845-d970ccb04497_story.html.
 - [27] Jim GILES. « Special Report Internet encyclopaedias go head to head ». In : *nature* 438.15 (2005), p. 900–901.
 - [28] Ricardo Jorge Tom   GON  ALVES et al. « DottedDB : Anti-Entropy without Merkle Trees, Deletes without Tombstones ». In : *2017 IEEE 36th Symposium on Reliable Distributed Systems (SRDS)*. 2017, p. 194–203. DOI : 10.1109/SRDS.2017.28.
 - [29] GOOGLE. *Google Docs*. URL : <https://docs.google.com/>.
 - [30] Glen GREENWALD et Ewen MACASKILL. *NSA Prism program taps in to user data of Apple, Google and others*. URL : <https://www.theguardian.com/world/2013/jun/06/us-tech-giants-nsa-data>.
 - [31] Victor GRISHCHENKO et Mikhail PATRAKEEV. « Chronofold : A Data Structure for Versioned Text ». In : *Proceedings of the 7th Workshop on Principles and Practice of Consistency for Distributed Data*. PaPoC ’20. Heraklion, Greece : Association for Computing Machinery, 2020. ISBN : 9781450375245. DOI : 10.1145/3380787.3393680. URL : <https://doi.org/10.1145/3380787.3393680>.
 - [32] Peter van HARDENBERG et Martin KLEPPMANN. « PushPin : Towards Production-Quality Peer-to-Peer Collaboration ». In : *7th Workshop on Principles and Practice of Consistency for Distributed Data*. PaPoC 2020. ACM, avr. 2020. DOI : 10.1145/3380787.3393683.

- [33] Claudia-Lavinia IGNAT. « Maintaining consistency in collaboration over hierarchical documents ». Thèse de doct. ETH Zurich, 2006.
- [34] Claudia-Lavinia IGNAT et al. « How Do User Groups Cope with Delay in Real-Time Collaborative Note Taking ». In : *European Conference on Computer Supported Cooperative Work 2015*. Sous la dir. de Nina BOULUS-RODJE et al. Proceedings of the 14th European Conference on Computer Supported Cooperative Work. Oslo, Norway : Springer International Publishing, sept. 2015, p. 223–242. DOI : 10.1007/978-3-319-20499-4_12. URL : <https://hal.inria.fr/hal-01238831>.
- [35] Claudia-Lavinia IGNAT et al. « Studying the Effect of Delay on Group Performance in Collaborative Editing ». In : *Proceedings of 11th International Conference on Cooperative Design, Visualization, and Engineering, CDVE 2014, Springer 2014 Lecture Notes in Computer Science*. Proceedings of 11th International Conference on Cooperative Design, Visualization, and Engineering, CDVE 2014. Seattle, WA, United States, sept. 2014, p. 191–198. DOI : 10.1007/978-3-319-10831-5_29. URL : <https://hal.archives-ouvertes.fr/hal-01088815>.
- [36] Paul R JOHNSON et Robert THOMAS. *RFC0677 : Maintenance of duplicate databases*. RFC Editor, 1975.
- [37] Gowtham KAKI et al. « Mergeable Replicated Data Types ». In : *Proc. ACM Program. Lang.* 3.OOPSLA (oct. 2019). DOI : 10.1145/3360580. URL : <https://doi.org/10.1145/3360580>.
- [38] Martin KLEPPMANN et Alastair R. BERESFORD. « A Conflict-Free Replicated JSON Datatype ». In : *IEEE Transactions on Parallel and Distributed Systems* 28.10 (oct. 2017), p. 2733–2746. ISSN : 1045-9219. DOI : 10.1109/tpds.2017.2697382. URL : <http://dx.doi.org/10.1109/TPDS.2017.2697382>.
- [39] Martin KLEPPMANN et al. « Local-First Software : You Own Your Data, in Spite of the Cloud ». In : *Proceedings of the 2019 ACM SIGPLAN International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software*. Onward! 2019. Athens, Greece : Association for Computing Machinery, 2019, p. 154–178. ISBN : 9781450369954. DOI : 10.1145/3359591.3359737. URL : <https://doi.org/10.1145/3359591.3359737>.
- [40] Nico KRUBER, Maik LANGE et Florian SCHINTKE. « Approximate Hash-Based Set Reconciliation for Distributed Replica Repair ». In : *2015 IEEE 34th Symposium on Reliable Distributed Systems (SRDS)*. 2015, p. 166–175. DOI : 10.1109/SRDS.2015.30.
- [41] Leslie LAMPORT. « Time, Clocks, and the Ordering of Events in a Distributed System ». In : *Commun. ACM* 21.7 (juil. 1978), p. 558–565. ISSN : 0001-0782. DOI : 10.1145/359545.359563. URL : <https://doi.org/10.1145/359545.359563>.
- [42] Mihai LETIA, Nuno PREGUIÇA et Marc SHAPIRO. « Consistency without concurrency control in large, dynamic systems ». In : *LADIS 2009 - 3rd ACM SIGOPS International Workshop on Large Scale Distributed Systems and Middleware*. T. 44. Operating Systems Review 2. Big Sky, MT, United States : Assoc. for Computing

- Machinery, oct. 2009, p. 29–34. DOI : 10.1145/1773912.1773921. URL : <https://hal.inria.fr/hal-01248270>.
- [43] Christopher MEIKLEJOHN et Peter VAN ROY. « Lasp : A Language for Distributed, Coordination-free Programming ». In : *17th International Symposium on Principles and Practice of Declarative Programming*. PPDP 2015. ACM, juil. 2015, p. 184–195. DOI : 10.1145/2790449.2790525.
 - [44] Madhavan MUKUND, Gautham SHENOY et SP SURESH. « Optimized or-sets without ordering constraints ». In : *International Conference on Distributed Computing and Networking*. Springer. 2014, p. 227–241.
 - [45] Brice NÉDELEC, Pascal MOLLI et Achour MOSTEFAOUI. « CRATE : Writing Stories Together with our Browsers ». In : *25th International World Wide Web Conference*. WWW 2016. ACM, avr. 2016, p. 231–234. DOI : 10.1145/2872518.2890539.
 - [46] Brice NÉDELEC, Pascal MOLLI et Achour MOSTÉFAOUI. « A scalable sequence encoding for collaborative editing ». In : *Concurrency and Computation : Practice and Experience* (), e4108. DOI : 10.1002/cpe.4108. eprint : <https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpe.4108>. URL : <https://onlinelibrary.wiley.com/doi/abs/10.1002/cpe.4108>.
 - [47] Brice NÉDELEC et al. « An adaptive peer-sampling protocol for building networks of browsers ». In : *World Wide Web* 21.3 (2018), p. 629–661.
 - [48] Brice NÉDELEC et al. « LSEQ : an adaptive structure for sequences in distributed collaborative editing ». In : *Proceedings of the 2013 ACM Symposium on Document Engineering*. DocEng 2013. Sept. 2013, p. 37–46. DOI : 10.1145/2494266.2494278.
 - [49] Hoang-Long NGUYEN, Claudia-Lavinia IGNAT et Olivier PERRIN. « Trusternity : Auditing Transparent Log Server with Blockchain ». In : *Companion of the The Web Conference 2018*. Lyon, France, avr. 2018. DOI : 10.1145/3184558.3186938. URL : <https://hal.inria.fr/hal-01883589>.
 - [50] Hoang-Long NGUYEN et al. « Blockchain-Based Auditing of Transparent Log Servers ». In : *32th IFIP Annual Conference on Data and Applications Security and Privacy (DBSec)*. Sous la dir. de Florian KERSCHBAUM et Stefano PARABOSCHI. T. LNCS-10980. Data and Applications Security and Privacy XXXII. Part 1 : Administration. Bergamo, Italy : Springer International Publishing, juil. 2018, p. 21–37. DOI : 10.1007/978-3-319-95729-6_2. URL : <https://hal.archives-ouvertes.fr/hal-01917636>.
 - [51] Petru NICOLAESCU et al. « Near Real-Time Peer-to-Peer Shared Editing on Extensible Data Types ». In : *19th International Conference on Supporting Group Work*. GROUP 2016. ACM, nov. 2016, p. 39–49. DOI : 10.1145/2957276.2957310.
 - [52] Petru NICOLAESCU et al. « Yjs : A Framework for Near Real-Time P2P Shared Editing on Arbitrary Data Types ». In : *15th International Conference on Web Engineering*. ICWE 2015. Springer LNCS volume 9114, juin 2015, p. 675–678. DOI : 10.1007/978-3-319-19890-3_55. URL : <http://dbis.rwth-aachen.de/~derntl/papers/preprints/icwe2015-preprint.pdf>.

- [53] Matthieu NICOLAS. « Efficient renaming in CRDTs ». In : *Middleware 2018 - 19th ACM/IFIP International Middleware Conference (Doctoral Symposium)*. Rennes, France, déc. 2018. URL : <https://hal.inria.fr/hal-01932552>.
- [54] Matthieu NICOLAS, G  rald OSTER et Olivier PERRIN. « Efficient Renaming in Sequence CRDTs ». In : *7th Workshop on Principles and Practice of Consistency for Distributed Data (PaPoC'20)*. Heraklion, Greece, avr. 2020. URL : <https://hal.inria.fr/hal-02526724>.
- [55] Matthieu NICOLAS et al. « MUTE : A Peer-to-Peer Web-based Real-time Collaborative Editor ». In : *ECSCW 2017 - 15th European Conference on Computer-Supported Cooperative Work*. T. 1. Proceedings of 15th European Conference on Computer-Supported Cooperative Work - Panels, Posters and Demos 3. Sheffield, United Kingdom : EUSSET, ao  t 2017, p. 1–4. DOI : 10.18420/ecscw2017_p5. URL : <https://hal.inria.fr/hal-01655438>.
- [56] Sylvie NO  L et Jean-Marc ROBERT. « Empirical study on collaborative writing : What do co-authors do, use, and like ? ». In : *Computer Supported Cooperative Work (CSCW) 13.1* (2004), p. 63–89.
- [57] G  rald OSTER et al. « Data Consistency for P2P Collaborative Editing ». In : *ACM Conference on Computer-Supported Cooperative Work - CSCW 2006*. Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work. Banff, Alberta, Canada : ACM Press, nov. 2006, p. 259–268. URL : <https://hal.inria.fr/inria-00108523>.
- [58] D. S. PARKER et al. « Detection of Mutual Inconsistency in Distributed Systems ». In : *IEEE Trans. Softw. Eng.* 9.3 (mai 1983), p. 240–247. ISSN : 0098-5589. DOI : 10.1109/TSE.1983.236733. URL : <https://doi.org/10.1109/TSE.1983.236733>.
- [59] Jim PICK. *Graf, Nikolaus*. URL : <https://www.serenity.re/en/notes>.
- [60] Jim PICK. *PeerPad*. URL : <https://peerpad.net/>.
- [61] Ravi PRAKASH, Michel RAYNAL et Mukesh SINGHAL. « An Adaptive Causal Ordering Algorithm Suited to Mobile Computing Environments ». In : *Journal of Parallel and Distributed Computing* 41.2 (1997), p. 190–204. ISSN : 0743-7315. DOI : <https://doi.org/10.1006/jpdc.1996.1300>. URL : <https://www.sciencedirect.com/science/article/pii/S0743731596913003>.
- [62] Nuno PREGUICA et al. « A Commutative Replicated Data Type for Cooperative Editing ». In : *2009 29th IEEE International Conference on Distributed Computing Systems*. Juin 2009, p. 395–403. DOI : 10.1109/ICDCS.2009.20.
- [63] Pierre-Antoine RAULT, Claudia-Lavinia IGNAT et Olivier PERRIN. « Distributed Access Control for Collaborative Applications Using CRDTs ». In : *Proceedings of the 9th Workshop on Principles and Practice of Consistency for Distributed Data*. PaPoC '22. Rennes, France : Association for Computing Machinery, 2022, p. 33–38. ISBN : 9781450392563. DOI : 10.1145/3517209.3524826. URL : <https://doi.org/10.1145/3517209.3524826>.
- [64] RIAK. *Riak KV*. URL : <http://riak.com/>.

-
- [65] Hyun-Gul ROH et al. « Replicated abstract data types : Building blocks for collaborative applications ». In : *Journal of Parallel and Distributed Computing* 71.3 (2011), p. 354–368. ISSN : 0743-7315. DOI : <https://doi.org/10.1016/j.jpdc.2010.12.006>. URL : <http://www.sciencedirect.com/science/article/pii/S0743731510002716>.
 - [66] Yasushi SAITO et Marc SHAPIRO. « Optimistic Replication ». In : *ACM Comput. Surv.* 37.1 (mar. 2005), p. 42–81. ISSN : 0360-0300. DOI : 10.1145/1057977.1057980. URL : <https://doi.org/10.1145/1057977.1057980>.
 - [67] Marc SHAPIRO et al. *A comprehensive study of Convergent and Commutative Replicated Data Types*. Research Report RR-7506. Inria – Centre Paris-Rocquencourt ; INRIA, jan. 2011, p. 50. URL : <https://hal.inria.fr/inria-00555588>.
 - [68] Marc SHAPIRO et al. « Conflict-Free Replicated Data Types ». In : *Proceedings of the 13th International Symposium on Stabilization, Safety, and Security of Distributed Systems*. SSS 2011. 2011, p. 386–400. DOI : 10.1007/978-3-642-24550-3_29.
 - [69] Haifeng SHEN et Chengzheng SUN. « A log compression algorithm for operation-based version control systems ». In : *Proceedings 26th Annual International Computer Software and Applications*. 2002, p. 867–872. DOI : 10.1109/CMPSAC.2002.1045115.
 - [70] Chengzheng SUN et al. « Achieving Convergence, Causality Preservation, and Intention Preservation in Real-Time Cooperative Editing Systems ». In : *ACM Trans. Comput.-Hum. Interact.* 5.1 (mar. 1998), p. 63–108. ISSN : 1073-0516. DOI : 10.1145/274444.274447. URL : <https://doi.org/10.1145/274444.274447>.
 - [71] Douglas B TERRY et al. « Managing Update Conflicts in Bayou, a Weakly Connected Replicated Storage System ». In : *SIGOPS Oper. Syst. Rev.* 29.5 (déc. 1995), p. 172–182. ISSN : 0163-5980. DOI : 10.1145/224057.224070. URL : <https://doi.org/10.1145/224057.224070>.
 - [72] Stéphane WEISS, Pascal URSO et Pascal MOLLI. « Logoot : A Scalable Optimistic Replication Algorithm for Collaborative Editing on P2P Networks ». In : *Proceedings of the 29th International Conference on Distributed Computing Systems - ICDCS 2009*. Montreal, QC, Canada : IEEE Computer Society, juin 2009, p. 404–412. DOI : 10.1109/ICDCS.2009.75. URL : <http://doi.ieeecomputersociety.org/10.1109/ICDCS.2009.75>.
 - [73] Stéphane WEISS, Pascal URSO et Pascal MOLLI. « Logoot-Undo : Distributed Collaborative Editing System on P2P Networks ». In : *IEEE Transactions on Parallel and Distributed Systems* 21.8 (août 2010), p. 1162–1174. DOI : 10.1109/TPDS.2009.173. URL : <https://hal.archives-ouvertes.fr/hal-00450416>.
 - [74] Stéphane WEISS, Pascal URSO et Pascal MOLLI. « Wooki : a P2P Wiki-based Collaborative Writing Tool ». In : t. 4831. Déc. 2007. ISBN : 978-3-540-76992-7. DOI : 10.1007/978-3-540-76993-4_42.
 - [75] C. WU et al. « Anna : A KVS for Any Scale ». In : *IEEE Transactions on Knowledge and Data Engineering* 33.2 (2021), p. 344–358. DOI : 10.1109/TKDE.2019.2898401.

- [76] Elena YANAKIEVA et al. « Access Control Conflict Resolution in Distributed File Systems Using CRDTs ». In : *Proceedings of the 8th Workshop on Principles and Practice of Consistency for Distributed Data*. PaPoC '21. Online, United Kingdom : Association for Computing Machinery, 2021. ISBN : 9781450383387. DOI : 10.1145/3447865.3457970. URL : <https://doi.org/10.1145/3447865.3457970>.
- [77] YJS. *Yjs : A CRDT framework with a powerful abstraction of shared data*. URL : <https://github.com/yjs/yjs>.
- [78] Weihai YU. « A String-Wise CRDT for Group Editing ». In : *Proceedings of the 17th ACM International Conference on Supporting Group Work*. GROUP '12. Sanibel Island, Florida, USA : Association for Computing Machinery, 2012, p. 141–144. ISBN : 9781450314862. DOI : 10.1145/2389176.2389198. URL : <https://doi.org/10.1145/2389176.2389198>.
- [79] Weihai YU et Claudia-Lavinia IGNAT. « Conflict-Free Replicated Relations for Multi-Synchronous Database Management at Edge ». In : *IEEE International Conference on Smart Data Services, 2020 IEEE World Congress on Services*. Beijing, China, oct. 2020. URL : <https://hal.inria.fr/hal-02983557>.
- [80] Weihai YU et Sigbjørn ROSTAD. « A Low-Cost Set CRDT Based on Causal Lengths ». In : *Proceedings of the 7th Workshop on Principles and Practice of Consistency for Distributed Data*. New York, NY, USA : Association for Computing Machinery, 2020. ISBN : 9781450375245. URL : <https://doi.org/10.1145/3380787.3393678>.
- [81] Marek ZAWIRSKI, Marc SHAPIRO et Nuno PREGUIÇA. « Asynchronous rebalancing of a replicated tree ». In : *Conférence Française en Systèmes d'Exploitation (CFSE)*. Saint-Malo, France, mai 2011, p. 12. URL : <https://hal.inria.fr/hal-01248197>.

Résumé

Afin d'assurer leur haute disponibilité, les systèmes distribués à large échelle se doivent de répliquer leurs données tout en minimisant les coordinations nécessaires entre noeuds. Pour concevoir de tels systèmes, la littérature et l'industrie adoptent de plus en plus l'utilisation de types de données répliquées sans conflits (CRDTs). Les CRDTs sont des types de données qui offrent des comportements similaires aux types existants, tel l'Ensemble ou la Séquence. Ils se distinguent cependant des types traditionnels par leur spécification, qui supporte nativement les modifications concurrentes. À cette fin, les CRDTs incorporent un mécanisme de résolution de conflits au sein de leur spécification.

Afin de résoudre les conflits de manière déterministe, les CRDTs associent généralement des identifiants aux éléments stockés au sein de la structure de données. Les identifiants doivent respecter un ensemble de contraintes en fonction du CRDT, telles que l'unicité ou l'appartenance à un ordre dense. Ces contraintes empêchent de borner la taille des identifiants. La taille des identifiants utilisés 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. Le but de cette thèse est de proposer des solutions pour pallier ce problème.

Nous présentons dans cette thèse deux contributions visant à répondre à ce problème : (i) Un nouveau CRDT pour Séquence, RenamableLogootSplit, qui intègre un mécanisme de renommage à sa spécification. Ce mécanisme de renommage permet aux noeuds du système de réattribuer des identifiants de taille minimale aux éléments de la séquence. Cependant, cette première version requiert une coordination entre les noeuds pour effectuer un renommage. L'évaluation expérimentale montre que le mécanisme de renommage permet de réinitialiser à chaque renommage le surcoût lié à l'utilisation du CRDT. (ii) Une seconde version de RenamableLogootSplit conçue pour une utilisation dans un système distribué. Cette nouvelle version permet aux noeuds de déclencher un renommage sans coordination préalable. L'évaluation expérimentale montre que cette nouvelle version présente un surcoût temporaire en cas de renommages concurrents, mais que ce surcoût est à terme.

Mots-clés: CRDTs, édition collaborative en temps réel, cohérence à terme, optimisation mémoire, performance

Abstract

Keywords: CRDTs, real-time collaborative editing, eventual consistency, memory-wise optimisation, performance

`main`: version du vendredi 29 juillet 2022 à 13 h 19

