Haut Conseil de l'Évaluation de la Recherche et de l'Enseignement Supérieur

DOCUMENT D'AUTOÉVALUATION

Équipe DELYS

Campagne d'évaluation 2023-2024 — Vague D

Table des matières

1 INFORMATIONS GÉNÉRALES SUR L'ÉQUIPE DELYS

Nom de l'équipe : DistributEd aLgorithms and sYstems (DELYS)

Responsable de l'équipe Pierre Sens

	2017	2018	2019	2020	2021	2022
PR	2	2	2	2	3	3
MCF HDR	0	0	0	0	0	0
MCF	5	5	5	5	5	5
DR	1	1	2	0	0	0
CR HDR	1	1	1	1	1	1
CR	0	0	0	0	0	0
Total permanents	9	9	10	8	9	9
Émérites	0	0	0	1	1	1
Doctorants	12	12	10	13	12	10
Ingénieurs CDD ou hors tutelles	1	0	0	0	0	0
Post-doc, ATER, etc.	1	0	4	2	2	0
Stagiaires	5	4	4	2	5	3
Total non permanents	19	16	18	17	19	13
Total avec émérites	**28**	**25**	**28**	**26**	**29**	**23**
Equivalent temps plein recherche	5.5	5.5	6.5	4.5	5.0	5.0

TABLE 1 – Personnels DELYS sur la période 2017-2022 (au 1er juillet de chaque année)

1.1 Les thématiques scientifiques et leurs enjeux

DELYS a été créée en 2018 et fait suite à l'équipe REGAL. De 2018 à 2022, DELYS était une équipe de recherche commune avec le centre Inria Paris (créée officiellement en tant EPC Inria le 1er janvier 2019). Depuis le 1er janvier 2023, elle n'est plus associée à Inria.

DELYS étudie les nouveaux systèmes distribués d'un point de vue pratique et théorique. Sa recherche s'intéresse à un large spectre de systèmes répartis : le cloud et Fog computing, les réseaux mobiles, les systèmes dynamiques tels que les réseaux de robots, les multi-cœurs.

Ces nouveaux systèmes informatiques distribués doivent tolérer les fautes, supporter la dynamique des nœuds et leur hétérogénéité, gérer une virtualisation à plusieurs niveaux. Les algorithmes conçus pour les systèmes distribués statiques doivent être donc être repensés pour tenir compte de la nature hautement instable du système.

DELYS concentre ses recherches autour de trois thématiques scientifiques.

Axe 1 : Algorithmes distribués

Nous avons obtenu des résultats tant sur les aspects fondamentaux que pratiques des algorithmes distribués. Nous avons étudié plusieurs problèmes clés des algorithmes distribués liés à la tolérance aux pannes, principalement à travers deux approches orthogonales : la détection des fautes et l'auto-stabilisation.

Détection de fautes. Les *détecteurs de fautes* (FD – *failure detector*) sont une abstraction fondamentale pour les algorithmes distribués. Les FD sont des oracles distribués qui fournissent des informations non fiables sur les défaillances des processus, souvent sous la forme d'une liste d'identités de processus corrects. Ils ont été largement utilisés pour résoudre les problèmes d'accord dans les systèmes asynchrones sujets à des fautes franches. Cependant, ils ont surtout été appliqués à des topologies de réseaux statiques où l'ensemble des nœud est connu à l'avance.

Nous avons donc étendu les détecteur aux systèmes dynamiques. Dans [Mauffret et al., 2019], prix du meilleur papier à ICDCN 2019, nous avons proposé le premier détecteur de défaillance pour résoudre le problème de l'exclusion mutuelle adapté aux systèmes dynamiques et prouvé qu'il est le détecteur de fautes le plus faible.

Nous avons également étudié l'élection d'un leader ultime, appelé Ω, qui est connu pour être le détecteur de fautes le plus faible pour résoudre le consensus avec une majorité de processus corrects. Ω est utilisé par des protocoles comme Paxos pour implémenter des machines à état répliqué. Dans [Dubois et al., 2019] nous avons montré que dans *n'importe quel* environnement par passage de messages, c'est-à-dire, sous n'importe quelle hypothèse sur le moment et l'endroit où les défaillances peuvent se produire, Ω est également le détecteur de défaillance le plus faible pour mettre en œuvre un service répliqué cohérent à terme. Ce résultat théorique est le fruit d'une collaboration avec Rachid Guerraoui (Professeur à l'EPFL) et Petr Kuznetsov (Professeur à l'IMT)

Dans [Favier et al., 2020a], nous avons proposé un nouvel algorithme qui élit à terme un leader correct pour chaque composant connecté d'un réseau dynamique où les nœuds peuvent se déplacer ou tomber en panne. Cet algorithme a reçu le prix du meilleur papier étudiant à la conférence NCA en 2021.

Auto-stabilisation. *L'auto-stabilisation* est une technique qui permet de résister aux fautes transitoires dans un système distribué [Altisen et al., 2019a]. Un algorithme auto-stabilisant est capable de retrouver un comportement correct en temps fini, quelle que soit la configuration initiale du système. L'auto-stabilisation permet également de tolérer les changements topologiques du réseau d'interconnexion. Dans ce cas, les changements topologiques sont considérés comme des défaillances transitoires des liens de communications.

Nous avons étudié les conditions dans lesquelles la stabilisation de l'élection d'un leader peut être résolue dans les systèmes de transmission de messages hautement dynamiques. Dans [Altisen et al., 2021b], nous fournissons des conditions nécessaires et suffisantes sous lesquelles ce problème peut être résolu en supposant que chaque processus peut atteindre tous les autres au moins une fois par le biais d'un *voyage* (un voyage peut être considéré comme un chemin dans le temps d'une source à une destination). Dans [Altisen et al., 2021a], nous affaiblissons le modèle en considérant des systèmes dynamiques où certains processus peuvent ne pas être des sources ou des destinataires.

Dans un contexte auto-stabilisant, nous avons également étudié le rassemblement (ou rendez-vous) et l'exploration qui sont deux problèmes fondamentales dans le domaine des systèmes distribués mobiles. Le rassemblement consiste à amener des agents qui partent initialement de positions différentes à se rencontrer tous ensemble en un temps fini. L'environnement peut être soit fini (modélisé par un graphe où les nœuds sont des emplacements), soit continu (modélisé par le plan). L'exploration consiste à amener les agents à visiter tout un environnement discret : chaque nœud du graphe doit être visité par au moins un agent. Depuis 2017, nous avons obtenu de nombreux résultats traitant de ces deux problèmes. Le problème de rassemblement a été traité dans l'environnement discret [Bournat et al., 2018b] et dans le plan [Bouchard et al., 2019a, Bouchard et al., 2020d]. Différents types d'exploration ont également été étudiés dans les graphes [Bournat et al., 2019, Devismes et al., 2019, Devismes et al., 2021] et le plan [Bouchard et al., 2020e].

Positionnement. Sur la période d'évaluation, nous avons collaboré avec 6 chercheurs de 5 pays.

Plusieurs équipes en France abordent des sujets proches de ceux de DELYS. WIDE (Irisa) explore les algorithmes distribués dans des grands réseaux dynamiques. WIDE se concentrent sur les résultats théoriques et les évaluations de performances sont principalement basées sur des simulations. DELYS a une approche plus système, mettant en œuvre et évaluant des algorithmes distribués dans des environnements simulés et réels. Les équipes COATI, DANTE, et FUN (Inria) s'intéressent également aux systèmes distribués dynamiques. Ils abordent des aspects différents en se concentrant souvent sur des fonctionnalités proches du réseau. Enfin, l'équipe de Petr Kuznetsov à Télécom ParisTech étudie la théorie des algorithmes distribués en se focalisant sur la détection des fautes. Cette équipe s'intéresse aux propriétés mathématiques des systèmes distribués.

A l'international, l'équipe LPD de l'EPFL, dirigée par Rachid Guerraoui, se concentre sur la fiabilité des systèmes distribués. Par rapport à cette équipe, nous nous concentrons davantage sur les problèmes de passage à l'échelle et son impact sur la dynamique. D'autres groupes, par exemple le groupe de Nancy Lynch "Theory of Distributed Systems" (TDS) au MIT, où l'équipe AC de Fabian Kuhn à l'Université de Freiburg s'intéressent également aux algorithmes pour les systèmes distribués dynamiques. TDS se concentre sur la tolérance aux fautes, tandis qu'AC étudie principalement sur les fondements mathématiques. Enfin, l'équipe de Alberto Lafuente et Mikel Larrea à San Sebastien (Espagne) étudie également la détection de fautes dans des environnements dynamiques en utilisant des hypothèses de synchronisation partielle. Cependant, aucun de ces groupes considère une approche auto-stabilisante ou des communications asynchrones.

Axe 2 : Systèmes répartis

Concernant les systèmes répartis, nous avons obtenu des résultats sur la cohérence des données répliquées, l'ordonnancement et le placement de tâches dans les grandes infrastructures réparties.

Gestion de données géo-localisées. La gestion et l'accès cohérent aux données réparties sur des datacenters est un problème majeur sur lequel nous avons fait les contributions suivantes.

Nous nous sommes intéressés au traitement efficace des requêtes lorsque les utilisateurs et les données sont distribués sur de multiples emplacements géographiques. En particulier, nous avons étudié comment le placement d'index, les modèles de requêtes et le stockage influent sur les performances. Nous avons ainsi proposé une nouvelle architecture de moteur de requêtes, appelé Proteus, qui permet à l'administrateur de prendre les décisions de placement appropriées. Ce travail a été réalisé conjointement avec B. King, par le biais d'une subvention Cifre

avec Scality.

La distribution et la réplication des données en bordure du réseau permettent une interrogation rapide, autonome et assure une meilleure disponibilité pour les applications, telles que les jeux, l'ingénierie coopérative ou le partage d'informations. Cependant, les utilisateurs exigent les meilleures garanties de cohérence possibles et un support spécifique pour la collaboration de groupe. Pour relever ce défi, nous avons conçu le système Colony qui garantit la Cohérence Transactionnelle Causale Plus (TCC+) [Toumlilt et al., 2021]. Ce travail, a été réalisé conjointement avec Pierre Sutra de Telécom SudParis.Les concepts explorés dans ce travail ont été en partie industrialisés, grâce au soutien de Inria Startup Studio, au sein de la start-up Concordant.

Enfin, dans le cadre d'accès concurrents à des objets répliqués, nous avons proposé une nouvelle méthodologie de preuve pour établir qu'un objet répliqué donné maintient un invariant. Notre approche permet de raisonner séparément sur les opérations individuelles. Nous avons développé un outil, Soteria, qui automatise les preuves en utilisant le solver SMT Boogie. Ce travail a été réalisé en collaboration avec Gustavo Petri d'ARM Research Cambridge (Royaume-Uni) et a été publié à ESOP 2019 [Nair et al., 2020a].

Gestion de ressources à large échelle. Depuis 2018, nous collaborons avec des chercheurs d'Orange Labs sur la *gestion des ressources dans les grands réseaux*. Nous avons étudié le problème de placement des fonctions de réseau dans le cadre du multi-slicing sous la forme d'un problème d'optimisation. Dans [Alves Esteves et al., 2020c], nous nous sommes concentrés sur le problème du placement de chaîne de fonctions de réseau virtuel (VNF). Nous avons pris en compte la localisation géographique des utilisateurs et proposé un modèle de données adapté à la programmation linéaire en nombres entiers. Nous avons également proposé et évalué dans [Alves Esteves et al., 2021a] une nouvelle approche hybride utilisant un apprentissage par renforcement profond assisté d'une heuristique pour résoudre le problème d'optimisation. Ce travail a fait l'objet de la thèse de doctorat de José Alves réalisée dans le cadre d'une Convention Cifre avec Orange Labs.

Nous avons également étudié l'allocation des ressources dans les environnements en nuage en ciblant *l'ordonnancement des tâches* d'applications scientifiques sur différents types de machines virtuelles (MV). Dans [Teylo et al., 2020], nous avons proposé un ordonnanceur dynamique qui prend en compte de l'hibernation potentielle de MV dans le cadre la plateforme Amazon EC2.

D'un point de vue plus théorique, nous avons étudié, en collaboration avec l'équipe RO du LIP6, l'ordonnancement des tâches en présence de fautes en introduisant un nouveau modèle non-probabiliste. Nous avons étudié deux problèmes : (i) le problème d'ordonnancement des requêtes avec erreurs, dont le but est de trouver suffisamment de créneaux sans erreurs avec le nombre minimum de requêtes, et (ii) le problème d'ordonnancement où nous cherchons les premiers créneaux sans erreurs avec le nombre minimum de requêtes. Nous considérons à la fois la version hors ligne et la version en ligne des problèmes. Ce travail, fait dans le cadre d'un projet LIP6 avec l'équipe RO, a été publié à IJCAI en 2018 [Arantes et al., 2018]. En collaboration à l'équipe Rap d'Inria Paris, nous avons également analysé l'efficacité des mécanismes de réplication dans les grands système de stockage. Nous avons étudié l'impact des stratégies de placement des copies d'un fichier en utilisant des méthodes probabilistes et l'analyse du champ moyen. Ce travail a fait l'objet d'une publication à la conférence SIGMETRICS [Sun et al., 2017b].

Positionnement. Sur la période d'évaluation, nous avons collaboré avec 16 chercheurs de 6 pays.

En France, les équipes WIDE, Myriads (Irisa), STACK (LN2S) ou ERODS (LIG) abordent des thématiques similaires à DELYS. Notre originalité est d'aborder un spectre allant des modèles de cohérence au placement sous un angle à la fois théorique et pratique avec des déploiements sur des plateformes réelles.

Des équipes internationales partagent des intérêts similaires à ceux de DELYS. On peut citer notamment l'équipe de Joe Hellerstein à UC Berkeley, et celles de Willy Zwaenepoel et Alan Fekete à U. Sydney.

Axe 3 : Système

Sur l'axe système, nous avons obtenu des avancées sur la gestion de la mémoire dans les nuages ainsi sur l'optimisation des systèmes pour des applications spécifiques.

Gestion de la mémoire.

Les conteneurs sont une technique de virtualisation légère clé dans les architectures en nuage. Nous avons proposé une nouvelle méthode, MemOpLight, qui permet d'équilibrer la charge mémoire entre des conteneurs s'exécutant sur une même machine physique. MemOpLight réaffecte dynamiquement la mémoire aux conteneurs en se basant sur un retour applicatif. MemOpLigth a été implémenté dans noyau Linux et a fait l'objet de la thèse de Francis Laniel qui a obtenu le prix du meilleur papier à la conférence NCA en 2020 [Laniel et al., 2020].

Depuis 2018, nous étudions également le problème de la contention mémoire dans les systèmes multi-cœurs

temps réel lorsque plusieurs applications s'exécutent simultanément sur une même machine. Nous avons proposé de nouvelles métriques pour quantifier les aspects qualitatifs de la consommation mémoire et défini un nouvel outil de profilage qui a fait l'objet d'une publication à RTSS en 2019 [Courtaud et al., 2019].

Cache dans le noyau linux.

Sur les aspects plus distribués, nous avons étudié les systèmes de stockage en mémoire de type clé-valeur afin d'améliorer leurs performances. En fournissant un accès rapide aux données populaires, ces systèmes sont très utilisés par les services réseau. Memcached, l'un des systèmes de stockage les plus populaires, a des performances qui restent limitées et qui sont inhérentes à la pile réseau du noyau Linux. Nous avons proposé dans [Ghigoff et al., 2021], publié à NSDI'21, BMC un cache intégré au noyau dédié à Memcached qui intercepte les requêtes d'accès aux données avant l'exécution de la pile réseau standard. Nos évaluations montrent que BMC améliore le débit jusqu'à 18 fois par rapport à Memcached classique et jusqu'à 6 fois par rapport à une version optimisée de Memcached.

Multi-cœur pour améliorer la résolution du SAT. Les machines multi-cœurs ouvrent de nouvelles possibilités pour les solveurs SAT. Grâce à notre expertise en matière de systèmes, nous avons développé avec l'équipe MOVE du LIP6 et le LRDE (Epita), *Painless* un nouvel outil permettant de construire des solveurs SAT parallèles efficaces (TACAS 2019 [Le Frioux et al., 2019]). Painless a remporté à trois reprises la compétition internationale annuelle des solveurs SAT.

Positionnement. Sur la période d'évaluation, nous avons collaboré avec 8 chercheurs de 3 pays.

En France quelques équipes abordent des thématiques proches des nôtres sur les aspects liés aux systèmes d'exploitation. Nos recherches sur l'ordonnancement et le temps réel ont été menées en collaboration avec l'équipe Whisper d'Inria Paris. Les travaux d'Alain Tchana au LIG sont proches des nôtres et visent à améliorer la gestion des MV. Les recherches de Vivien Quéma dans le groupe ERODS du LIG considèrent également l'impact du NUMA dans les architectures multi-cœurs. Ces travaux sont complémentaires aux nôtres. Enfin, depuis 2018, nous collaborons avec l'équipe de Willy Zwaenepoel de l'Université de Sydney.

2 INTRODUCTION DU PORTFOLIO

Le portfolio de DELYS est composé des documents suivants :

1. *Élément 1 (autre) :* il s'agit du retour d'évaluation Inria fait par les 3 experts internationaux (Pascal Felber, Université de Neuchatel - Suisse, Paola Flocchini, Université d'Ottawa - Canada, Bertrand Lecun, Google) lors de l'évaluation du thème "Distributed Systèmes and Middleware" d'Inria en 2022.
2. *Élément 2 (publication) :* On Implementing Stabilizing Leader Election with Weak Assumptions on Network Dynamics [Altisen et al., 2021a] publié à la conférence PODC 2021 sur l'implémentation d'une élection de leader dans un système dynamique. Cette publication est le fruit d'une collaboration avec le Labri et Verimag.
3. *Élément 3 (publication) :* Highly-available and consistent group collaboration at the edge with colony [Toumlilt et al., 2021] publié à la conférence Middleware 2021 sur système Colony qui implémente la nouvelle Cohérence Transactionnelle Causale Plus.
4. *Élément 4 (publication) :* The Weakest Failure Detector to Solve the Mutual Exclusion Problem in an Unknown Dynamic Environment [Mauffret et al., 2019] (meilleur papier Système à la conférence ICDCN 2019) qui définit le détecteur de fautes le plus faible pour résoudre le problème de l'exclusion mutuelle dans un système réparti dynamique.
5. *Élément 5 (publication) :* Accelerating Memcached using Safe In-kernel Caching and Pre-stack Processing [Ghigoff et al., 2021] publié dans la conférence NSDI en 2021 qui présente le cache BMC interne au noyau linux pour améliorer les performances de memcached.

3 AUTOÉVALUATION DU BILAN

3.1 Autoévaluation de l'équipe

Domaine 2. Attractivité

Référence 1. L'unité est attractive par son rayonnement scientifique et s'insère dans l'espace européen de la recherche.

Prix et distinctions. Depuis sa création en 2018, deux doctorants de l'équipe ont obtenu un prix. Alejandro Tomsic, a obtenu en 2019 le prix de meilleur thèse française en Système et Réseaux décerné par le GDR RSR et l'ACM Sigops France, pour sa thèse intitulé "Exploring the design space of highly-available distributed transactions". Sreeja Nair, ancienne doctorante de DELYS a obtenu le "Sephora Berrebi Scholarship for Women in Advanced ´ Mathematics & Computer Science" (2020, 3d edition). Pierre Sens a recu le prix 2022 du chercheur sénior en système décerné par le GDR RSD.

Par ailleurs, plusieurs de nos articles ont été récompensés :

- "Best paper award" à la conférence CLOSER 2018 [Bruneliere et al., 2018]
- "Best paper award", distributed computing Track, à la conférence ICDCN 2019 [Mauffret et al., 2019].
- "Best paper award" à la conférence NCA 2020 [Laniel et al., 2020]
- "Best student paper award" à la conférence NCA 2021 [Favier et al., 2020a]
- "Best student paper award" à la conférence OPODIS 2021 [Blin et al., 2021]
- "Best demo award" à la conférence CNSM 2021 [Alves Esteves et al., 2021c]

Invitations. Plusieurs membres d'équipe ont été invités à présenter leurs travaux dans *12 keynotes* de conférences et workshops :

- *Fault tolerance in dynamic distributed systems*. Invited keynote speaker, "Insights for the Future of Computing", LIG, Grenoble, Avril 2018.
- *Robustness : a New Form of Heredity Motivated by Dynamic Networks*, Invited speaker, 9th Workshop on GRAph Searching, Theory & Applications (GRASTA 2018), Berlin, Germany, Sept. 2018.
- *Just-Right Consistency : As available as possible, As consistent as necessary, Correct by design*. Invited speaker, Verification of Distributed Systems workshop, Essaouira, Morocco, May 2018.
- *Just-Right Consistency : As available as possible, As consistent as necessary, Correct by design*. Keynote presentation, DotScale, the European Tech Conference on Scalability, Distributed Systems & DevOps, Aubervillers, June 2018.
- *Just-Right Consistency*. Keynote presentation, Conf on Advances and Computing and Communication Engineering, Paris, June 2018.
- *Just-Right Consistency : As available as possible, Synchronous when necessary, Correct by design*. Invited Keynote talk, I/O Labs annual workshop, Châtillon, France, Oct. 2018.
- *Life after consistency*. Workshop of the EuroSys Program Committee, Dec. 2018.
- *Living on the edge, safely ; or : Life without consensus* at the 7th International Conference on Networked Systems, in Marrakech, Morroco, June 2019.
- *The programming continuum, from core to edge* at the Workshop on Verification of Distributed Systems (VDS), June 2019, Marrakech, Morroco.
- Dagstuhl Seminar on *Programming Languages for Distributed Systems and Distributed Data Management* October 2019.
- *Living Without Consensus*, at the seminar "Taking Stock of Distributed Computing," at Collège de France, April 2019.
- *Fault Tolerance in Dynamic Distributed Systems* at the 19th IEEE International Symposium on Network Computing and Applications (NCA 2020)

Par ailleurs, dans le cadre de visite dans les laboratoire à l'étranger, des membres de l'équipe ont donné les séminaires suivants :

- *Probabilistic Byzantine Tolerance Scheduling in Hybrid Cloud Environments*. Research Seminar, University of Fluminense, Brazil, October 2019

- *Fault tolerance in large and dynamic distributed systems*. Research Seminar, University of Fluminense, Brazil, October 2019
- *A Communication-Efficient Causal Broadcast Protocol*. Research Seminar, University of Fluminense, PUC-Rio, Brazil, October 2019

Activités éditoriale et participation à des comités de programme.

L'équipe participe au *comité éditorial de 3 revues* :

- Associate editor of International Journal of High Performance Computing and Networking (IJHPCN)
- Special Issue on Stabilization, Safety, and Security, Journal on Theory of Computing Systems (ToCS).
- Associate Editor for Letters of the IEEE Computer Society (LOCS).

Sur la période les membres de l'équipe ont participé à plus de *40 comités de programme* dont ISSRE, SRDS, DISC, PODC, Middleware, EuroSys, OSDI..., conférences majeures dans le domaine des systèmes et des algorithmes répartis.

Organisation de conférences.

L'équipe a participé à l'organisation de *4 conférences internationales* :

- Organiser of Dagstuhl Workshop on "Data Consistency in Distributed Systems : Algorithms, Programs, and Databases" (19-0117), February 2018. General Chair.
- PC Track Chair (track A : Theoretical and Practical Aspects of Stabilizing Systems) of the 20th International Symposium on Stabilization, Safety, and Security of Distributed Systems (SSS 2018).
- Latin-Americna Symposium og Dependable Computing (LADC), 2018, 2022. Tutorial Chair.
- 24th International Symposium on Stabilization, Safety, and Security of Distributed Systems (SSS), 2022. General Chair.

Par ailleurs, des membres de l'équipes sont membres de *3 comités de pilotage des conférences* : PaPoc, SSS et SBAC-PAD.

Pilotage de la recherche et d'expertise scientifique par des membres de l'équipe. Les membres de l'équipe ont été impliqués dans plusieurs instances de pilotage et d'animations de la recherche :

- Membre du Panel PE6 (Computer Science) of European Research Council (ERC) Starting Grants 2018.
- Membre CoNRS Section 6 (2016–2021)
- Membre Comité Exécutif du Labex SMART, Chair of Track 4, Autonomic Distributed Environments for Mobility (2012–2019)
- Vice-président de la Société informatique de France (SIF) (2019–2021)
- Membre de ACM Europe working group on European Research Visibility (RAISE).

Référence 2. L'unité est attractive par la qualité de sa politique d'accompagnement des personnels.

L'équipe a un recrutement diversifié des doctorants. Depuis 2018, nous avons accueilli, entre autre, deux doctorants normaliens (ENS Paris Saclay), 4 doctorants étrangers (3 brésiliens, 1 chilien) fruit de collaborations internationales, 4 étudiants en CIFRE.

Les doctorants de l'équipe sont co-encadrés par plusieurs permanents et certains doctorants ont été encadrés conjointement avec une autre équipe du LIP6 (2 avec NPA, 2 avec ALSOC, 1 avec Whisper) ou en co-tutelle avec une équipe étrangère (1 cotutelle avec une Université Brésilienne).

Nous avons accueilli également 6 professeurs dont Andrezj Pelc, ainsi que 3 doctorants d'autres équipes (1 de France, 2 du Brésil) pour des séjours entre 1 mois et 8 mois.

Référence 3. L'unité est attractive par la reconnaissance de ses succès à des appels à projets compétitifs.

L'équipe a une activité contractuelle soutenue et diversifiée.

Projets européens et internationaux. L'équipe a participé à *6 projets internationaux* :

- 1 projet européen H2020 (LightKone) en partenariat avec l'Université Catholique de Louvain (Belgium), Technische Universitaet Kaiserslautern (Germany), INESC TEC - Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciencia (Portugal), Faculdade de Ciencias E Tecnologiada Universidade Nova de Lisboa (Portugal), Universitat Politecnica De Catalunya (Spain), Scality (France), Gluk Advice B.V. (Netherlands)

- 1 projet Capes-Cofecub en partenariat avec Paris XI (LRI), le LIG, SUPELEC, Universidade de São Paulo - Instituto de Matemática e Estatística - Brazil, Unicamp - Instituto de Computação - Brazil
- 1 projet du ministère Espagnol de la recherche en partenariat avec le Labri, Irisa, University of the Basque Country UPV - Spain, EPFL - LSD - Switzerland, Friedrich-Alexander-Universitat Erlangen-Nurenberg - Deutschland, University of Sydney - Australia
- 3 projets STIC Amsud avec Universidade Federal do Rio Grande do Sul (Brazil)- Mãrcio Dorn, Universidad Nacional de San Luis (Argentina) Universidad de Santiago de Chile (Chile), Portales and Universidad Tecnica Federico Santa Maria (Chile), Universidade Federal de Uberlandia, Universidade Federal do Rio Grande do Norte and Instituto Federal Sul-Rio-Grandense (Brazil), Universidad de la Republica (Uuruguay), Universidade Federal Fluminense (UFF), Université d'Avignon, Mine Paristech, Université de Bordeaux, Université de Montpellier

Projets PIA et ANR. L'équipe a participé au LABEX SMART en tant membre de comité exécutif. Elle est également impliquée dans 2 PEPR (Cloud et Ensemble). Sur la période, DELYS a participé à *4 projets ANR dont 2 en tant que coordinateur*.

- le projet ESTATE (2016–2021) (coordinateur) en partenariat avec le LaBRI et Verimag
- Le projet RainbowFS (2016–2020) (coordinateur) en partenariat avec Scality SA, CNRS-LIG, Télécom Sud-Paris, Université Savoie-Mont-Blanc.
- Le projet AdeCoDS (2019–2023) en partenariat avec l' Université de Paris, ARM et Orange.
- Le projet SeMaFoR - (2021–2024) en partenariat avec LS2N-IMT Atlantique et AlterWay.

Domaine 3. Production scientifique

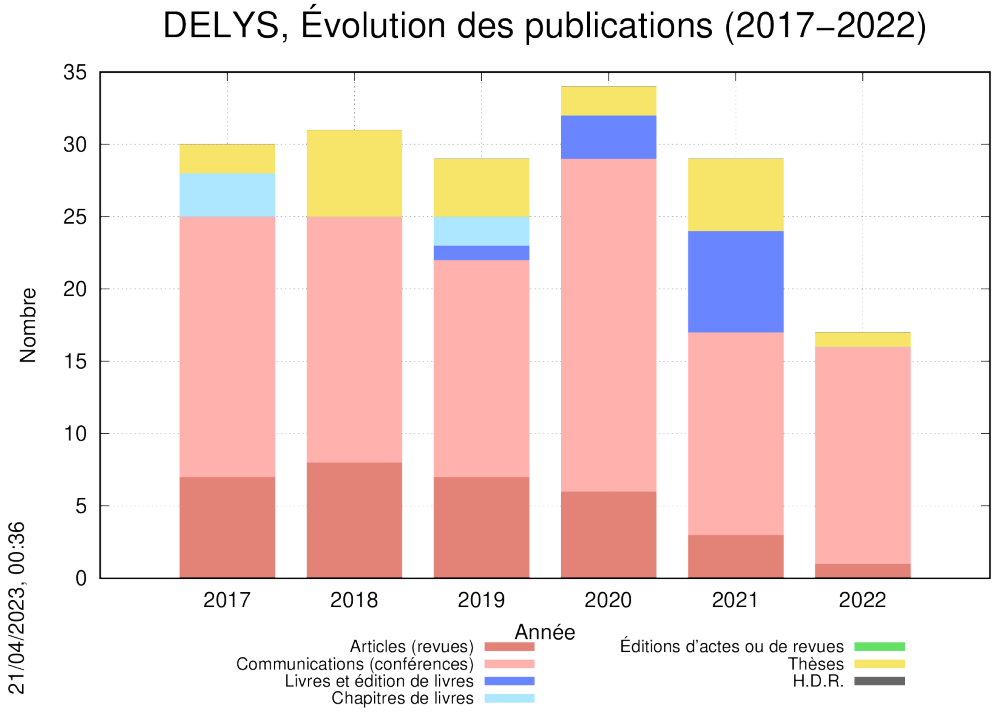

FIGURE 1 – Évolution des publications entre 2017 et 2022

	2017	2018	2019	2020	2021	2022
Articles (revues)	1.27	1.45	1.07	1.33	0.60	0.20
Communications (conférences)	3.27	3.09	2.30	5.11	2.80	3.00

TABLE 2 – Publications par ETPR par an entre 2017 et 2022

Référence 1. La production scientifique de l'unité satisfait à des critères de qualité.

L'équipe DELYS a une stratégie de publication visant entre autres les meilleures revues et conférences dans le domaine des systèmes et algorithmes répartis. En système, les meilleures conférences sont souvent plus sélectives que les revues.

DELYS a publié dans les revues majeures suivantes :

- Algorithmica : 1 [Bouchard et al., 2020e]
- Distributed Computing : 2 [Bouchard et al., 2019a, Dubois et al., 2019]
- FGCS : 1 [Al-Shara et al., 2018]
- JCSS : 1 [Censor-Hillel and Rabie, 2020]
- JPDC : 4 [Altisen et al., 2019c, Blin et al., 2018, de Araujo et al., 2018b, Rodrigues et al., 2018a]
- The Computer Journal : 3 [Graciela de Moraes Rossetto et al., 2018, Devismes et al., 2021, Bonnaire et al., 2017]
- TCS : 2 [Bournat et al., 2019, Casteigts et al., 2020]
- TPDS : 1 [Mosli Bouksiaa et al., 2019]
- TNSM : 1 [Alves Esteves et al., 2021a]

L'équipe a également publié dans les conférences majeures suivantes :

- ESOP : 1 [Nair et al., 2020a]
- EuroSys : 1 [Lepers et al., 2020]
- ICALP : 1 [Bouchard et al., 2018c]
- IJCAI : 1 [Arantes et al., 2018]
- ICPP : 1 [de Araujo et al., 2018a]
- Middleware : 2 [Tomsic et al., 2018, Toumlilt et al., 2021]
- NSDI : 1 [Ghigoff et al., 2021]
- PODC : 2 [Altisen et al., 2021a, Bouchard et al., 2020b]
- RTSS : 1 [Courtaud et al., 2019]
- SPAA : 1 [Bouchard et al., 2020d]
- SRDS : 2 [Johnen et al., 2021, Blin et al., 2022c]
- SIGMETRICS : 1 [Sun et al., 2017b]
- TACAS : 1 [Le Frioux et al., 2019]
- Usenix ATC : 2 [Bouron et al., 2018, Gouicem et al., 2020]

Parmi l'ensemble des publications depuis 2018, 27 ont été co-écrites avec des chercheurs internationaux.

Référence 2. La production scientifique de l'unité est proportionnée à son potentiel de recherche et correctement répartie entre ses personnels.

Depuis 2018, les doctorants ont été co-auteurs de 11 revues sur les 25 de l'équipe et de 62 conférences sur 85.

Référence 3. La production scientifique de l'unité respecte les principes de l'intégrité scientifique, de l'éthique et de la science ouverte. Elle est conforme aux directives applicables dans ce domaine.

L'équipe accorde une grande importance aux expérimentations, à l'absence de biais lors des mesures et à l'aspect reproductif des expériences. Lorsque des données sont collectées, les jeux de données sont fournies librement à la communauté. À titre d'exemple, nous avons collecté dans la plate-forme Planetlab la latence entre des machines géo-réparties. Les traces et le code permettant d'obtenir sont disponibles sur ce site `https://gitlab.lip6.fr/psens/latency-trace-planetlab`.

Domaine 4. Inscription des activités de recherche dans la société

Référence 1. L'unité se distingue par la qualité et la quantité de ses interactions avec le monde non-académique.

CRDTs. Nous avons participé à l'invention des CRDT (*Conflict-Free Replicated Data Types*) [1]. Un CRDT est un type de données répliquées que les développeurs peuvent utiliser pour construire des applications distribuées. Les répliques d'un CRDT peuvent être mises à jour en parallèle sans synchronisation et sont garanties de converger vers une valeur correcte, grâce à leurs propriétés mathématiques.

Les CRDT ont été largement adoptés par l'industrie. Plusieurs bases de données utilisent des CRDT : la base de données hautement disponible Riak utilise des CRDT dans ses clusters, Redis utilise des CRDT pour étendre sa fonctionnalité à la géo-distribution et Microsoft Azure propose des CRDT. Plusieurs applications industrielles utilisent directement les CRDT, notamment Facebook, Apple, TomTom, SoundCloud.

Référence 2. L'unité développe des produits à destination du monde culturel, économique et social.

Start-up. Durant la période, nous avons lancé la création de la start-up *Concordant.io* pour industrialiser les technologies développées dans l'Axe 2. Il s'agit d'une base de données CRDT décentralisée qui offre des garanties TCC+. Concordant a obtenu le soutien de l'Inria Startup Studio (ISS). Nous avons développé un premier prototype pour que les utilisateurs intéressés puissent développer des applications de de démonstration. Il s'agit d'un logiciel libre, disponible sur gitlab et github. Concordant s'est arrêté en avril 2022.

Brevet. Un brevet a été délivré sous le titre "Système de calcul distribué mettant en œuvre une mémoire transactionnelle matérielle de type non-spéculatif et son procédé d'utilisation pour le calcul distribué," en septembre 2019, tant en France qu'aux USA.

Référence 3. L'unité partage ses connaissances avec le grand public et intervient dans des débats de société.

L'équipe a participé à l'animation de la fête de la science en 2017 et 2018 (animation d'un atelier sur les algorithmes distribués).

4 RÉFÉRENCES BIBLIOGRAPHIQUES EXTERNES

[1] Marc Shapiro, Nuno Preguiça, Carlos Baquero, and Marek Zawirski. Convergent and commutative replicated data types. (104) :67–88, June 2011.

5 RÉFÉRENCES BIBLIOGRAPHIQUES SIGNIFICATIVES DE DELYS

[Al-Shara et al., 2018] Al-Shara, Z., Alvares, F., Bruneliere, H., Lejeune, J., Prud'Homme, C., and Ledoux, T. (2018). CoMe4ACloud : An End-to-End Framework for Autonomic Cloud Systems. *Future Generation Computer Systems*, 86 :339–354.

[Altisen et al., 2019a] Altisen, K., Devismes, S., Dubois, S., and Petit, F. (2019a). *Introduction to Distributed Self-Stabilizing Algorithms*, volume 8 of *Synthesis Lectures on Distributed Computing Theory*. Morgan & Claypool.

[Altisen et al., 2021a] Altisen, K., Devismes, S., Durand, A., Johnen, C., and Petit, F. (2021a). On Implementing Stabilizing Leader Election with Weak Assumptions on Network Dynamics. In *PODC '21 : ACM Symposium on Principles of Distributed Computing*, pages 21–31, Virtual Event, Italy. ACM.

[Altisen et al., 2021b] Altisen, K., Devismes, S., Durand, A., Johnen, C., and Petit, F. (2021b). Self-stabilizing Systems in Spite of High Dynamics. In *22nd International Conference on Distributed Computing and Networking, ICDCN'21*, ICDCN '21 : International Conference on Distributed Computing and Networking 2021, pages 156–165, Nara, Japan.

[Altisen et al., 2019c] Altisen, K., Devismes, S., Durand, A., and Petit, F. (2019c). Gradual stabilization. *Journal of Parallel and Distributed Computing*, 123 :26–45.

[Alves Esteves et al., 2020c] Alves Esteves, J. J., Boubendir, A., Guillemin, F., and Sens, P. (2020c). Location-based Data Model for Optimized Network Slice Placement. In *NetSoft 2020 - 6th IEEE International Conference on Network Softwarization*, pages 404–412, Ghent / Virtual, Belgium. IEEE.

[Alves Esteves et al., 2021a] Alves Esteves, J. J., Boubendir, A., Guillemin, F., and Sens, P. (2021a). A Heuristically Assisted Deep Reinforcement Learning Approach for Network Slice Placement. *IEEE Transactions on Network and Service Management*, pages 1–1.

[Alves Esteves et al., 2021c] Alves Esteves, J. J., Boubendir, A., Guillemin, F., and Sens, P. (2021c). DRL-based Slice Placement under Realistic Network Load Conditions. In *CNSM 2021 - 17th International Conference on Network and Service Management*, Izmir, Turkey.

[Arantes et al., 2018] Arantes, L., Bampis, E., Kononov, A., Letsios, M., Lucarelli, G., and Sens, P. (2018). Scheduling under Uncertainty : A Query-based Approach. In *IJCAI 2018 - 27th International Joint Conference on Artificial Intelligence*, pages 4646–4652, Stockholm, Sweden.

[Blin et al., 2018] Blin, L., Boubekeur, F., and Dubois, S. (2018). A Self-Stabilizing Memory Efficient Algorithm for the Minimum Diameter Spanning Tree under an Omnipotent Daemon. *Journal of Parallel and Distributed Computing*, 117 :50–62.

[Blin et al., 2021] Blin, L., Feuilloley, L., and Le Bouder, G. (2021). Optimal Space Lower Bound for Deterministic Self-Stabilizing Leader Election Algorithms. In *OPODIS 2021 - International Conference on Principles of Distributed Systems*, LIPIcs, Strasbourg, France.

[Blin et al., 2022c] Blin, L., Johnen, C., Le Bouder, G., and Petit, F. (2022c). Silent Anonymous Snap-Stabilizing Termination Detection. In *2022 41st International Symposium on Reliable Distributed Systems (SRDS)*, pages 156–165, Vienna, Austria. IEEE.

[Bonnaire et al., 2017] Bonnaire, X., Cortés, R., Kordon, F., and Marin, O. (2017). ASCENT : a Provably-Terminating Decentralized Logging Service. *The Computer Journal*, 60(12) :1889–1911.

[Bouchard et al., 2019a] Bouchard, S., Bournat, M., Dieudonné, Y., Dubois, S., and Petit, F. (2019a). Asynchronous approach in the plane : a deterministic polynomial algorithm. *Distributed Computing*, 32(4) :317–337.

[Bouchard et al., 2018c] Bouchard, S., Dieudonné, Y., and Lamani, A. (2018c). Byzantine Gathering in Polynomial Time. In *45th International Colloquium on Automata, Languages, and Programming (ICALP 2018)*, Prague, Czech Republic.

[Bouchard et al., 2020b] Bouchard, S., Dieudonné, Y., and Pelc, A. (2020b). Want to Gather ? No Need to Chatter ! In *PODC '20 - 39th Symposium on Principles of Distributed Computing*, pages 253–262, Salerno / Virtual, Italy. ACM.

[Bouchard et al., 2020d] Bouchard, S., Dieudonné, Y., Pelc, A., and Petit, F. (2020d). Almost Universal Anonymous Rendezvous in the Plane. In *SPAA '20 : 32nd ACM Symposium on Parallelism in Algorithms and Architectures*, pages 117–127, Virtual Event, United States. ACM.

[Bouchard et al., 2020e] Bouchard, S., Dieudonné, Y., Pelc, A., and Petit, F. (2020e). Deterministic Treasure Hunt in the Plane with Angular Hints. *Algorithmica*, 82(11) :3250–3281.

[Bournat et al., 2019] Bournat, M., Datta, A. K., and Dubois, S. (2019). Self-stabilizing robots in highly dynamic environments. *Theoretical Computer Science*, 772 :88–110.

[Bournat et al., 2018b] Bournat, M., Dubois, S., and Petit, F. (2018b). Gracefully Degrading Gathering in Dynamic Rings. In *Stabilization, Safety, and Security of Distributed Systems - 20th International Symposium, SSS 2018*, volume 11201 of *Lecture Notes in Computer Science*, pages 349–364, Tokyo, Japan. Springer.

[Bouron et al., 2018] Bouron, J., Chevalley, S., Lepers, B., Zwaenepoel, W., Gouicem, R., Lawall, J., Muller, G., and Sopena, J. (2018). The Battle of the Schedulers : FreeBSD ULE vs. Linux CFS. In *2018 USENIX Annual Technical Conference*, Boston, MA, United States.

[Bruneliere et al., 2018] Bruneliere, H., Al-Shara, Z., Alvares, F., Lejeune, J., and Ledoux, T. (2018). A Model-based Architecture for Autonomic and Heterogeneous Cloud Systems. In *CLOSER 2018 - 8h International Conference on Cloud Computing and Services Science*, volume 1, pages 201–212, Funchal, Portugal. Best Paper Award.

[Casteigts et al., 2020] Casteigts, A., Dubois, S., Petit, F., and Robson, J. (2020). Robustness : A new form of heredity motivated by dynamic networks. *Theoretical Computer Science*, 806 :429–445.

[Censor-Hillel and Rabie, 2020] Censor-Hillel, K. and Rabie, M. (2020). Distributed Reconfiguration of Maximal Independent Sets. *Journal of Computer and System Sciences*, 112 :85–96.

[Courtaud et al., 2019] Courtaud, C., Sopena, J., Muller, G., and Gracia, D. (2019). Improving Prediction Accuracy of Memory Interferences for Multicore Platforms. In *RTSS 2019 - 40th IEEE Real-Time Systems Symposium*, Hong-Kong, China. IEEE.

[de Araujo et al., 2018a] de Araujo, J. P., Arantes, L., Duarte Júnior, E. P., Rodrigues, L. A., and Sens, P. (2018a). A Communication-Efficient Causal Broadcast Protocol. In *ICPP 2018 - 47th International Conference on Parallel Processing*, Eugene, Oregon, United States.

[de Araujo et al., 2018b] de Araujo, J. P., Arantes, L., Duarte Júnior, E. P., Rodrigues, L. A., and Sens, P. (2018b). VCube-PS : A causal broadcast topic-based publish/subscribe system. *Journal of Parallel and Distributed Computing*.

[Devismes et al., 2021] Devismes, S., Lamani, A., Petit, F., Raymond, P., and Tixeuil, S. (2021). Terminating Exploration Of A Grid By An Optimal Number Of Asynchronous Oblivious Robots. *The Computer Journal*, 64(1) :132–154.

[Devismes et al., 2019] Devismes, S., Lamani, A., Petit, F., and Tixeuil, S. (2019). Optimal torus exploration by oblivious robots. *Computing*, 101(9) :1241–1264.

[Dubois et al., 2019] Dubois, S., Guerraoui, R., Kuznetsov, P., Petit, F., and Sens, P. (2019). The weakest failure detector for eventual consistency. *Distributed Computing*, 32(6) :479–492.

[Favier et al., 2020a] Favier, A., Guittonneau, N., Arantes, L., Fladenmuller, A., Lejeune, J., and Sens, P. (2020a). Topology Aware Leader Election Algorithm for Dynamic Networks. In *PRDC 2020 - 25th IEEE Pacific Rim International Symposium on Dependable Computing*, 2020 IEEE 25th Pacific Rim International Symposium on Dependable Computing (PRDC), pages 1–10, Perth, Australia.

[Ghigoff et al., 2021] Ghigoff, Y., Sopena, J., Lazri, K., Blin, A., and Muller, G. (2021). BMC : Accelerating Memcached using Safe In-kernel Caching and Pre-stack Processing. In *NSDI'21 - 18th USENIX Symposium on Networked Systems Design and Implementation*, pages 487–501, Virtual event, United States. USENIX Association.

[Gouicem et al., 2020] Gouicem, R., Carver, D., Lozi, J.-P., Sopena, J., Lepers, B., Zwaenepoel, W., Palix, N., Lawall, J., and Muller, G. (2020). Fewer Cores, More Hertz : Leveraging High-Frequency Cores in the OS Scheduler for Improved Application Performance. In *2020 USENIX Annual Technical Conference*, Boston / Virtual, United States. USENIX.

[Graciela de Moraes Rossetto et al., 2018] Graciela de Moraes Rossetto, A., Geyer, C., Arantes, L., and Sens, P. (2018). Impact FD : An Unreliable Failure Detector Based on Process Relevance and Confidence in the System. *The Computer Journal*.

[Johnen et al., 2021] Johnen, C., Arantes, L., and Sens, P. (2021). FIFO and Atomic broadcast algorithms with bounded message size for dynamic systems. In *SRDS 2021 - 40th International Symposium on Reliable Distributed Systems*, Chicago / Virtual, United States.

[Laniel et al., 2020] Laniel, F., Carver, D., Sopena, J., Wajsburt, F., Lejeune, J., and Shapiro, M. (2020). MemOpLight : Leveraging application feedback to improve container memory consolidation. In *NCA 2020 - 19th IEEE International Symposium on Network Computing and Applications*, pages 1–10, Cambridge / Virtual, United States.

[Le Frioux et al., 2019] Le Frioux, L., Baarir, S., Sopena, J., and Kordon, F. (2019). Modular and Efficient Divide-and-Conquer SAT Solver on Top of the Painless Framework. In Vojnar, T. and Zhang, L., editors, *TACAS 2019 - 25th International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, volume 11427 of *Lecture Notes in Computer Science*, pages 135–151, Prague, Czech Republic.

[Lepers et al., 2020] Lepers, B., Gouicem, R., Carver, D., Lozi, J.-P., Palix, N., Aponte, M.-V., Zwaenepoel, W., Sopena, J., Lawall, J., and Muller, G. (2020). Provable Multicore Schedulers with Ipanema : Application to Work Conservation. In *Eurosys 2020 - European Conference on Computer Systems*, Heraklion / Virtual, Greece. Virtual (online) conference.

[Mauffret et al., 2019] Mauffret, E., Jeanneau, É., Arantes, L., and Sens, P. (2019). The Weakest Failure Detector to Solve the Mutual Exclusion Problem in an Unknown Dynamic Environment. In *20th International Conference on Distributed Computing and Networking (ICDCN 2019)*, Bangalore, India. Extended version : https ://hal.archives-ouvertes.fr/hal-01661127v3.

[Mosli Bouksiaa et al., 2019] Mosli Bouksiaa, M. S., Trahay, F., Lescouet, A., Voron, G., Dulong, R., Guermouche, A., Brunet, E., and Thomas, G. (2019). Using differential execution analysis to identify thread interference. *IEEE Transactions on Parallel and Distributed Systems*, 30(12) :2866–2878.

[Nair et al., 2020a] Nair, S. S., Petri, G., and Shapiro, M. (2020a). Proving the safety of highly-available distributed objects. In *ESOP 2020 - 29th European Symposium on Programming*, Dublin, Ireland.

[Rodrigues et al., 2018a] Rodrigues, L. A., Duarte Júnior, E. P., and Arantes, L. (2018a). A distributed k-mutual exclusion algorithm based on autonomic spanning trees. *Journal of Parallel and Distributed Computing*, 115 :41–55.

[Sun et al., 2017b] Sun, W., Simon, V., Monnet, S., Robert, P., and Sens, P. (2017b). Analysis of a Stochastic Model of Replication in Large Distributed Storage Systems : A Mean-Field Approach. In *ACM Sigmetrics 2017- International Conference on Measurement and Modeling of Computer Systems*, pages 51–51, Urbana-Champaign, Illinois, United States. ACM.

[Teylo et al., 2020] Teylo, L., Arantes, L., Sens, P., and Drummond, L. M. A. (2020). A dynamic task scheduler tolerant to multiple hibernations in cloud environments. *Cluster Computing*.

[Tomsic et al., 2018] Tomsic, A. Z., Bravo, M., and Shapiro, M. (2018). Distributed transactional reads : the strong, the quick, the fresh & the impossible. In *2018 ACM/IFIP/USENIX International Middleware Conference*, Proceedings of 2018 ACM/IFIP/USENIX International Middleware Conference, page 14, Rennes, France. ACM/IFIP/USENIX, ACM.

[Toumlilt et al., 2021] Toumlilt, I., Sutra, P., and Shapiro, M. (2021). Highly-available and consistent group collaboration at the edge with colony. In *Middleware 2021 : 22nd International Middleware Conference*, pages 336–351, Québec / Virtual, Canada. ACM.

A ANNEXE — MEMBRES PERMANENTS AU 31/12/2022

La table ci dessous liste les membres permanents de l'équipe DELYS.

NOM	Prénom	Corps	Employeur
ARANTES	Luciana	MCF	Sorbonne Université
DARCHE	Philippe	MCF	Université Paris-Cité
DUBOIS	Swan	MCF	Sorbonne Université
FOLLIOT	Bertil	PR	Sorbonne Université
LEJEUNE	Jonathan	MCF	Sorbonne Université
MAKPANGOU	Mesaac	CR (HDR)	Inria
PETIT	Franck	PR	Sorbonne Université
SENS	Pierre	PR	Sorbonne Université
SOPENA	Julien	MCF	Sorbonne Université

ÉLÉMENT DE PORTFOLIO 01

Autre

1 DÉFINITION DE CET ÉLÉMENT

Titre de l'élément : Retour d'évaluation Inria

Fichier de élément : Expert-Inria-Report-Delys.pdf

2 MOTIVATIONS DU CHOIX DE CET ÉLÉMENT

Ce document présente l'évaluation faite par trois experts internationaux des systèmes distribués sur les activités de DELYS de 2018 à 2021. Ce rapport s'inscrit dans le cadre de la politique d'évaluation d'Inria où tous les quatre ans, les équipes Inria d'une même thématique scientifique sont évaluées par un panel d'experts internationaux. Le déroulé d'une évaluation suit les étapes suivantes :

1. L'équipe rédige un rapport scientifique transmis à trois experts du panel.
2. Les experts rédigent un rapport.
3. Le responsable d'équipe apporte des réponses écrites au rapport des experts, qui modifie leur rapport final.
4. L'équipe fait un retour au comité des projets de son centre Inria qui émet un avis sur l'équipe (renouvellement ou arrêt)
5. La commission d'évaluation (CE) nationale Inria émet un avis sur l'équipe (renouvellement ou arrêt)
6. Le direction du centre Inria de l'équipe décide en accord avec la direction scientifique nationale de la suite de l'équipe.

3 PRÉSENTATION DE CET ÉLÉMENT

Ce document est le retour d'évaluation Inria fait par les 3 experts lors de l'évaluation du thème "Distributed System and Middleware" d'Inria dont le séminaire d'évaluation a eu lieu en Juillet 2022.

Les trois experts extérieurs qui ont évalué l'équipe DELYS sont :

- Pascal Felber, Professeur Université de Neuchatel, Suisse
- Paola Flocchini, Professeur University of Ottawa,Canada
- Bertrand Lecun, Chercheur, Google

L'équipe a écrit un rapport d'évaluation sur la période 2018–2021 pour la reconduction de l'équipe projet Inria Commune DELYS. Suite à ce rapport, les évaluateurs ont produit le document présenté. En juin 2022, Pierre Sens a fait un retour d'évaluation au comité de projet d'Inria Paris qui a émis un avis *favorable* au renouvellement de l'équipe. La Commission d'Evaluation nationale d'Inria a également émis un avis *favorable* à la reconduction de l'équipe.

La direction du centre Inria Paris en accord avec la direction scientifique d'Inria n'a pas souhaité reconduire l'équipe projet commune qui s'est donc arrêtée au 31 décembre 2022 et continue sous la forme d'une équipe LIP6.

ÉLÉMENT DE PORTFOLIO 02

Publication

1 DÉFINITION DE CET ÉLÉMENT

Titre de l'élément : On Implementing Stabilizing Leader Election with Weak Assumptions on Network Dynamics

URL de l'élément : `https://hal.science/hal-02979166`

2 MOTIVATIONS DU CHOIX DE CET ÉLÉMENT

Les systèmes dynamiques sont au cœur des recherches de notre équipe. Nous étudions notamment les aspects algorithmiques selon plusieurs modèles de dynamiques. Au cours de la période d'évaluation, nous avons été porteurs de l'ANR ESTATE (ANR-16 CE25-0009-03) dont le but était de poser un cadre algorithmique permettant l'*Autonomic Computing* ou plus précisément, de concevoir un modèle qui comprenne des bases algorithmiques minimales permettant l'émergence de systèmes distribués hautement dynamiques dotés d'aptitudes auto-comportementales. Dans le cadre de l'ANR ESTATE, nous nous sommes plus particulièrement intéressés à la conception d'algorithmes auto-stabilisants pour des réseaux hautement dynamiques. Les résultats présentés ici s'inscrivent dans une étroite collaboration entre partenaires de l'ANR ESTATE, à savoir VERIMAG (Grenoble), le LaBRI (Bordeaux) et bien sûr DELYS. Ils constituent l'aboutissement d'une série de travaux traitant de mêmes problématiques, notamment [2, 3] pour ne citer que les plus représentatifs.

3 PRÉSENTATION DE CET ÉLÉMENT

Dans cet article [1], nous considérons l'auto-stabilisation et sa forme affaiblie appelée pseudo-stabilisaion. Nous étudions les conditions dans lesquelles l'élection d'un leader (pseudo- et auto-) stabilisé peut être résolue dans des réseaux soumis à des changements topologiques très fréquents. Pour modéliser une telle dynamique élevée, nous utilisons le paradigme des graphes dynamiques (DG) et étudions une taxonomie de neuf classes de DG importantes.

Nos résultats montrent que l'élection d'un leader auto-stabilisé ne peut être réalisée que dans les classes où tous les processus sont des sources, c'est-à-dire des processus qui sont infiniment souvent capables d'atteindre tous les autres par inondation. Nous montrons également que, parmi ces classes, le temps de convergence des solutions pseudo- et donc auto-stabilisantes ne peut être limité que dans la classe où toutes les sources sont réellement ponctuelles, c'est-à-dire toujours capables d'atteindre tous les autres processus en un temps borné. En outre, même l'élection d'un leader pseudo-stabilisant ne peut être résolue dans toutes les classes restantes, sauf dans la classe où au moins un processus est une source ponctuelle. Nous illustrons ce résultat en proposant un algorithme d'élection de leader pseudo-stabilisant pour cette dernière classe. Nous montrons que dans ce dernier cas, le temps de convergence des algorithmes d'élection de leader pseudo-stabilisant ne peut pas être limité. Néanmoins, notre solution est spéculative puisque son temps de convergence peut être borné lorsque la dynamique n'est pas trop erratique, précisément lorsque tous les processus sont des sources ponctuelles.

4 RÉFÉRENCES BIBLIOGRAPHIQUES

[1] Karine Altisen, Stéphane Devismes, Anaïs Durand, Colette Johnen, and Franck Petit. On Implementing Stabilizing Leader Election with Weak Assumptions on Network Dynamics. In *PODC '21 : ACM Symposium on Principles of Distributed Computing*, pages 21–31, Virtual Event, Italy, July 2021. ACM.

[2] Karine Altisen, Stéphane Devismes, Anaïs Durand, Colette Johnen, and Franck Petit. Self-stabilizing Systems in Spite of High Dynamics. In *22nd International Conference on Distributed Computing and Networking, ICDCN'21*, ICDCN '21 : International Conference on Distributed Computing and Networking 2021, pages 156–165, Nara, Japan, January 2021.

[3] Karine Altisen, Stéphane Devismes, Anaïs Durand, and Franck Petit. Gradual stabilization. *J. Parallel Distributed Comput.*, 123 :26–45, 2019.

ÉLÉMENT DE PORTFOLIO 03

Publication

1 DÉFINITION DE CET ÉLÉMENT

Titre de l'élément : Highly-available and consistent group collaboration at the edge with colony

URL de l'élément : `https://hal.inria.fr/hal-03353663`

2 MOTIVATIONS DU CHOIX DE CET ÉLÉMENT

Ce document illustre nos travaux sur la cohérence de données. Il correspond à la publication [2] publié à la conférence Middleware en 2021 qui introduit la nouvelle Cohérence Transactionnelle Causale Plus (TCC+) implémentée dans le système Colony. Ce travail a été réalisé conjointement avec Pierre Sutra de Telécom SudParis et a fait l'objet de la thèse d'Ilyas Toumlilt [1]. Le système Colony a été réalisé en partenariat avec Technische Universität Kaiserslautern (UniKL), Allemagne.

Les concepts de Colony ont été développés dans le cadre de deux projets auxquels DELYS a participé (Le projet européen LightKone et l'ANR RainbowFS). Colony a été valorisé, grâce au soutien de Inria Startup Studio, au sein de la start-up Concordant.

3 PRÉSENTATION DE CET ÉLÉMENT

La distribution et la réplication des données en bordure du réseau permettent une interrogation rapide, autonome des données et assure une meilleure disponibilité pour les applications, telles que les jeux, l'ingénierie coopérative ou le partage d'informations. Cependant, les utilisateurs exigent des garanties de cohérence les meilleures possibles ainsi qu'un support pour la collaboration de groupe.

Cet article présente la Cohérence Transactionnelle Causale Plus (TCC+) adapté aux configurations géo-distribuées. TCC+ propose un modèle hybride où une cohérence forte à base de *Snapshot Isolation* est appliquée au sein des groupes en périphérie du réseau ayant bonne connectivité et une cohérence plus relâchée est réalisée entre les groupes de zones géographiques distantes. Colony s'appuie sur une topologie de communication logique en arbre dont les racines sont répliquées dans un cloud central.

4 RÉFÉRENCES BIBLIOGRAPHIQUES

[1] Ilyas Toumlilt. *Colony : a Hybrid Consistency System for Highly-Available Collaborative Edge Computing*. Theses, Sorbonne Université, December 2021.

[2] Ilyas Toumlilt, Pierre Sutra, and Marc Shapiro. Highly-available and consistent group collaboration at the edge with colony. In *Middleware 2021 : 22nd International Middleware Conference*, pages 336–351, Québec / Virtual, Canada, December 2021. ACM.

ÉLÉMENT DE PORTFOLIO 04

Publication

1 DÉFINITION DE CET ÉLÉMENT

Titre de l'élément : The Weakest Failure Detector to Solve the Mutual Exclusion Problem in an Unknown Dynamic Environment [3]

URL de l'élément : `https://hal.science/hal-01661127v3`

2 MOTIVATIONS DU CHOIX DE CET ÉLÉMENT

Cette publication est représentative des travaux que nous menons sur la détection de fautes dans les systèmes dynamiques. La publication [3] a obtenu le prix du meilleur papier *Système* à la conférence ICDCN en 2019. Ces travaux ont été effectués dans le cadre du Labex SMART dont l'équipe était membre du comité exécutif.

Notre objectif est de proposer des bases algorithmiques pour concevoir des services robustes dans le cadre de systèmes dynamiques. Ces systèmes dynamiques couvrent un large spectre d'architectures réparties récentes et émergentes allant des Fogs aux réseaux de capteurs.

Cet article s'intéresse à l'exclusion mutuelle, service de synchronisation essentiel dès que deux entités partagent une ressource. Pour assurer la cohérence, il faut garantir un accès exclusif aux ressources partagées. On retrouve ce type de service dans de nombreuses plateformes [1]. Le problème de l'exclusion mutuelle a été largement étudié dans les systèmes répartis classiques où l'ensemble de nœuds est initialement connu et fixe [4]. Dans le cadre de systèmes tels que des réseaux de capteurs, les nœuds peuvent arriver, partir (suite à des défaillances) ou même se déplacer. Il est important alors d'avoir d'avoir des informations fiables sur les nœuds présents pour assurer la vivacité et la sûreté des algorithmes d'exclusion mutuelle.

3 PRÉSENTATION DE CET ÉLÉMENT

L'article définit le détecteur de fautes le plus faible pour résoudre le problème de l'exclusion mutuelle dans un système réparti dynamique. Résoudre l'exclusion mutuelle est impossible dans les systèmes asynchrones (i.e., sans borne sur les délais de transmission et de traitement des messages) en présence de défaillances. Les détecteurs de défaillance ont été introduits pour de contourner cette impossibilité et dans [2], il a été démontré que le détecteur de défaillance $\mathcal{T}$ est le plus faible pour résoudre l'exclusion mutuelle dans un système statique avec une majorité de nœuds corrects. Cet article étend ce résultat aux systèmes dynamiques en définissant le détecteur $(\mathcal{T}+\Sigma^l)$ qui permet de résoudre l'exclusion mutuelle dans des systèmes dynamiques où les nœuds peuvent tomber en panne puis revenir. En outre, il est démontré que ce détecteur est nécessaire et suffisant pour la résolution du problème.

4 RÉFÉRENCES BIBLIOGRAPHIQUES

[1] Michael Burrows. The chubby lock service for loosely-coupled distributed systems. In Brian N. Bershad and Jeffrey C. Mogul, editors, *OSDI'06), November 6-8, Seattle, WA, USA*, pages 335–350. USENIX Association, 2006.

[2] Carole Delporte-Gallet, Hugues Fauconnier, Rachid Guerraoui, and Petr Kouznetsov. Mutual exclusion in asynchronous systems with failure detectors. *JPDC*, 65(4) :492–505, apr 2005.

[3] Etienne Mauffret, Élise Jeanneau, Luciana Arantes, and Pierre Sens. The Weakest Failure Detector to Solve the Mutual Exclusion Problem in an Unknown Dynamic Environment. In *20th International Conference on Distributed Computing and Networking (ICDCN 2019)*, Bangalore, India, January 2019. Extended version : https ://hal.archives-ouvertes.fr/hal-01661127v3.

[4] Julien Sopena, Luciana Bezerra Arantes, and Pierre Sens. Performance evaluation of a fair fault-tolerant mutual exclusion algorithm. In *SRDS'06)*.

ÉLÉMENT DE PORTFOLIO 05

Publication

1 DÉFINITION DE CET ÉLÉMENT

Titre de l'élément : Accelerating Memcached using Safe In-kernel Caching and Pre-stack Processing

URL de l'élément : `https://hal.inria.fr/hal-03361644`

2 MOTIVATIONS DU CHOIX DE CET ÉLÉMENT

Ce document illustre nos travaux au niveau du système d'exploitation en lien également au sein des autres thématiques de l'équipe. Cette contribution, effectuée dans le cadre d'une collaboration avec Orange Labs, s'intègre directement dans les couches réseaux du système.

Ces travaux continuent toujours avec Orange Lab afin d'embarquer l'implémentation de BMC directement au sein d'une carte réseau programmable.

3 PRÉSENTATION DE CET ÉLÉMENT

La publication [1] dans la conférence NSDI en 2021 présente le cache BMC interne au noyau linux pour améliorer les performances de memcached.

En fournissant un accès ultrarapide aux données les plus populaires, les bases données clé/valeur en mémoire sont l'un des composants critiques pour assurer un passage à l'échelle des grands services Internet. Or, Memcached, l'un des standards dans l'industrie pour ce type de bases données, souffre de limitations de performance inhérentes à la pile réseau de Linux et ne parvient pas à utiliser pleinement la puissance des nouvelles interfaces réseau à haut débit. Bien que la pile réseau de Linux puisse être contournée à l'aide du framework DPDK, ces approches nécessitent une refonte complète de la pile logicielle et induisent une utilisation élevée de l'unité centrale, même lorsque la charge du client est faible.

Pour résoudre ce problème, nous avons conçu BMC, un cache intégré directement dans le noyau Linux qui sert les requêtes avant l'exécution de la pile réseau standard. Les demandes adressées au cache BMC sont traitées comme faisant partie des interruptions de la carte réseau, ce qui permet de faire évoluer les performances en fonction du nombre de cœurs desservant les files d'attente de la carte réseau. Pour garantir la sécurité, le BMC est mis en œuvre à l'aide de l'eBPF tout en maintenant un très haut niveau de performance. Ainsi, pour des requêtes de petite taille de type Facebook, BMC multiplie par 18 le débit d'un serveur Memcached. En outre, nos résultats montrent également que le BMC a un surcoût négligeable et ne détériore pas le débit lorsqu'il traite des charges de travail non ciblées.

4 RÉFÉRENCES BIBLIOGRAPHIQUES

[1] Yoann Ghigoff, Julien Sopena, Kahina Lazri, Antoine Blin, and Gilles Muller. BMC : Accelerating Memcached using Safe In-kernel Caching and Pre-stack Processing. In *NSDI'21 - 18th USENIX Symposium on Networked Systems Design and Implementation*, pages 487–501, Virtual event, United States, April 2021. USENIX Association.