Persistance de noyau dans les systèmes dynamiques à grande échelle

Vincent Gramoli*†, Anne-Marie Kermarrec*, Achour Mostéfaoui*, Michel Raynal*et Bruno Sericola*

*IRISA, INRIA et Université Rennes 1, Campus de Beaulieu, 35042 Rennes Cedex, France.
†INRIA Futurs, Parc Club Université, 91893 Orsay Cedex, France.
vgramoli@irisa.fr

La plupart des systèmes distribués modernes sont à la fois à grande échelle et dynamiques. Malgré l'avènement de tels systèmes, aucune des solutions que nous avons rencontrées ne permet de maintenir la persistance des données. Cet article étudie la portion des nœuds que quelqu'un doit contacter et la fréquence avec laquelle il doit le faire, en fonction du va-et-vient du système, afin d'assurer, avec une probabilité fixée, la persistance des données. Plus précisément, cet article met en relation le nombre d'éléments à contacter et la fréquence de contact, prouve ce résultat et analyse cette information à grande échelle.

Keywords: Persistance, Donnée, Probabilité, Noyau, Va-et-vient, Grande échelle

1 Introduction

Contexte. Maintenir la persistance d'une donnée dans un système distribué est une nécessité pour beaucoup d'applications. Bien que de nombreuses solutions aient été proposées dans un cadre statique, cela reste un probléme ouvert dans un cadre dynamique. Un système dynamique est un système ou les participants (*nœuds*) quittent et (re)joignent le système arbitrairement souvent. Pour pallier non seulement au dynamisme et au passage à l'échelle, les garanties probabilistes y remplacent naturellement les garanties déterministes fortes. Quantifier les garanties qu'il est possible d'obtenir probabilistiquement reste primordiale pour des applications dynamiques à grande échelle.

Plus précisément, un des problèmes fondamentaux d'un tel contexte consiste à assurer en dépit du dynamisme, que les données critiques ne soient pas perdues. L'ensemble des nœuds détenant une copie de la donnée critique est parfois appelé un *noyau*. Plusieurs noyaux peuvent coéxister, chacun associé à une donnée spécifique. Pourvu qu'un noyau reste suffisamment longtemps présent dans le système, la donnée peut être transmise de nœuds en nœuds à l'aide d'un protocole de "transfert de données" aboutissant à la création d'un nouveau noyau. Cependant, l'utilisation d'un tel protocole nécessite de faire un choix sur la fréquence du transfert de données pour éviter un surcoût mais assurer que la donnée ne disparaisse pas.

Contributions. Cet article assure probabilistiquement la maintenance d'un noyau. Les paramètres pris en considération sont la taille du noyau, le pourcentage de nœuds qui rentrent et sortent par unité de temps et la durée d'observation du système. Soit S le système à un temps τ. Il est composé de n nœuds dont q nœuds représentent le noyau Q d'une donnée critique. Soit S', le système au temps $\tau+\delta$. D'après l'évolution du système certains nœuds de S peuvent avoir quitté le système au temps $\tau+\delta$. Une question importante est la suivante : Étant donné un ensemble Q' de q nœuds de S' quel est la probabilité que Q et Q' s'intersectent ? Cet article exprime cette probabilité sous forme d'une fonction dont les paramètres caractérisent le système dynamique.

Travaux similaires. Des travaux portent sur le maintien du routage [RD01] en système dynamique. Néanmoins le degré de réplication nécessaire dépend complètement du nombre de pannes ou de départs: une donnée répliquée $k+1$ fois tolère seulement k pannes/départs. D'autres travaux [MTK06] étudient comment accéder à une donnée avec forte probabilité mais ne prennent pas en compte le dynamisme dans l'analyse.

2 Modèle

Le système est composé de n nœuds. Il est dynamique, dans le sens où cn nœuds rejoignent et quittent le système par unité de temps. Ces nœuds peuvent être vus comme étant remplacés, c représente donc le *va-et-vient* par nœud et par unité de temps. Un nœud quitte le système soit en se déconnectant volontairement ou lors d'une panne définitive. Lorsqu'un nœud rejoint le système il est considéré comme nouveau et la connaissance qu'il possédait du système est considérée comme perdue.

3 Relation entre les paramètres clefs du système dynamique

Cette section présente la relation des différents paramètres du système. Le Lemme suivant donne la portion C de nœuds qui sont remplacés après une période de δ unités de temps en fonction du va-et-vient c.

Lemma 3.1 *Soit C la portion de nœuds initiaux qui est remplacée après δ unités de temps. On obtient $C = 1-(1-c)^{\delta}$.*

Ce Lemme est facilement prouvable par récurrence sur δ (cf. [GKM+06]). Dans la suite de l'article on appellera α, le nombre de nœuds remplacés dans le système après δ unités de temps avec $\alpha = \lceil Cn \rceil = \lceil (1-(1-c)^{\delta})n \rceil$. Étant donné un noyau de q nœuds qui détiennent la donnée au temps τ, le Théorème suivant nous donne la probabilité de ne pas trouver cette donnée en contactant aléatoirement q nœuds après une période de δ unités de temps et le va-et-vient de α noeuds.

Theorem 3.2 *Soient $x_1,...,x_q$, des nœuds du système au temps $\tau' = \tau+\delta$. La probabilité qu'aucun de ces nœuds n'appartient au noyau initial est*

$$\frac{\sum_{k=a}^{b}\left[\binom{n+k-q}{q}\binom{q}{k}\binom{n-q}{\alpha-k}\right]}{\binom{n}{q}\binom{n}{\alpha}},$$

avec $\alpha = \lceil (1-(1-c)^{\delta})n \rceil$, $a = \max(0, \alpha-n+q)$ et $b = \min(\alpha, q)$.

Proof. Le problème à résoudre peut être modélisé de la façon suivante: Une urne contient n boules telles que, initiallement, q parmi les n sont vertes et $n-q$ sont noires. Les boules vertes représentent le noyau initial $Q(\tau)$ et sont représentées par l'ensemble Q. On tire aléatoirement et uniformément $\alpha = \lceil Cn \rceil$ boules de l'urne, on les peint en rouge et on les replace dans l'urne. Ces boules représentent les nœuds initiaux qui ont été remplacés après δ unités de temps.

Ainsi, l'état du système obtenu au temps $\tau' = \tau+\delta$ est différent. Soit A, l'ensemble des boules peintes en rouge et soit Q' le noyau Q après que certaines boules aient été peintes en rouge. Soit E l'ensemble des boules vertes, $E = Q' \setminus A$ au temps τ' et soit β le nombre de boules dans l'ensemble $Q' \cap A$. Nous savons que β possède une loi de distribution hypergéométrique, i.e., pour $a \leq k \leq b$ où $a = \max(0, \alpha-n+q)$ et $b = \min(\alpha, q)$, on a:

$$\Pr[\beta = k] = \frac{\binom{q}{k}\binom{n-q}{\alpha-k}}{\binom{n}{\alpha}}. \quad (1)$$

Finalement, on tire aléatoirement et successivement sans remise, q boules $x_1,...,x_q$ de l'urne (système au temps τ'). Le problème consiste à calculer la probabilité de l'événement {aucune des boules sélectionnées $x_1,...,x_q$ n'est verte}, pouvant s'écrire $\Pr[x_1 \notin E, ..., x_q \notin E]$.

Comme $\{x \in E\} \Leftrightarrow \{x \in Q'\} \cap \{x \notin Q' \cap A\}$, on obtient (en prenant le contraire) $\{x \notin E\} \Leftrightarrow \{x \notin Q'\} \cup \{x \in Q' \cap A\}$, ainsi on peut conclure $\Pr[x \notin E] = \Pr[\{x \notin Q'\} \cup \{x \in Q' \cap A\}]$. Étant donné que les événements $\{x \notin Q'\}$ et $\{x \in Q' \cap A\}$ sont disjoints, on obtient $\Pr[x \notin E] = \Pr[x \notin Q'] + \Pr[x \in Q' \cap A]$. Le système contient n boules. Le nombre de boules dans Q', A et $Q' \cap A$ est égal à q, α et β, respectivement. Par conséquent, on obtient, $\Pr[x_1 \notin E \mid \beta = k] = 1 - \frac{q-k}{n}$. Comme il n'y a pas de répétition, on obtient simplement de la même manière,

$$\begin{aligned}
\Pr[x_1 \notin E, ..., x_q \notin E \mid \beta = k] &= \sum_{k=a}^{b} \prod_{i=1}^{q} \left(1 - \frac{q-k}{n-i+1}\right), \\
&= \sum_{k=a}^{b} \frac{\binom{n-q+k}{q}}{\binom{n}{q}}, \\
\Pr[x_1 \notin E, ..., x_q \notin E] &= \sum_{k=a}^{b} \frac{\binom{n-q+k}{q}}{\binom{n}{q}} \Pr[\beta = k], \\
&= \frac{\sum_{k=a}^{b} \left[\binom{n+k-q}{q}\binom{q}{k}\binom{n-q}{\alpha-k}\right]}{\binom{n}{q}\binom{n}{\alpha}}.
\end{aligned}$$

□

3.1 Relation entre la taille du noyau q et la probabilité ε.

Étant donné une valeur C requise par le concepteur d'une application, d'autres paramètres influent sur le coût du maintien d'un noyau ou la garantie d'avoir un tel noyau. Le surcoût est facilement mesurable en utilisant le nombre q de nœuds à contacter.

On considère la valeur ε donnée par le Théorème 3.2. Cette valeur peut être interprétée de la façon suivante: $p = 1 - \varepsilon$ est la probabilité que, au temps $\tau' = \tau + \delta$, une des q requêtes exécutée (aléatoirement) par un nœud s'adresse à un nœud du noyau. Un problème important est alors de savoir quel est la relation entre ε et q et amène la question: Dans quelle mesure, augmenter q diminue-t'il ε ?

Cette relation est décrite sur la Figure 1 pour un système de $n = 10^4$ nœuds. Chaque courbe correspond à un différent taux de nœuds ayant quitté le système. La courbe montre que lorsque $10^{-3} \leq \varepsilon \leq 10^{-2}$, la probabilité $p = 1 - \varepsilon$ croît rapidement vers 1. Par exemple, la courbe représentant $C = 10\%$ montre qu'un noyau de $q = 224$ (resp. $q = 274$) nœuds assure une probabilité d'intersection de $1 - \varepsilon = 0.99$ (resp. 0.999).

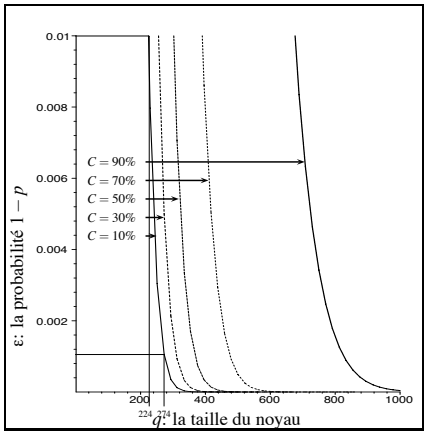

Fig. 1: Probabilité ε de ne pas contacter un noeud du noyau en fonction de sa taille q.

On remarque, que le résultat obtenu est similaire au paradoxe de l'anniversaire qui est paradoxal au sens de l'intuition et non de la logique. Pour que la probabilité que deux personnes dans la même pièce soient nés le même jour (toute année confondue) soit plus grande que $1/2$, il suffit qu'il y ait 23 personnes dans cette pièce. Lorsqu'il y a 50 personnes dans la pièce la probabilité est de 97% et devient 99.9996% pour 100 personnes.

Dans notre cas, ce phénomène se traduit par une forte augmentation de la probabilité d'accéder à un nœud du noyau lorsque la taille du noyau est seulement légèrement augmentée. Ainsi le concepteur du système peut augmenter la probabilité de contacter un nœud du noyau au prix d'un faible surcoût.

3.2 Relation entre la taille du noyau q et la période δ

Précédemment, on a montré qu'une application doit fixé C pour pouvoir définir le nombre q de nœuds à contacter permettant d'obtenir une probabilité p de réussite. Il existe un dernier compromis qu'un concepteur d'application doit faire: en fonction d'une probabilité $p = 1 - \varepsilon$, il s'agit de décider de la taille q, au dépend de la période δ après laquelle les q nœuds sont contactés. Ainsi, il est nécessaire de comparer précisément q à δ pour une valeur ε fixée.

Probabilité d'intersection	Va-et-vient $C = 1-(1-c)^{\delta}$	Taille du noyau $q(n=10^3)$	$q(n=10^4)$	$q(n=10^5)$
99%	statique	66	213	677
	10%	70	224	714
	30%	79	255	809
	60%	105	337	1071
	80%	143	478	1516
99.9%	statique	80	260	828
	10%	85	274	873
	30%	96	311	990
	60%	128	413	1311
	80%	182	584	1855

Fig. 2: La taille du noyau en fonction de la taille du système et du taux de va-et-vient.

Nous mettons en relation la taille et la durée de vie d'un noyau lorsque la probabilité de contacter un nœud du noyau est de 99% ou 99,9%. Ces valeurs ont été choisies comme pouvant refléter les choix d'un concepteur d'applications. Pour chacune des deux probabilités, nous avons représenté, sur la Figure 2, la taille d'un quorum en fonction du nombre de nœuds ayant quitté le système durant la période δ. La Figure 2 met en valeur cette relation dans un système statique et dans un système dynamique (pour différentes valeurs de C). Le système statique implique qu'aucun nœud ne quitte ou ne rejoint le système durant la période δ alors que dans le système dynamique une certaine portion C des nœuds quitte et rejoint le système durant cette période δ. Dans un souci de simplicité de présentation, nous avons représenter C plutôt que δ. Les résultats obtenus mènent à deux observations intéressantes.

D'une part, lorsque δ est assez grand pour que 10% des nœuds soient remplacés, alors la taille du noyau nécessaire est étonamment proche de celle du cas statique (873 contre 828 lorsque $n = 10^5$ pour une probabilité de 0,999). De plus, $q = 990$ est suffisant lorsque C passe à 30%. D'autre part, lorsque la période δ est suffisamment large pour que 80% du système soit remplacé, le nombre de nœuds q à contacter reste bas comparé à la taille dy système. Par exemple, dans le cas ou 6 000 nœuds sur 10 000 sont remplacés, alors seulement 413 nœuds doivent être contactés pour obtenir une intersection avec probabilité 0,999.

4 Conclusion

Maintenir une donnée critique dans un système ou les nœuds partent et arrivent dynamiquement est un problème difficile. Dans cet article, nous avons défini la notion de noyau persistant permettant de maintenir la donnée indépendamment de la structure sous-jacente utilisée.

Nos résultats apportent une aide aux concepteurs d'application pour ajuster les paramètres en fonction du dynamisme et de la garantie recherchée. Un de nos résultats montre qu'augmenter légèrement la taille du noyau suffit à augmenter la garantie de beaucoup, même avec un fort dynamisme. Ce travail ouvre la voie à de nouvelles recherches tant dans le domaine de la cohérence mémoire que l'évaluation de l'intensité du va-et-vient.

Références

[GKM+06] Vincent Gramoli, Anne-Marie Kermarrec, Achour Mostéfaoui, Michel Raynal, and Bruno Sericola. Core persistence in peer-to-peer systems: Relating size to lifetime. In *Proceedings of the OTM'06 International Workshop on Reliability in Decentralized Distributed Systems*, volume 4278 of *LNCS*. Springer, April 2006.

[MTK06] Ken Miura, Taro Tagawa, and Hirotsugu Kakugawa. A quorum-based protocol for searching objects in peer-to-peer networks. *IEEE Transactions on Parallel and distributed Systems*, 17(1):25–37, January 2006.

[RD01] Antony Rowstron and Peter Druschel. Pastry: Scalable, distributed object location and routing for large-scale peer-to-peer systems. In Rachid Guerraoui, editor, *Middleware 2001*, volume 2218 of *LNCS*, pages 329–350. Springer-Verlag, 2001.