

Service de Cache pour les Grilles de Calcul ¹

Yonny Cardenas, Jean-Marc Pierson et Lionel Brunie

LIRIS, CNRS UMR 5205

INSA de Lyon, Bât. B. Pascal, 7 av. Jean Capelle, 69621 Villeurbanne cedex,
France

<prenom>.<nom>@liris.cnrs.fr

<http://liris.cnrs.fr>

Résumé. Nous proposons un système fédérateur de caches pour les grilles que les applications de la grille utilisent comme un service de cache uniforme. Le système est fondé sur le concept de l'activité de données où les applications partagent et réutilisent l'information sémantique liée à l'activité des données sous la forme de métadonnées. Ces métadonnées représentent la connaissance sur les données et sur leur gestion. Elles permettent d'optimiser, suivant le contenu et l'utilisation de ces données, leur placement, leur recherche, leur durée de vie et leur pertinence vis-à-vis de leur exploitation.

1 Introduction

Les grilles de calcul sont la conséquence de l'évolution des systèmes distribués. Leur principale caractéristique est la capacité à réunir dynamiquement des ressources distribuées pour offrir des moyens de calcul et de stockage à grande échelle. Dans ce cadre, plusieurs institutions peuvent établir une « organisation virtuelle » à partir de ces infrastructures informatiques (Foster et al. 2003). Ainsi, les grilles de calcul mettent en oeuvre la technologie permettant l'intégration et le partage de données entre différentes organisations.

Les mécanismes d'accès aux données qui sont disponibles aujourd'hui dans les intergiciels de grille OGSA-WSRF (Globus 2005) sont assez rudimentaires pour le traitement et la gestion de données complexes, ce qui est le cas par exemple dans le domaine médical, domaine cible du projet RagTime ¹. Cela amène potentiellement à une prolifération de copies des données alors même que les mécanismes d'utilisation des métadonnées sur la structure, la disponibilité et la localisation de ces données ne sont pas assez avancés. La gestion des métadonnées sur la grille souffre de l'absence de mécanismes de haut niveau pour le partage d'information. Ce partage permettrait aux applications d'exploiter efficacement (en particulier pour des applications d'extraction des connaissances) la disponibilité de certaines données présentes sur la grille. Le service de cache pour grilles présenté ici a comme double objectif de permettre un accès global aux données dispersées dans la grille et d'optimiser l'usage des ressources de données (stockage et communication des données).

¹ Ce travail est soutenu par la Région Rhône-Alpes (project RagTime), et l'ACI Masse de Données.

2 L'extraction des connaissances et l'activité des données

Les applications dans les plateformes de grille de calcul ont besoin de déterminer les données les plus appropriées pour la résolution d'un problème particulier. Elles ont besoin de connaître la signification ou la sémantique des entités de données, c'est-à-dire des propriétés pour la description à bas niveau des données. Les métadonnées sont des étiquettes sémantiques utilisées pour la description des données. Elles sont utilisées pour indiquer des caractéristiques fondamentales comme le type, la taille, le format, etc. De plus, elles peuvent être utilisées pour enrichir les descriptions de données telles que l'origine, les paramètres de création, la provenance par rapport aux algorithmes employés, et les annotations. Ces informations permettent d'établir la nature des données et leur cycle de vie.

Notre idée est basée sur l'importance de cette notion d'activité des données. Les données qui sont intéressantes sont celles qui se déplacent (copies, téléchargement, etc), et plus généralement celles qui ont des activités de type création, mise à jour, etc. La présence des métadonnées augmente la probabilité que les données attachées aient une plus grande activité. Les données avec le plus d'activité sont les données avec les descriptions les plus riches.

3 Caches coopératifs et grilles de calcul

Les systèmes de cache sont des techniques qui permettent de gérer l'accès aux copies des données qui ont une haute fréquence d'accès, c'est à dire une forte activité. Les applications « grille » utilisent souvent des données qui ont un coût élevé en termes de ressources de calcul et de communication pour les collecter, les accéder, ou les transporter. Les systèmes de cache permettent qu'elles soient facilement et rapidement disponibles pour des applications similaires ou qui ont des objectifs proches.

Plusieurs propositions de cache collaboratifs ont été conçues, la plupart pour le trafic web (Barish et al. 2000). Ce trafic web a cependant des comportements très différents par rapport aux modèles d'accès de la grille, qui se caractérisent par un grand mouvement de masses de données. L'utilisation de la sémantique pour la gestion de caches collaboratifs a montré son intérêt dans le travail (Pierson 2002) pour les caches web. Ce même principe peut être élargi pour une gestion collective efficace de cache dans la grille de calcul.

4 Proposition

Nous proposons un service de cache pour les grilles de calcul géré suivant l'activité des données, s'appuyant sur l'utilisation sémantique de métadonnées liées à ces données.

Le système de cache offre l'accès aux données pour les applications de la grille en utilisant des descriptions sémantiques qui correspondent aux métadonnées associées à ces données. Celles-ci permettent de mesurer l'activité de ces données. Le système est élargi à toute la grille avec l'utilisation d'une architecture de caches coopératifs.

L'architecture générale du service de cache uniforme est constituée d'une part par des caches locaux et autonomes, et d'autre part par une couche transversale distribuée réalisant la coordination et la collaboration au niveau global. Les composants distribués de la couche

transversale assurent une vision et une gestion globale sur l'ensemble de ces caches autonomes au niveau de toute la grille. Ces composants distribués fournissent des capacités dynamiques de découverte, d'enregistrement, d'accès, de validation et d'invalidation des données des caches locaux, en utilisant les métadonnées associées.

Les composantes essentielles de l'architecture sont : le *Local Grid Cache* (Section 4.1) et le *Collective Grid Cache* (Section 4.2). Ils implémentent les trois fonctionnalités principales du service de cache pour les grilles.

- *Fonctionnalité d'Enregistrement* : Capacité de recevoir et gérer des métadonnées sur les données et leur activité, en utilisant des catalogues de métadonnées. Le système peut recevoir des métadonnées provenant de différents types de sources (applications ou autres services de la grille).
- *Fonctionnalité d'Accès* : Capacité de rechercher des données (requête applicative) dans le cache local ou dans l'ensemble des caches distribués de la grille, en s'appuyant sur les métadonnées associées aux données, récupération et déplacement depuis le système de stockage.
- *Fonctionnalité de Gestion* : Capacité de gérer les ressources (stockage, base de données) et services (transport de données, communications) qui a donc besoin le cache pour mettre en oeuvre ces fonctions. Elle comprend aussi la capacité de coopérer entre caches en s'appuyant sur l'utilisation de protocoles de communication inter-cache. Celui-ci supporte l'échange de résumés du contenu des catalogues des caches locaux.

4.1 Local Grid Cache

Il implémente les fonctionnalités du service de cache en ce qui concerne chaque organisation individuellement (niveau local). Il expose ces fonctionnalités depuis la grille grâce à des interfaces standards établies pour l'interopérabilité grille. Il est l'élément basique pour tout ce qui concerne la capture de métadonnées sur l'activité des données près des applications à l'intérieur de l'organisation. Il traite les requêtes de données depuis les applications soit en utilisant les catalogues locaux, soit en collaboration avec les autres caches grâce au composant collectif du service de cache. Finalement, il traduit des opérations d'accès aux données en provenance de la grille en opérations supportées par les systèmes de cache à l'intérieur de l'organisation et vice versa.

4.2 Collective Grid Cache

Il implémente la fonctionnalité de cache au niveau collectif. Il est l'élément fédérateur pour tout ce qui concerne la collaboration entre l'ensemble des caches distribués. Pour faire la gestion globale, il exécute les actions suivantes :

Il réunit l'information de l'activité globale des données à partir des rapports provenant de l'ensemble des caches distribués. Il gère cette information avec un catalogue distribué des métadonnées d'activité.

Il fait des analyses de l'information dans ce catalogue distribué des métadonnées d'activité et en extrait l'information utile pour la gestion du système uniforme de cache. Par exemple, il compare des informations qui ont été réunies par la fonctionnalité d'enregistrement afin d'établir la présence de plusieurs copies de la même entité de données

sur plusieurs caches. A partir de cela, il peut réduire le nombre de copies. Pour cela il sélectionne le cache selon l'activité de données en question (copie) dans le cache local.

5 Conclusion

Nous avons présenté un service de cache uniforme pour les environnements de grille de données où les applications ont besoin de mécanismes spécifiques pour l'accès efficace aux ressources de données. Nous proposons une architecture pour le service de cache dans les grilles basée sur l'exploitation de métadonnées.

Notre service de cache uniforme offre de manière classique deux fonctions basiques aux applications de la grille : en premier lieu, les applications s'adressent au service en tant que fournisseurs de données et métadonnées au cache, lequel à son tour gère l'accès à ces données depuis d'autres applications ; en second lieu, les applications font des requêtes de données au cache uniforme, lequel les résoud sur toute la grille grâce aux mécanismes de collaboration. Une troisième fonction se spécialise en collaboration inter-cache pour une gestion efficace des ressources de stockage et de communications utilisées par l'ensemble des caches. Un prototype de ce service de cache pour les grilles a été développé sur la plateforme OGSA-Globus version 3.2.1 (Globus 2005).

Références

- Foster, I., Kesselman, C., Nick, J., and Tuecke, S. (2002) , The Physiology of the Grid : an Open Grid Services Architecture for Distributed Systems Integration, Global Grid Forum technical report.
- Barish G, Obraczka K.(2000), World Wide Web Caching : Trends and Techniques, IEEE Communications Magazine 2000; 38(5) : 178-184.
- J-M. Pierson, L. Brunie, and D. Coquil (2002), Semantic collaborative web caching, Web Information Systems Engineering, (ACM/IEEE WISE 2002), pp. 30,39

Summary

We suggest a federal cache system for grid computing the applications use as an uniform cache service. The system is derived from the notion of data activity where the applications share and reuse semantic information through metadata related with these activities. These metadata represent the knowledge of data and their management. Metadata allow to optimize the placement, the lifecycle and the pertinence of data for future utilization.