Calcul des seuils optimaux d'une politique à hysteresis pour un modèle de cloud

Mohamed Mehdi Kandi, Hind Castel-Taleb, Farah Ait Salaht, Emmanuel Hyon

Calcul des seuils optimaux d'une politique à hysteresis pour un modèle de cloud

M.M. Kandi[1], H. Castel-Taleb[1], F. Aït-salaht[2,3], E. Hyon[3,4]

[1] SAMOVAR, UMR 5157, Télécom Sud Paris, Evry, France,
Castel@it-sudparis.eu

[2] Crest Ensai, farah.ait-salaht@ensai.fr

[3] Sorbonne Universités, UPMC Univ Paris 06, UMR 7606, LIP6,

[4] Université Paris Ouest, Nanterre, France, emmanuel.hyon@u-paris10.fr

Mots-clés : *Cloud computing, optimisation d'énergie, Politique à hysteresis, calcul des seuils.*

1 Introduction

Le Cloud Computing est une technologie qui permet d'accéder, à la demande, à un ensemble de services fournis par des ressources informatiques virtualisées, mutualisées et partagées. Un des concepts clé des architectures de cloud est la virtualisation qui permet une utilisation flexible et rentable des ressources, et donc un meilleur rendement énergétique tout en garantissant les performances. Lorsque la charge du système est faible, les machines virtuelles (VMs) non utilisées sont désactivées (mises en veille) et consomment alors moins d'énergie. Par contre, lorsque la charge augmente les VMs sont activées afin de compenser le surplus de travail. Toutes ces considérations énergétiques sont cruciales dans les environnements de cloud [6] tout comme pour les data centers puisqu'elles représentent la majeure partie des coûts de fonctionnement et impactent donc la compétitivité du cloud.

Notre objectif dans ce travail consiste à gérer les activations et désactivations des serveurs de façon à garantir à la fois Qualité de Service (QoS) et consommation énergétique. Nous nous focalisons sur les politiques à hysteresis qui sont des politiques avec séquences de seuils déclenchant soit l'activation soit la désactivation des machines virtuelles selon l'intensité de la charge du système. Ces activations et désactivations dynamiques des serveurs (ou machines virtuelles) permettent de représenter la variabilité du nombre de ressources disponibles dans le cloud en fonction de la scalabilité de la demande, tandis que la notion d'hysteresis permet d'éviter les effets de ping-pong ce qui la rend plus pertinente qu'une politique à un seul seuil. Les politiques à hysteresis présentent des facilités de mise en oeuvre et sont optimales dans une majorité de modèles Markoviens [1]. C'est pourquoi ce type de politiques est largement appliqué dans les modèles d'analyse de cloud et a déjà été étudié dans [2, 3, 7] par exemple.

Notre but est de calculer efficacement les seuils optimaux d'activation et désactivation en fonction des coûts, en considérant à la fois des aspects d'évaluation de performances ainsi que des aspects énergétiques qui généralement n'ont été abordés que séparément dans les travaux précédents.

2 Formulation du problème : modèle d'optimisation

Afin de représenter l'aspect dynamique de ce système, nous considérons le modèle de file d'attente suivant : le modèle considéré est à temps continu et ponctué par les arrivées ou les départs du système. Nous supposons que le processus d'arrivée dans la file est poissonnien de taux λ et que la file est de taille finie. Le système est composé de K serveurs (VMs) supposés

homogènes dont le service est exponentiel de taux μ. On suppose qu'il y a au moins un serveur actif dans le système et au plus K. De plus, sachant que le modèle est à capacité finie, un certain nombre de clients seront amenés à être perdus lorsque la capacité de la file est atteinte. L'état du système est donné par $\mathcal{X} = (x, s)$ où x et s représentent respectivement le nombre de clients dans le système et le nombre de serveurs actifs.

Pour les coûts, nous considérons deux types de coûts : les coûts liés à la QoS du client et les coûts liés à la consommation énergétique. Certains d'entre eux sont instantanés et d'autres s'accumulent au cours du temps. Ainsi, nous supposons que c_P est le coût de la perte d'un client, c_{Sa} est le coût associé à l'activation d'un serveur, et c_{Sd} celui associé à la désactivation d'un serveur, tandis que c_H est le coût par unité de temps du nombre de client dans la file et c_U le coût par unité de temps associé à l'utilisation d'un serveur actif.

Nous cherchons à minimiser les coûts moyens. Il y a plusieurs manières de calculer ces seuils soit par des approches de programmation dynamique stochastique soit en exprimant le coût moyen en fonction des seuils puis en appliquant un algorithme de recherche (le plus souvent locale) comme dans [5]. C'est cette dernière approche que nous appliquons et adaptons ici. Elle se base sur le fait qu'une fois l'ensemble des valeurs des seuils fixées le système peut-être décrit comme une chaîne de Markov. Il suffit alors de calculer la loi invariante pour ensuite en tirer les performances moyennes. La résolution avec formule close de la chaîne est proposée dans [4].

L'article [5] propose trois heuristiques de recherche locale pour des modèles comportant uniquement un seul type de seuil. Nous proposons une adaptation de chacune de ces trois heuristiques pour des modèles à deux types de seuil. Ces heuristiques sont implémentées et comparées entre elles. Nous constatons alors que ces heuristiques sont très efficaces mais qu'elles ne retournent pas forcément la solution optimale (calculée par énumération sur de petits exemples) contrairement à celles de Kranenbourg [5]. Les résultats (pour une taille de 16 serveurs et de 40 clients maximum) se trouvent dans le tableau ci-dessous et plaident pour des recherches plus poussées sur l'adaptation d'une méthode exacte efficace pour notre modèle.

-	Pourcentage de solutions optimales retournées	Temps moyen d'exécution	Temps maximal d'exécution
Heuristique 1	94%	0.441 sec	2.424 sec
Heuristique 2	95, 5%	0.386 sec	1.285 sec
Heuristique 3	86, 6%	0.119 sec	0.510 sec
Énumération	100%	143.515 sec	382.749 sec

Références

[1] I.J.B.F. Adan, V.G. Kulkarni, and A.C.C. van Wijk. Optimal control of a server farm. *INFOR*, 51(4) :241–252, 2013.

[2] F. Aït-Salaht and H. Castel-Taleb. Bounding aggregations on phase-type arrivals for performance analysis for performance analysis of clouds. In *MASCOTS 2016*. IEEE, 2016.

[3] V. Goswami, S.S. Patra, and G.B. Mund. Performance analysis of cloud with queue-dependent virtual machines. In *1st RAIT*, 2012.

[4] M.M. Kandi. Garantie de la QoS et de la consommation énergétique dans le cloud. Master's thesis, UPMC, 2016.

[5] A.A. Kranenburg and G.J. van Houtum. Cost optimization in the (S-1, S) lost sales inventory model with multiple demand classes. *Oper. Res. Lett.*, 35(4) :493–502, 2007.

[6] P.J. Kuehn and M.E. Mashaly. Automatic energy efficiency management of data center resources by load-dependent server activation and sleep modes. *Ad Hoc Networks*, 2015.

[7] Isi Mitrani. Managing performance and power consumption in a server farm. *Annals of Operations Research*, 202 :121–134, 2013.