Étude de l'impact de la flexibilité sur la QoS et la puissance consommée dans un datacenter vert

Damien Landré, Laurent Philippe, Jean-Marc Pierson

Étude de l'impact de la flexibilité sur la QoS et la puissance consommée dans un datacenter vert

Damien Landré[1,2], Laurent Philippe[1], Jean-Marc Pierson[2]

[1] Institut FEMTO-ST, département DISC, France {lphilipp,damien.landre}@femto-st.fr
[2] Institut de Recherche en Informatique de Toulouse, France
{damien.landre,jean-marc.pierson}@irit.fr

Mots-clés : *datacenter, énergies renouvelables, workload, qualité de service*

1 Introduction

La demande en ressources énergétiques des datacenters n'a cessé de croître, représentant aujourd'hui plus de 1% de la consommation totale d'électricité [3]. Les projets ANR DATAZERO [4] (2015-2019) et DATAZERO 2 (2020-2024) ont pour objectif de concevoir et exploiter un datacenter autonome fonctionnant uniquement aux énergies renouvelables locales. L'élément central de ce projet est la mise en place d'une négociation entre la partie électrique et la partie IT du datacenter afin de concilier la prévision d'offre et de demande en énergie. Elle se fait régulièrement au moyen d'échanges de profils de puissance entre les deux parties. Un profil de puissance est représenté par une série temporelle de valeurs de puissance par pas de temps de 15 minutes à 1 heure. Du côté de la partie IT, le calcul des profils de puissance s'appuie sur la prévision de la charge de travail (workload) soumise par les utilisateurs et sur les ressources informatiques disponibles, sous la contrainte d'une qualité de service [1] mesurée par la part de charge non traitée car dépassant sa deadline. Dans un travail précédent [1], un algorithme dichotomique a été proposé. Il détermine pour chaque pas de temps du profil la valeur de puissance nécessaire au traitement du workload, sous les contraintes d'une qualité de service et de temps. Le gain en efficacité a été étudié avec des heuristiques déterminant une configuration des machines minimisant la puissance consommée. Nous étudions ici l'impact de la flexibilité sur la qualité de service et la réduction de la consommation énergétique.

2 Objectif

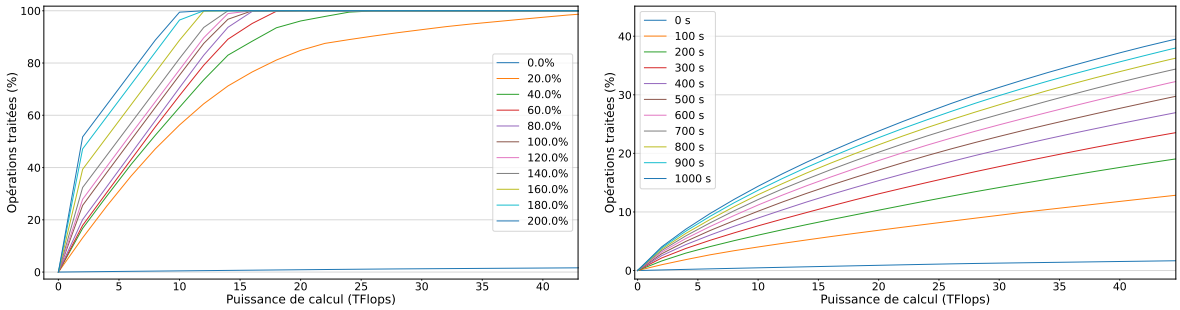
Il est primordial d'être efficace dans la demande en énergie pour ne pas mettre en péril la robustesse du datacenter. Plusieurs leviers permettent d'accroître cette efficacité, par exemple minimiser le nombre de machines allumées. Par ailleurs, le workload varie et est flexible [5]. Ce sont les pics de charge qui déterminent la configuration des machines nécessaire pour traiter l'enemble de la charge : il faut suffisamment de machines pour traiter ces pics. Lisser ces pics permet alors de diminuer le besoin de puissance maximum demandé par les machines tout en respectant la contrainte de qualité de service. Le lissage des pics est cependant contraint par les deadlines associées à la charge. L'objectif est donc d'étudier l'impact de la flexibilité des deadlines sur la qualité de service et la puissance consommée par les machines. Pour ceci nous contraignons plus ou moins les deadlines et observons la part de la charge non traitée avec un workload réel.

Peu de traces de workloads réels couvrant de larges périodes sont disponibles pour les datacenters. Pour cette étude, nous considérons donc le workload Metacentrum [2] issu du Parallel Workload Archive [1]. Le workload est filtré selon certains critères sur une période d'un an.

1. www.cs.huji.ac.il/labs/parallel/workload/

3 Résultats

Pour évaluer l'impact de la durée des deadlines sur la qualité de service, nous associons des deadlines différentes sur chacun des jobs du workload. Le premier jeu de deadlines (non-identiques) repose sur l'exécution réelle sur les machines de Metacentrum. Elles sont calculées sur la base d'un pourcentage du temps d'attente plus du temps de calcul. Le second jeu applique des deadlines identiques à tous les jobs. La Figure 1 montre la qualité de service obtenue en fonction de la puissance de calcul et de la flexibilité appliquée aux opérations à traiter, identique pour toutes (a) ou non (b), sur une période de 240 heures. Avec une flexibilité élevée associée au workload, il est possible d'atteindre un niveau donné de qualité de service avec moins de puissance de calcul. Par exemple (Figure 1a), pour atteindre 100% d'opérations traitées, il faut 10 TFlops si les deadlines sont à 200% de l'exécution réelle, 16 TFlops pour 100% et 52 TFlops pour 20%. Si les deadlines sont fixes (Figure 1b), avec une puissance de calcul de 45 TFlops et une deadline de 1000s, on peut traiter 40% des opérations. Cette valeur tombe à 30% pour une durée de 500s. Nous avons également noté que, sans flexibilité (deadlines à 0), il faudrait une puissance de calcul de 176 000 TFlops. Nous voyons donc que la flexibilité a un impact fort sur la qualité de service et peut être un levier efficace pour réduire le besoin de puissance de calcul et à fortiori le besoin de puissance électrique. À noter cependant que la flexibilité joue directement sur la satisfaction des utilisateurs, ce qui peut être un frein. Une solution pourrait alors être d'avoir des classes de jobs avec des flexibilités variables.



(a) (b)

FIG. 1 – Qualité de service en fonction de la puissance de calcul et de la flexibilité des deadlines. (a) : Flexibilité non-identique. (b) : Flexibilité identique (en secondes).

Références

[1] L.-C. Canon, D. Landré, L. Philippe, J.-M. Pierson, and P. Renaud-Goud. Assessing power needs to run a workload with quality of service on green datacenters. In *European Conference on Parallel Processing*, pages 229–242. Springer, 2023.

[2] D. Klusáček, Š. Tóth, and G. Podolníková. Real-life experience with major reconfiguration of job scheduling system. In *Job Scheduling Strategies for Parallel Processing : JSSPP 2016, Chicago, USA*, pages 83–101. Springer, 2017.

[3] E. Masanet, A. Shehabi, N. Lei, S. Smith, and J. Koomey. Recalibrating global data center energy-use estimates. *Science*, 367(6481) :984–986, 2020.

[4] J.-M. Pierson et al. Datazero : Datacenter with zero emission and robust management using renewable energy. *IEEE Access*, 7 :103209–103230, 2019.

[5] A. Radovanović et al. Carbon-aware computing for datacenters. *IEEE Transactions on Power Systems*, 38(2) :1270–1280, 2022.