ThERMOS :

Towards an Energy-awaRe micrOService-oriented orchestrator

(Vers un orchestrateur de micro-services orienté énergie).

Laboratoire : LS2N

Début : 01/09/2017

Financement : 50% IMT-A

Cofinancement : 50% sur fonds propres

Encadrement :

Pr Menaud Jean-Marc, LS2N, menaud@imt-atlantique.fr

Pr Südholt Mario, LS2N, mario.sudholt@imt-atlantique.fr

Mots clés en français:

Energies renouvelables, Cloud, micro-services, edge computing.

Mots clés en anglais :

Renewable Energy, Cloud, micro-services, edge computing.

Contexte

La thèse s'inscrit dans le contexte général du développement de la thématique « maîtrise énergétique des centres de données » développée à IMT-Atlantique depuis 2006 et fait écho à l'infrastructure CPER SeDuCe : un micro-datacenter alimenté par des énergies renouvelables. Cette thèse s'inscrit également dans les axes de recherche de l'équipe Ascola (puis Stack) IMT-A/LS2N/INRIA tant sur les aspects énergies que composants logiciels (micro-services), ainsi que dans le thème transverse GreenIT du LS2N. Enfin, ce sujet de thèse est une suite naturelle des recherches menées dans les projets CominLabs EPOC, FSN Hosanna, I&R EcoCLoud et du Carnot Cheddar.

Nul besoin de rappeler que les centres de données (hébergeant les services d'Internet) consomment de nos jours près de 2% de la production d'électricité mondiale [6], que leurs nombres et leurs tailles

sont en forte croissance ces dernières années et que cette croissance, tirée par l'arrivée massive de l'internet des objets, devrait continuer à un rythme important. A titre d'exemple, [1] en 2015 les centres de données ont consommés 416 térawatts/h, à comparer à la consommation électrique de la Grande-Bretagne qui est de l'ordre de 300 térawatts/h.

Tant pour des aspects économiques qu'environnementaux, les GAFA (Google, Apple, Facebook, Amazon) s'intéressent de plus en plus à alimenter leurs centres de données par des énergies renouvelables. Google a ainsi annoncé fin 2016 [2] que leurs objectifs 2017 seront d'utiliser pour leurs centres de données 100% d'énergies renouvelables. Cependant, la production électrique des énergies renouvelables étant fluctuante et les besoins en calcul constants, le choix a été fait par Google de contracter directement avec des fournisseurs, leur garantissant ainsi la disponibilité énergétique dont ils ont besoin et leur évitant ainsi d'avoir à gérer en interne cette fluctuation. En effet, cette gestion reste complexe du fait de l'intermittence de la production énergétique des sources d'énergies renouvelables [3, 4].

D'un point de vue scientifique, les solutions actuellement proposées se basent sur l'approche « follow the (sun, wind, tide etc…) ». Ces approches se basent sur deux piliers d'une part sur une infrastructure matérielle distribuées sur un large territoire (des micro-centres de données) alimentées pour toute ou partie par des énergies renouvelables et d'autres part par des applications pouvant se déplacer d'un micro-centre à un autre en fonction de critères techniques et de disponibilités énergétiques [5]. Cependant la migration applicative reste une opération difficilement exploitable à large échelle. En effet, migrer le contenu d'un serveur vers un autre sur un réseau 10Gb/s ne peut se réaliser en moins d'une dizaine d'heure. Or, le cycle de fluctuation énergétique et les modèles de prédictions associés, sont plus proches de l'heure que de la demi-journée.

Si la migration applicative n'est pas viable, une seule solution alternative consiste à distribuer intelligemment l'ensemble des applications sur un ensemble choisi de micro data-center permettant de couvrir leurs besoins énergétiques. Faire évoluer les systèmes et applications dans ce sens nécessite des méthodes et des techniques flexibles d'adaptation logicielle. Dans cette thèse, nous exploiterons la notion récente des « microservices » [7, 8] pour permettre une structuration en composants legers soutenant les adaptations nécessaires.

Objectifs

Le sujet de thèse proposé vise a étudier cette voie alternative. De nombreux défis sont à relever :

- Répliquer une application « cloudifiée » sur un ensemble de serveurs implique une augmentation importante des ressources consommées (énergie, processeur) et mobilisées (mémoire, disque). Proposer des solutions pour limiter cette surconsommation est indispensable.

- Sélectionner les micro data-centers sur lesquels déployer les réplicas est un problème de placement NP-Complet multi-objectif devant prendre en compte les temps d'accès, la disponibilité énergétique et les profils d'utilisation de l'application. Développer une heuristique efficace est nécessaire à la résolution de ce problème.

- Une fois déployé, sélectionner sous de multiples critères (énergétique, qualité de service…) le bon réplica à solliciter par requête de manière distribuée, reste une question scientifique ouverte. Un mécanisme a bus de message et broker de service peuvent constituer une solution intéressante à étudier.

Pour répondre à l'ensemble des défis, nous proposons dans ce sujet de thèse d'étudier l'intérêt de combiner une architecture à base de microservice, un système de placement distribué orienté énergie et un système autonomique de découverte de services et de répartition des requêtes.

Les microservices [7, 8] sont un style d'architecture logicielle à partir duquel une application complexe est décomposée en plusieurs processus indépendants et faiblement couplés, souvent spécialisés dans une seule tâche. Ces micro-services, embarqués dans des technologies Cloud (Container, VM), permettraient de limiter fortement l'augmentation des ressources consommées et mobilisées pour l'hébergement des réplicas. Les question scientifiques abordées dans cette décomposition sont à la fois d'ordre du Génie Logiciel (architecture applicative) et des systèmes (impact énergétique d'une telle architecture). Le système de placement à définir hériterait de nos travaux sur le placement de machines virtuelles sur des data-center mono site développé dans les projets Hosanna et EPOC. Le principal défi reste la modélisation du problème et notamment la prise en compte des profils de production énergétique des ENR couplés aux micro-data-centers répartis sur le territoire et des profiles de consommations des applications. Enfin, la répartition des requêtes utilisateur sur les réplicas sera effectué par un broker distribué et générique pour permettre une sélection intelligentes des réplicas à utiliser pour fournir le service.

Compétences requises

Etudiant Ingénieur ou M2 Recherche avec de bonnes compétences dans les domaines des systèmes distribués et du génie logiciel.

Compétences : Cloud computing, Monitoring, Architecture logicielle, Energie.

Références

[1] Tom Bawden. Global warming: Data cen- tres to consume three times as much energy in next decadde. January 2016.

[2] https://www.theguardian.com/environment/2016/dec/06/google-powered-100-renewable-energy-2017

[3] Towards energy-proportional Clouds partially powered by renewable energy , Menaud and all, Computing 2017.

[4] Goiri, Í., Katsak, W., Le, K., Nguyen, T. D., & Bianchini, R. (2014). Designing and managing data centers powered by renewable energy. IEEE Micro, 34(3), 8-16.

[5] Liu, Z., Lin, M., Wierman, A., Low, S. H., & Andrew, L. L. (2011, June). Greening geographical load balancing. In Proceedings of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems (pp. 233-244). ACM.

[6] Epa report to congress on server and data center energy efficiency. Technical report, Environmental Protection Agency, US Congress, 2007.

[7] J. Lewis, M. Fowler. Microservices – a definition of this new architectural term. https://martinfowler.com/articles/microservices.html

[8] P. Leitner, J. Cito, E. Stöckli. Modelling and Managing Deployment Costs of Microservice-Based Cloud Applications. IEEE/ACM 9th International Conference on Utility and Cloud Computing (UCC), 2016.