Utilisation du Cloud StratusLab : tests de performance des clusters virtuels.

Cécile Cavet (1), Maude Lejeune (2), Fabrice Dodu (3), Michèle Detournay (4)
(1) cecile.cavet@apc.univ-paris7.fr, APC, Univ Paris Diderot, CNRS/IN2P3, CEA/Irfu, Obs de Paris, Sorbonne Paris Cité, France
(2) maude.lejeune@apc.univ-paris7.fr, APC, Univ Paris Diderot, CNRS/IN2P3, CEA/Irfu, Obs de Paris, Sorbonne Paris Cité, France
(3) fabrice.dodu@apc.univ-paris7.fr, APC, Univ Paris Diderot, CNRS/IN2P3, CEA/Irfu, Obs de Paris, Sorbonne Paris Cité, France
(4) michele.detournay@apc.univ-paris7.fr, APC, Univ Paris Diderot, CNRS/IN2P3, CEA/Irfu, Obs de Paris, Sorbonne Paris Cité, France

Overview:

Cloud computing offers IT resources on demand. This recent infrastructure in computing framework recovers three services named IaaS (Infrastructure-as-a-Service), PaaS (Platform-as-a-Service) and SaaS (Software-as-a-Service). In this work, we have focused on the IaaS layer which allows to provide virtual machine, cluster and storage on demand.
We have tested the performance of the StratusLab Cloud (european research project) to determine advantages and disadvantages of Cloud computing for numerical simulation and data analysis. The academic StratusLab Cloud is a public open-source Cloud and it provides full Cloud solution: OpenNebula virtual infrastructure manager, computing resources and end-user client to instantiate and manage virtual machines.
We have used classical benchmarks usually performed on supercomputers and we have compared our results with previous studies made on public paying Cloud like Amazon EC2, GoGrid and IBM. We have run benchmarks on memory bandwidth, I/O access, MPI communications, scientific applications-like and weak scalability. Furthermore, we have tested file transfer abilities with IRODS (Integrated Rule-Oriented Data System) software and we have run a numerical code to study strong scalability by measuring speedup, efficiency and Karp-Flatt metric.
We have concluded that the virtual cluster of the StratusLab Cloud does not allow high performance in MPI involving tests. The slow inter-node communication is due to low virtual network i.e. virtualization layer overhead and Gigabit interconnection as shown in previous studies performed on equivalent paying Cloud. However, the non-MPI results of performed benchmarks are in agreement with classical cluster performances. Thus, the Cloud computing is an interesting solution for prototyping high performance computing applications and for running other types of scientific applications because it provides an easy and quick access to any kind of platform with reasonable resources and performances.

Enjeux scientifiques, besoin en calcul, stockage et visualisation :

Le Cloud computing offre des ressources informatiques à la demande (voir la définition complète du NIST [1]). Cette infrastructure récente vient s'ajouter aux offres de calcul usuelles que sont la grappe et la grille de calcul. Le Cloud computing recouvre trois services intitulés IaaS (« Infrastructure-as-a-Service », ressources de calcul), PaaS (« Platform-as-a-Service », applications pour le Web) and SaaS (« Software-as-a-Service », portails Web) ; il propose un accès public (tout utilisateur), privé (utilisateurs internes à une structure) ou hybride ; et il fournit un service gratuit (Cloud académique) ou payant. Il permet une grande flexibilité dans son utilisation et devient par conséquent une plateforme potentielle pour le calcul scientifique.
Dans cette étude, nous nous sommes intéressés au Cloud public IaaS qui permet de créer des machines, des clusters, du stockage et du réseau virtuel selon le besoin des utilisateurs. Nous avons évalué les performances de StratusLab [2] qui est un Cloud académique issu d'un projet européen débuté en 2010 et, plus généralement, nous avons déterminé les atouts du Cloud computing de type IaaS pour la simulation numérique et le traitement de données dans le contexte de la recherche en astroparticule.
Afin de déterminer les capacités du Cloud académique, nous avons généré un cluster virtuel en utilisant les ressources fournies par StratusLab et nous avons réalisé les tests de performance classiques qui permettent l'évaluation des calculateurs et des supercalculateurs. Avec cette méthode, nous avons testé la bande-passante de la mémoire, l'accès aux entrées/sorties, les communications MPI et des applications de type calcul de haute performance (« High Performance Computing », HPC) incluant l'extensibilité faible. De plus, nous avons testé le transfert de fichiers avec le logiciel IRODS (« Integrated Rule-Oriented Data System », [3]) et nous avons réalisé des simulations numériques à l'aide d'un code de calcul afin d'étudier l'extensibilité forte. Cette dernière étude nous a permis d'utiliser les concepts d'accélération et d'efficacité afin de quantifier la parallélisation et les communications MPI du code de calcul utilisé.
En comparant les résultats obtenus avec le cluster virtuel du Cloud StratusLab avec ceux obtenus sur un cluster classique et en procédant de même avec les résultats obtenus par des études précédentes effectuées sur des Clouds payants, nous avons pu déterminer les atouts et les inconvénients du Cloud computing pour la recherche scientifique et, plus particulièrement, pour les applications astroparticules.

Développements, utilisation des infrastructures :

StratusLab [2] est une solution de Cloud complète. En effet, ce logiciel open-source est constitué du gestionnaire d'infrastructure virtuelle OpenNebula ; du client StratusLab permettant aux utilisateurs de lancer, de gérer et d'utiliser les machines virtuelles ; et de ressources de calcul qui sont virtualisées via l'hyperviseur KVM. Pour cette étude, nous avons utilisé essentiellement les serveurs de calcul du Laboratoire d'Accélération Linéaire (LAL) sur lesquelles fonctionnent en partie le Cloud StratusLab. (les autres machines étant hébergées par GRNet en Grèce). L'ensemble de ces machines forme une grappe de 10 nœuds de calcul, chaque noeud possédant 24 cœurs et 36 GB de RAM. La connexion Ethernet inter-nœuds est à 1 GbE/s (*).
Nous avons créé des images disque de la distribution Scientific Linux 6.2 à l'aide du logiciel VirtualBox. Ces images disque contextualisées pour StratusLab sont disponibles sur le site de référencement des images de systèmes d'exploitation, le Marketplace [4] (l'image utilisée pour cette étude a pour identifiant HqcwtHmPvMtNZbIYJjoEdpJ8F05).
Afin d'évaluer les résultats obtenus sur StratusLab, nous avons reproduit les mêmes tests sur le Cluster Arago [5]. Il est installé au Centre François Arago (FACe) et est en phase de production depuis mai 2012. Il est constitué de 11 nœuds de calcul, chaque nœud possédant 16 cœurs et 48 GB de RAM. La connexion Ethernet inter-nœuds est à 10 GbE/s. Le système de gestion des fichiers est GlusterFS et le gestionnaire de soumission de job est Torque/Maui.
Dans le but de pouvoir comparer les performances des deux types de clusters, les clusters virtuels ont été configurés, dans la mesure du possible, en s'approchant des caractéristiques du cluster Arago. Ainsi, pour l'ensemble des tests, le cluster virtuel est constitué d'un nombre de nœuds de calcul variables (au maximum 8 nœuds), chaque nœud possédant 8 cœurs et 8 GB de RAM. Cependant, le système de gestion des fichiers du cluster virtuel est NFS. Mais pour les types de test que nous avons effectué sur les deux clusters, la différence entre NFS et GlusterFS n'est pas significative. Les librairies OpenMPI et ATLAS sont installées afin de pouvoir utiliser MPI et BLAS respectivement. Nous n'avons pas installé de gestionnaire de soumission de job étant donné que le cluster virtuel était réservé à un seul utilisateur.

(*) Il est nécessaire de préciser que le principe du Cloud étant une grande quantité de ressources sans contraintes matérielles, les caractéristiques techniques des machines ne sont en générales pas indiquées. Nous sommes donc reconnaissant aux administrateurs de StratusLab de nous les avoir fournies.

Outils, le cas échéant complémentarité des ressources, difficultés rencontrées :

Afin de déterminer les performances de StratusLab, nous avons analysé les résultats d'études précédentes réalisées sur les Clouds payants suivants (*) :

- L'Amazon Web Services [7] offre des machines virtuelles à la demande, le Cloud Elastic Compute (EC2), et du stockage, le Simple Storage Service (S3). L'EC2 permet de générer des instances dont les caractéristiques sont prédéfinies (« small », « large » et « extra large »). Cela a pour conséquence que, contrairement à StratusLab, il n'est pas possible de choisir exactement le nombre de cœurs, et la taille de la RAM et du disque du cluster virtuel. De plus, dans les grappes d'instances la connexion Ethernet est à 1 GbE/s. Mais Amazon propose depuis fin 2011 une nouvelle instance (Cluster Compute Eight Extra Large) qui inclue une connexion 10 GbE/s destinée aux applications HPC.
- GoGrid [8] permet de la même manière qu'Amazon de lancer des instances prédéfinies et il propose aussi une connexion à 1 GbE/s.
- IBM [9] offre des instances similaires aux Clouds précédents mais il propose une connexion plus faible à 100 MbE/s.

Nous avons utilisé différentes mesures qui permettent d'évaluer l'extensibilité (« scalability ») d'une application. Mais il faut d'abord discerner deux types d'extensibilité : l'extensibilité forte et faible. Leurs définitions sont les suivantes :

- L'extensibilité forte fait référence à la mesure de la diminution du temps de calcul pour un problème dont la taille totale est fixée. Dans cette approche, il est possible de quantifier la parallélisation et les communications MPI en mesurant l'accélération (rapport entre le temps de calcul séquentiel et parallèle) et l'efficacité (métrique classique et de Karp-Flatt, normalisation plus ou moins complexe de l'accélération par le nombre de cœurs de calcul utilisés).
- L'extensibilité faible fait référence à la même mesure mais pour un problème dont la taille par processeur est fixée. Dans ce cas la performance mesurée est comparée à une valeur théorique liée directement aux caractéristiques du matériel.

Nous avons utilisé ces concepts dans certains cas d'étude (voir ci-dessous le code de calcul Ramses et le test NBP).

(*) Les connexions Ethernet décrites ici sont celles déterminées par He et al. en 2010 [6]. Elles peuvent avoir évolué depuis comme dans le cas du Cloud EC2.

Résultats scientifiques :

Afin de déterminer les performances du Cloud StratusLab, nous avons réalisé les tests classiques suivants :

- **Bande-passante de la mémoire :**

Nous avons utilisé le code STREAM afin d'évaluer la bande-passante de la mémoire. En effet, le test réalise quatre opérations sur des grandes quantités de données stockées dans la mémoire. En prenant une taille de tableau de travail de 500 MB et en effectuant les opérations 1000 fois (*), nous obtenons, pour la version séquentielle et parallèle OpenMP du test, un comportement similaire pour les deux types de cluster : la bande-passante de la mémoire en fonction du nombre de processus (« threads ») présente la même croissance jusqu'à ~22 GB/s ce qui correspond à 4 processus et ensuite la bande-passante sature pour les deux clusters car leurs caches ont probablement atteint leurs capacités maximales.

- **Accès aux entrées/sorties :**

Nous avons étudié ensuite l'accès aux entrées/sorties en utilisant IOzone. Ce test détermine le temps nécessaire à l'écriture et la lecture de fichiers de petites et grandes tailles. Nous avons testé ces opérations dans le répertoire /home/user porté par NFS (**). L'écriture a un débit moyen de ~900 MB/s pour les petits fichiers (de 64 KB à 2 GB) et ~300 MB/s pour les grands fichiers (de 2 GB à 4 GB) et la lecture qui présente moins de variations entre les deux catégories de fichier est à ~3GB/s. Ces résultats sont comparables aux performances de GlusterFS sur le cluster Arago et supérieurs aux ~50 MB/s (fichier de 1 MB) trouvés par Evangelinos et Hill [10] en 2008 sur EC2.

- **Communications MPI :**

Avec le test Intel MPI Benchmark (IMB), nous avons pu vérifier les communications MPI. Nous avons réalisé l'opération « Send/Receive » entre deux machines virtuelles (**). Ce test envoie des chaînes de paquets de données de plus en plus volumineux (de 1 KB à 4 MB). Nous avons constaté que la bande-passante du cluster virtuel (40 MB/s) est ~16 fois moins importante que sur le cluster Arago ce qui correspond au rapport entre les connexions Ethernet des deux clusters et au réseau virtuel (voir le test NPB ci-dessous). Hill et Humphrey [11] ont montré en 2009 qu'EC2 présente un taux de 90 à 40 MB/s dépendant du type de l'instance et ainsi StratusLab ne présente pas de différence avec EC2 dans ce cas.

- **Calcul haute performance :**

Nous avons ensuite réalisé deux tests qui sont utilisés dans la branche du calcul haute performance (HPC) pour évaluer les plateformes de calcul : le NASA Parallel Benchmark (NPB) et le High Performance LINPACK (HPL).

Le NPB effectue 8 opérations de type applications scientifiques (voir Figure 1). Nous avons utilisé la version MPI du test afin d'étudier en même temps les communications inter-noeuds. En comparant la moyenne des temps de calcul (**) avec les résultats de He et al. (2010) [6], nous déterminons que le StratusLab est quasiment aussi performant qu'EC2 et GoGrid (connexion à 1 GbE/s) et meilleur qu'IBM (connexion à 100 MbE/s). Le cluster Arago présente des résultats bien meilleurs révélant l'importance d'une connexion à 10 GbE/s et la dégradation des performances due à la couche de virtualisation.

Le test HPL est utilisé en HPC pour classer les supercalculateurs mondiaux (Top 500). Il résout des systèmes denses d'équations linéaires. Ce test fait intervenir l'extensibilité faible (voir ci-dessus) car la dimension du système à résoudre augmente avec le nombre de nœuds mais la taille des blocs résolus par les cœurs est fixe. Les conditions initiales du test sont déterminées avec un outil analytique [12] et la performance au pic théorique ($R_{pic,\ théo}$) est établie avec HPL-calculator [13]. Dans le cluster virtuel (**), un nœud est constitué de 8 cœurs à 2,67 GHz et 4 opérations à virgule flottante par cycle d'horloge et donc $R_{Rpic,\ théo}$ = 71 GFLOPS. Nous mesurons pour 1 nœud une efficacité de 45 % et cette efficacité décroit quand le nombre de nœuds augmente (voir Figure 2). Ces valeurs sont un peu plus faibles que pour EC2 et GoGrid comme montré dans l'étude de Iosup et al. (2011) [14]. Mais dans tous les cas, les Clouds restent en dessous des 60 % d'efficacité nécessaires à l'obtention de bonnes performances pour les applications scientifiques.

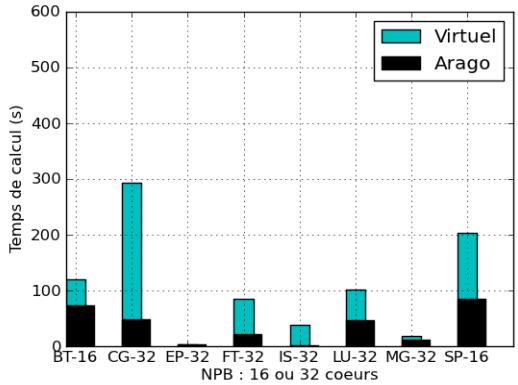

Figure 1 : test NPB sur les clusters virtuel et Arago.
Temps de calcul en fonction du nombre de cœurs de calcul.

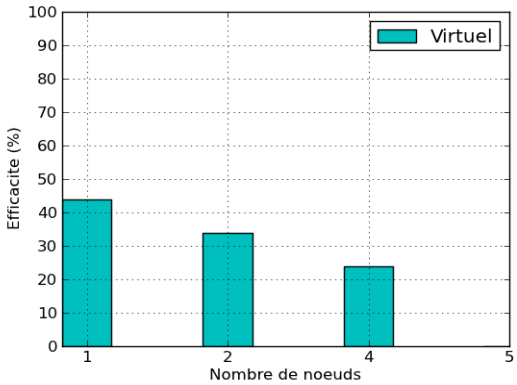

Figure 2 : test HPL sur le cluster virtuel.
Efficacité en fonction du nombre de nœuds de calcul.

- **Transfert de fichier :**

Nous avons aussi utilisé le logiciel IRODS [3] qui est un système de gestion de données distribuées afin de déterminer les contraintes du transfert de gros volumes de données entre le cluster virtuel et un serveur de stockage extérieur. Nous avons transféré 123 GB de données depuis une zone IRODS située au CC-IN2P3 vers StratusLab et le Cluster Arago. Sur le cluster classique, le transfert est ~100 MB/s ce qui est ~4 fois plus performant que sur le cluster virtuel.

- **Extensibilité forte :**

Nous avons finalement exécuté sur StratusLab des simulations numériques avec le code Ramses [15] qui est utilisé en cosmologie pour étudier la formation des grandes structures. Ce code numérique est parallélisé avec MPI et implémente une grille cartésienne. Nous avons réalisé une étude d'extensibilité forte (voir ci-dessus) afin de voir la réponse du Cloud à une véritable application scientifique. Dans ce but, nous avons avons utilisé 8 nœuds de calcul et les caractéristiques usuelles. L'accélération (voir Figure 3) est quasiment linéaire jusqu'à ~10 cœurs pour le cluster virtuel et ~20 cœurs pour le cluster Arago. Elle atteint son maximum pour ~64 cœurs et ~80 cœurs respectivement donnant la limite du nombre de cœurs de calcul à utiliser sur ces infrastructures. L'efficacité selon la métrique classique n'est pas très pertinente dans ce cas et celle de Karp-Flatt (voir Figure 4), qui reste faible pour un nombre de cœur inférieur aux maxima, nous informe que la partie séquentielle du code n'est pas responsable d'une limite dans le nombre de cœurs. Ainsi l'utilisation en phase de production (c'est-à-dire afin de maximiser la diminution de temps de calcul) d'un code numérique parallélisé MPI reste encore réservée aux supercalculateurs.

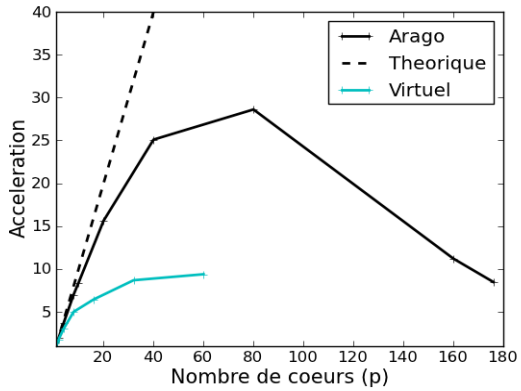

Figure 3 : simulations numériques avec Ramses sur les clusters virtuel et Arago.
Accélération en fonction du nombre de cœurs de calcul.

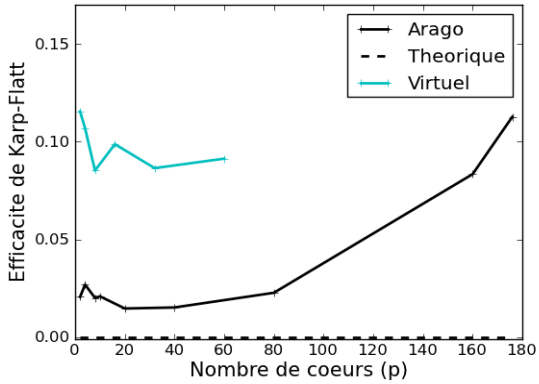

Figure 4 : simulations numériques avec Ramses sur les clusters virtuel et Arago.
Efficacité de Karp-Flatt en fonction du nombre de cœurs de calcul.

(*) Pour ce test, nous avons lancé une seule machine virtuelle avec 16 cœurs et 32 GB de RAM.
(**) Pour ce test, nous avons lancé un cluster virtuel avec 4 nœuds et les caractéristiques usuelles.

Perspectives :

Nous avons pu voir qu'une partie des tests classiques que nous avons effectués présentent de bonnes performances sur le Cloud StratusLab. En effet, la bande-passante de la mémoire virtuelle et l'accès aux entrées/sorties ont produit des taux élevés. En ce qui concerne les tests et les applications impliquant des communications MPI, nous avons vu que les résultats sont moins performants que sur les calculateurs classiques ce qui révèle l'importance d'une connexion Ethernet à 10 GB/s et que la virtualisation des ressources dégrade les performances et cela est sensiblement visible au niveau du réseau. Ainsi le Cloud Computing de type IaaS est une solution intéressante pour prototyper des applications de calcul haute performance et pour évaluer différentes applications scientifiques car il offre un accès facile et et rapide à des ressources dimensionnables en fonction des besoins avec des performances raisonnables.
Dans la continuité de cette étude, nous aimerions tester d'autres plateformes de Cloud académique. Mais nous avons pu constater qu'il n'existe que très peu de Cloud IaaS publics et gratuits qui permettent le même type de possibilités qu'offre le Cloud StratusLab. En effet, il existe le projet américain Science Clouds [16] qui propose une solution de Cloud via la plateforme Future Grid [17] mais l'accès y est plus restreint. Le CC-IN2P3 propose depuis récemment un Cloud IaaS privé qui utilise le gestionnaire d'infrastructure virtuelle OpenStack et qu'il sera donc intéressant de tester.
Dans un autre cadre, nous avons aussi commencé à porter des applications astroparticules sur le Cloud StratusLab. Les résultats préliminaires montrent que le Cloud Computing est bien adapté, particulièrement dans les applications nécessitant beaucoup de mémoire. Le retour de différents types de projets scientifiques permettra d'avoir une idée globale des possibilités qu'offre le Cloud IaaS pour la recherche en astroparticule.

Références :

[1] Définition du NIST : http://www.nist.gov/itl/cloud/
[2] Cloud StratusLab : http://stratuslab.eu/index.php
[3] IRODS : https://www.irods.org/
[4] Marketplace : http://marketplace.stratuslab.eu/metadata
[5] Cluster Arago : https://www.apc.univ-paris7.fr/FACeWiki/pmwiki.php?n=Face-cluster.Face-cluster
[6] Q. He, S. Zhou, B. Kobler, D. Duffy, T. McGlynn, Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing (2010)
[7] Amazon Web Services LLC, Amazon Elastic Compute Cloud (Amazon EC2) : http://aws.amazon.com/ec2
[8] Cloud Gogrid : http://www.gogrid.com
[9] Cloud IBM : http://www-935.ibm.com/services/fr/gts/cloud/index.html
[10] C. Evangelinos, C.N. Hill, Proc. 1st Workshop on Cloud Computing and Its Applications, CCA'08, Chicago, IL, USA, pp. 1–6 (2008)
[11] Hill et Humphrey, Proceedings of the 10th IEEE/ ACM International Conference on Grid Computing (Grid 2009). Oct 13-15 2009. Banff, Alberta, Canada (2009)
[12] Advanced Clustering : http://www.advancedclustering.com/faq/how-do-i-tune-my-hpldat-file.html
[13] Hpl-caculator : http://hpl-calculator.sourceforge.net
[14] A. Iosup, S. Ostermann, N. Yigitbasi, R. Prodan, T. Fahringer, D. Epema, IEEE Transactions on Parallel and Distributed Systems 22, 931–945 (2011)
[15] Code Ramses : http://irfu.cea.fr/Phocea/Vie_des_labos/Ast/ast_sstechnique.php?id_ast=904
[16] Science Clouds : http://scienceclouds.org/
[17] Future Grid : https://portal.futuregrid.org/