Chapitre III : Allocation de ressources dans les systèmes Cloud computing

Introduction

L'allocation des ressources est un sujet qui a été développé dans de nombreux domaines informatiques, tels que les IoT, l'Edge computing et le Cloud computing.

Le but de l'Edge et le Cloud computing est de réunir les ressource non-utilisé pour réaliser des pools de ressources partagés, rendre l'accès à la demande des utilisateurs aux ressources commodément, améliorer l'utilisation des ressources, etc. Cependant, les ressources disponibles des fournisseurs et les exigences des consommateurs en ressources sont à la fois variées dynamiquement. Par conséquent, l'un des principaux enjeux est définir une méthode d'allocation des ressources pour le traitement des données au bord et/ou au niveau de Cloud computing.

III.1°) Définition et Concepts

L'évolution rapide d'applications des infrastructures informatiques (edge et Cloud) repose sur des technologies clés dont la technologie de base est la gestion des ressources.

« Les ressources peuvent être défini d'une manière général comme tout matériel ou immatériel, les objets à usage restreint, ou de toute aide à maintenir les moyens de subsistance des objets » **[17]**.

Les Ressources de Cloud et du edge computing peuvent être vus comme n'importe quelle ressource (physique ou virtuel) que les utilisateurs peuvent bénéficier suivant un ensemble technique mis en place. En général, les ressources sont situées dans un centre des données qui sont partagés par plusieurs clients, et doivent être attribués aux utilisateurs et ajustés dynamiquement en fonction d'une sollicitation.

❖ Classification des ressources dans le Cloud computing

Les termes ressources incluent la fourniture d'un centre de données, matériel, logiciels, applications des fournisseurs d'infrastructure pour traiter et stocker des données sur Internet. Le concept principal du Cloud est de fournir des ressources à travers des services pour les personnes publiques ou privées ou les gens d'affaires utilisant les nuages. Ainsi de nombreux types de ressources sont fournis aux utilisateurs par Cloud pour l'exécution de leurs applications. Parmi ces types de ressources nous avons :

➢ Ressources informatiques

Elles sont constituées par la collecte de mémoire, réseau, Processeur, périphériques d'entrée / sortie dans les infrastructures informatique (Edge et Cloud computing) ; celles-ci sont appelées machines physique(PM).Ainsi, les ressources informatiques sont successibles d'être allouées ou achetées par les utilisateurs selon leurs besoins .Cependant, Le concept de VM(Virtual machine) a révolu le PM où cette

dernière crée un logiciel virtuel sur lequel l'utilisateur peut s'exécuter machine virtuelle dans différents systèmes d'exploitation, applications et plates-formes.

➢ Ressources de réseautage

Le réseautage est constitué de différentes types de ressources, parmi lesquelles nous avons : la bande passante, le stockage, la communication, le trafic, etc.…

Dans le domaine du réseau, on constate très souvent des problèmes (manquements ou insuffisances) liés à ces ressources, ce qui nous oblige à implanter des protocoles pour améliorer la qualité des plateformes de nuages.

➢ Ressources énergétiques

Dans le Cloud, la consommation d'énergie est l'un des facteurs le plus important pour pouvoir satisfaire la demande de ressources des utilisateurs. L'énergie consommée par le système pour fournir et allouer de ressources est beaucoup moins que l'énergie consommée par le système qui est inactif, dans l'attente d'une ressource alloué. Cela a conduit à une autre technologie, à savoir le nuage vert d'informatique.

❖ Classification des ressources dans l'Edge computing

La première étape dans l'évaluation des avantages d'une solution de pointe consiste à décider quels types de ressources peuvent être gérés mieux comparé à un système centralisé.

Une justification évidente de l'utilisation des architectures de bord est réduire le temps de réponse, ce qui peut être fait si des ressources de calcul et de communication sont fournies et utilisées adéquatement [20].

L'un des ressources la plus préoccupante est le stockage, il peut être bénéfique pour la sécurité ou la rapidité du fait de la personnalisation récupérer et sécuriser les mécanismes de stockage. Un type moins évident de la ressource est d'avoir accès à un type spécial de données (par ex. disponibilité de capteurs) offrant des avantages locaux dans une application. Des exemples sont l'utilisation de caméras ou de capteurs de localisation. La quantité et le type de données capturées affectent à leur tour les ressources de calcul et de communication (à quelle fréquence mélanger les données, quelle quantité traiter ou filtrer avant de mélanger), et implicitement le choix de l'endroit et de la quantité des autres ressources à déployer.

Une autre catégorie que nous considérons est l'énergie en tant que ressource, qui est clairement influencée par la quantité de calcul, communication, stockage et capture de données. Enfin, certains travaux considèrent les ressources dans une manière générique en utilisant des termes abstraits tels que «ressource virtuelle» «Valeur» ou simplement en tant qu'éléments sans unité dans un modèle.

III.2°) Gestion des ressources dans les systèmes Cloud Computing

Il existe de nombreuses types ressources dans les plateformes informatique de nuage (Cloud et edge) tel que : la mémoire, le processeur, l'espace de stockage, la bande passante et les équipements informatiques.ces ressources sont allouées aux consommateurs pour satisfaire leurs besoins en terme de ressources. Cependant une bonne gestion des ressources est nécessaire pour qu'elles puissent être utilisées d'une manière fiable et optimale par les utilisateurs.

Ainsi de nombreuses méthodes sont élaborées pour mieux gérer les ressources :

➢ La mutualisation des ressources

C'est la pratique qui consiste à partager l'utilisation d'un ensemble de ressources par des consommateurs (ou entités quelconques) n'ayant aucun lien entre eux. Les ressources peuvent être de diverses natures : logicielles ou matérielles (machines, équipements réseau, énergie électrique). Cette pratique dépend du désir des entreprises de délocaliser leurs services informatiques vers des infrastructures informatiques **[4].**

Ces infrastructures informatiques doivent neutraliser les problèmes (sécurité, la disponibilité, l'intégrité, la fiabilité et l'uniformité d'accès aux données) liés à leurs exploitations et utilisations.

➢ La virtualisation des ressources

La virtualisation est un mécanisme qui tente de masquer les caractéristiques physiques d'une ressource informatique de manière à simplifier les interactions entre cette ressource et d'autres systèmes, d'autres applications et les utilisateurs ; permet également d'obtenir une ressource physique comme plusieurs ressources logiques et, inversement, de percevoir plusieurs ressources physiques comme une seule ressource logique. Il offre une vue logique plutôt que physique, de la puissance de calcul, de la capacité de stockage, et des autres ressources informatiques et permet aussi de diminuer le gaspillage des ressources.

La virtualisation repose sur trois éléments importants :

- ✓ L'abstraction des ressources informatiques ;
- ✓ La répartition des ressources par l'intermédiaire de différents outils, de manière à ce que celles-ci puissent être utilisées par plusieurs environnements virtuels ;
- ✓ La création d'environnements virtuels.

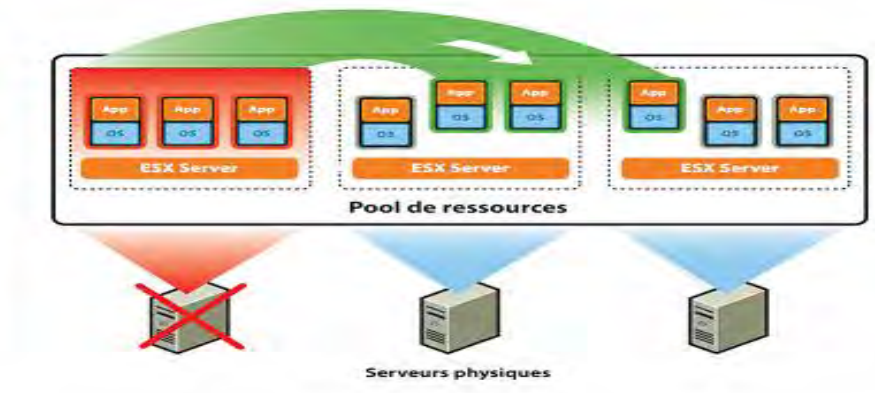


*Figure 9 Architecture de virtualisation des ressources **[18]***

III.2.1°) Allocation des ressources dans les systèmes Cloud computing et Edge Computing

De plus en plus, on note une croissance importante de la popularité des systèmes qui offrent des ressources informatique à la demande, basé sur la facturation à l'usage et l'équité des ressources partagées selon la demande des utilisateur pour que ces derniers puissent ajuster (augmenter ou diminuer) leur taux de consommation de ressources en fonction de leurs besoins. Ainsi ces systèmes peuvent supporter plusieurs consommateurs simultanément sur les mêmes infrastructures.

Figure 10 Module de gestion de Ressource [19]

Le composant de base de la gestion des ressources est le processus de découverte des ressources qui détermine les types de ressources disponibles appropriés selon les exigences du client.
Ce processus est géré par le fournisseur de services Cloud. Les informations complètes sur la disponibilité des ressources sont déterminées par la procédure de découverte des ressources.
La découverte de ressources offre une méthode permettant de déterminer l'état des ressources gérées. Elle fonctionne avec la distribution des ressources pour fournir des informations sur l'état des ressources au serveur.
La gestion des Ressources est un processus efficace et efficient qui gère les ressources ainsi fournie ET garanties le QoS aux utilisateurs Cloud tel que la haute disponibilité et le partage des ressources.
Elle permet de gérer les ressources physiques tels que les noyaux CPU, espace disque et la bande passante du réseau.

Il existe deux(2) principaux acteurs dans les systèmes Cloud computing : le fournisseur et l'utilisateur ou le consommateur.

D'une part, les fournisseurs détiennent des ressources informatiques massives dans leurs grands centres de données et les louent à des utilisateurs. D'autre part, il ya les utilisateurs qui ont des applications avec des charges variées, louent des ressources de la part des fournisseurs pour exécuter leurs applications. La **Figure 11** montre l'interaction entre les fournisseurs et les utilisateurs. Tout d'abord, l'utilisateur envoie une demande au fournisseur qui contient ces besoins en ressources. Lorsque ce dernier reçoit la demande, il cherche des ressources pour satisfaire la demande et alloue ces ressources à l'utilisateur demandeur, généralement sous forme des machines virtuelles (VM). Ensuite, l'utilisateur utilise les ressources assigné a lui pour exécuter leur applications et paie les ressources qui a utilisé. Lorsque l'utilisateur termine avec ces ressources, ils les retournent aux fournisseurs **[21].**

Cependant l'avantage majeur de l'edge et du Cloud est que les acteurs ont leur propre intérêt. C'est-à-dire les fournisseurs visent à maximiser que possible leurs revenus avec un investissement minimum autrement dit ils veulent maximiser l'utilisation de leurs ressources informatiques tandis que les utilisateurs veulent accomplir leur travail à un coût minimal ou, en d'autres termes, ils veulent maximiser leur performance économique et applicative.

Vue l'importance donné à la confidentialité du matériel ou des données dans les systèmes informatiques, l'allocation optimale des ressources devient de plus en plus difficile à réaliser (ex. les fournisseurs ne veulent pas exposer combien et quel genre de machines ils ont et comment ils sont connectés

au moment ou les utilisateurs ne veulent pas exposer les détails de leur charge de travail à d'autres personnes y compris les fournisseurs).

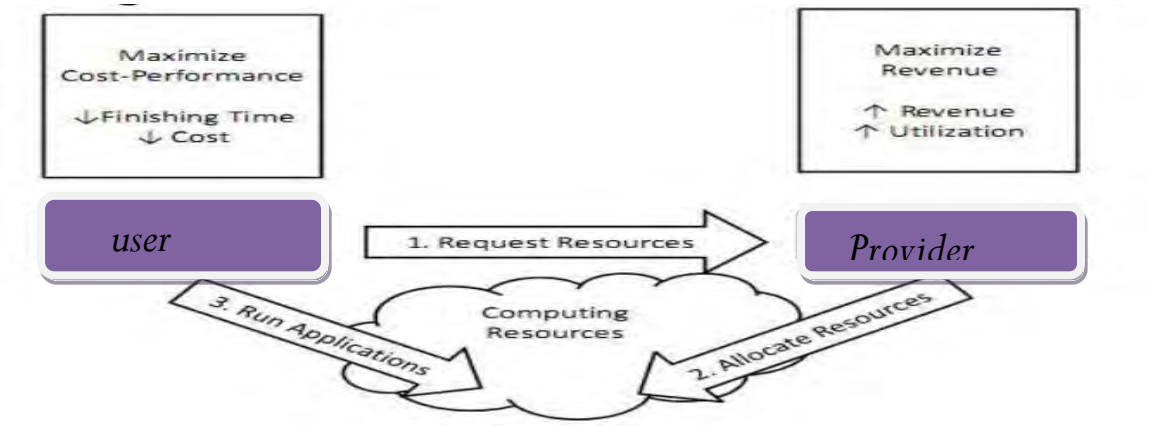


Figure 11 Scénario d'utilisation des ressources

Notre architecture ci-dessus représente les principaux concepts liés à l'allocation des ressources. Ainsi de nombreuses définitions ont été données à l'allocation de ressources :

"L'allocation des ressources est un élément important du nuage informatique. Son efficacité va influencer directement sur la performance de l'ensemble de l'environnement du nuage. Il requiert le type et la quantité de ressources nécessaires pour chaque application afin de terminer le travail de l'utilisateur**"[22]**.

"Dans le nuage informatique, l'allocation des ressources (RA) est le processus d'attribution des ressources disponibles pour les requises des applications d'Edge ou de Cloud via Internet. L'allocation des ressources affame des services si l'allocation n'est pas gérée avec précision"**[20].**

Allouer des ressources aux consommateurs du nuage informatique est une tâche énorme. Le provisioning des ressources du nuage se fait par tenu en compte des accords de niveau de service (SLA). Cependant une allocation efficace des ressources, devrait éviter les situations suivantes :

- **Sur-provisioning** : se pose lorsque le fournisseur alloue pour le consommateur des ressources plus que la demande.

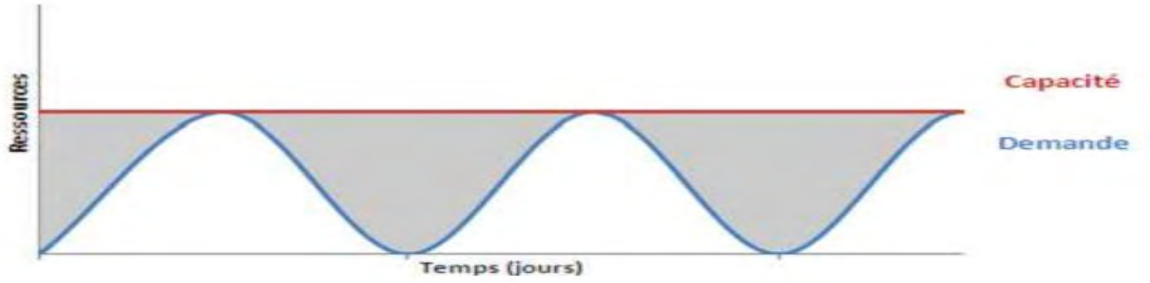


Figure 12 Besoins en ressources informatiques, surestimation [42]

Le premier graphique (Figure 12 ci-dessus) présente le cas d'un utilisateur (entreprise ou particulier) qui a les ressources informatiques pour absorber tous les pics. Les zones grisées montrent la part de budget perdue dû à une sous utilisation des ressources.

- ***Sous-provisioning*** *:* se produit lorsque l'allocation des ressources est moins que la demande.

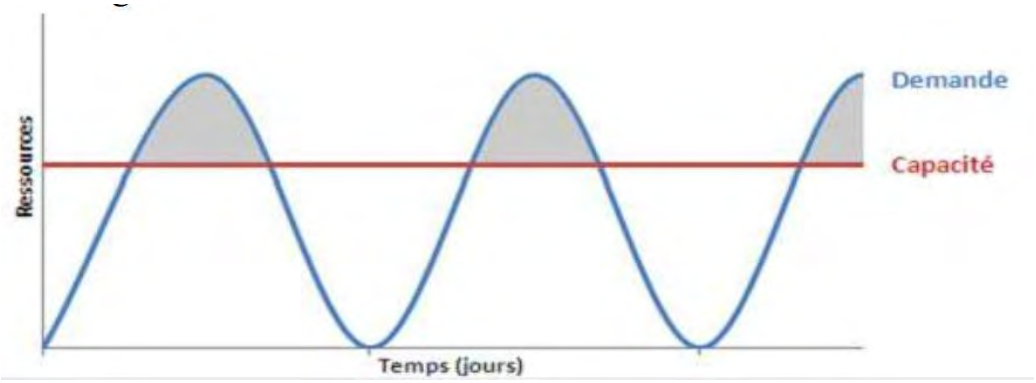


Figure 13 Besoins en ressources informatiques, sous-estimation [42]

Ce deuxième graphique (Figure13 ci-dessus), présente le cas d'un utilisateur (entreprise ou particulier) qui ne peut pas absorber tous les pics. Si ce dernier utilise ces ressources pour des prestations, il risque de perdre des clients, car ces derniers ne seront pas satisfaits de la prestation offerte. Les zones grisées montrent la part de temps où les capacités informatiques ont été sous-estimées. Cette zone est généralement difficile à estimer d'un point de vue technique, mais encore plus d'un point de vue budgétaire.

- **Situation conflictuelle de ressource** : se pose lorsque deux applications tentent d'accéder à la même ressource en même temps.
- **Le manque en ressources** : se pose lorsque les ressources sont limitées.
- **La fragmentation des ressources** : cette situation se pose lorsque les ressources sont isolées. (Il y'a suffisamment de ressources, mais l'allocation est impossible).

III.2.2°) Stratégies d'allocation des ressources (RAS)

Une stratégie d'allocation des ressources (RAS) peut être défini comme un mécanisme qui vise à garantir que les ressources physiques et ou virtuels sont correctement assignés aux utilisateurs.

Les paramètres d'entrée au serveur d'accès à distance et le mode de ressource d'allocation varient en fonction des services, de l'infrastructure et de la nature des applications qui exigent des ressources.

❖ Critères et classification de RAS

Le tableau et le diagramme ci-dessous décrit respectivement les critères du RAS et la classification des ressources.

Tableau 1 Les critères de RAS

Critères	Descriptions
Performance	la performance se définit par le plus petit temps de réponse pour l'exécution d'une tache résultants des applications aux utilisateurs
Disponibilité	La disponibilité désigne le ratio de temps pendant lequel le système est en état de fonctionner correctement sur une période de temps donnée, autrement dit, le fournisseur de service doit répondre au besoin des utilisateurs dés que ces derniers effectuent des demandes.
Fiabilité	La fiabilité désigne le processus de la demande et de la réception de ressources jusqu'à l'exécution sans rencontrer le moindre problème.
Temps de réponse	Un temps de réponse désigne la durée d'exécution d'une opération sur le système informatique. c'est un critère très important pour les applications interactives.
Débit	C'est une mesure de la charge instantanée supportée par le logiciel et son infrastructure. Ainsi le nombre d'applications exécutées par unité de temps devrait être élevé.
Sécurité	Le système doit être sécurisé pour les applications de traitement des transactions où la sécurité est considérée comme un critère important.

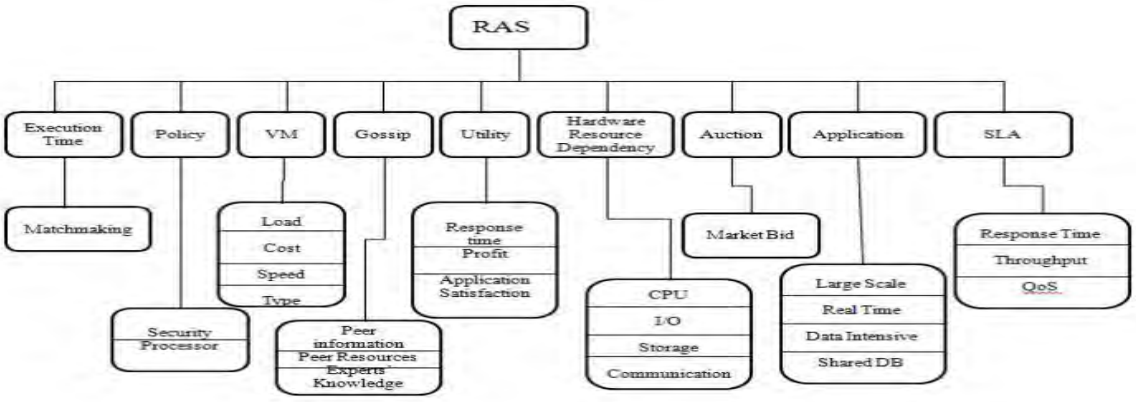


Figure 14 Resource Allocation Stratégies [23]

➢ Exécution Time

Différents types de mécanismes d'allocation de ressources sont proposé en nuage. Le temps d'exécution et la planification préemptable sont pris en compte pour l'allocation de ressources. Il surmonte le problème de la ressource contention et augmente l'utilisation des ressources en utilisant différents modes de location de capacités informatiques. Mais, estimer le temps d'exécution d'un travail est une tâche ardue pour un utilisateur et des erreurs sont commises. La stratégie d'allocation est basée sur les critères de type Any-Schedulability pour attribuer des travaux à des

ressources opaques de manière hétérogène. Ce travail n'utilise pas une connaissance détaillée des politiques de planification utilisées sur les ressources et est soumises à une AR (réservation anticipée).

➢ Policy

L'un des défis de l'affectation des ressources est la fragmentation en environnement multi-cluster est contrôlée par le travail qui utilisait la politique de processeur la plus adaptée pour l'allocation de ressources. La politique la plus adaptée alloue un travail au cluster, ce qui produit un reste distribution du processeur, conduisant au plus grand nombre d'attributions de travail ultérieures immédiates. Cela nécessite un processus de recherche complexe, impliquant l'activité d'allocation simulée, pour déterminer le cluster cible. Les clusters sont supposés être homogènes et réparties géographiquement. Le nombre de processeurs dans chaque cluster est compatible binaire. Des résultats expérimentaux montrent que la politique la plus adaptée aux complexités temporelles plus élevées, mais les frais généraux sont négligeables par rapport au fonctionnement à long terme du système. Cette politique est pratique à utiliser dans un système réel.

➢ Virtual Machine (VM)

C'est un système capable de redimensionner automatiquement son infrastructure de ressources. Le système composé d'un réseau virtuel et de machines virtuelles capables de migrer en direct à travers une infrastructure physique multi-domaines. En utilisant la disponibilité dynamique des ressources d'infrastructure et de demande dynamique d'application. Un environnement de calcul virtuel est capable de se déplacer automatiquement à travers l'infrastructure et mettre à l'échelle ses ressources. Mais le travail ci-dessus ne considère que la politique de planification non préemptable. Plusieurs chercheurs ont développé des ressources efficaces d'allocations pour les tâches en temps réel sur un système multiprocesseur. Les utilisateurs peuvent configurer et démarrer les ressources requises et ils ne doivent payer que pour les ressources nécessaires. Il est mis en œuvre en permettant aux utilisateurs d'ajouter et / ou de supprimer une ou plusieurs instances des ressources sur la base de chargement de la machine virtuelle et conditions spécifiées par l'utilisateur. Ce qui précède RAS mentionné sur IaaS diffère de RAS sur SaaS dans la plateforme d'allocation parce que SaaS fournit uniquement l'application à l'utilisateur.

➢ Gossip

L'environnement d'allocation de ressource diffère en termes de clusters, serveurs, les nœuds, leur référence de localité et leur capacité. Le protocole général Gossip est proposé pour une allocation équitable de la CPU aux clients des nuages. Le protocole implémente un schéma distribué qui alloue les ressources Cloud vers un ensemble d'applications qui ont une demande de mémoire indépendante du temps et maximise de manière dynamique une fonction d'utilité globale de l'infrastructure informatique. Des résultats ont montré que le protocole produit une allocation optimale lorsque la demande est inférieure à la mémoire disponible, et la qualité de l'attribution ne change pas avec le nombre d'applications et du nombre de machines. Mais ce travail nécessite des fonctionnalités supplémentaires pour l'allocation des ressources.

Ce système est robuste en cas de panne de la machine qui couvre plusieurs clusters et Datacenter.

➢ Utility Function

De nombreuses propositions gèrent de manière dynamique les ressources en IaaS en optimisant une fonction objective telle que minimiser la fonction de coût, la fonction de performance des coûts et atteindre les objectifs de QoS. La fonction objective est définie comme propriété utilitaire sélectionnée en fonction de mesures de temps de réponse, nombre de QoS, objectifs atteints et profit, etc.

Pour les systèmes informatiques en nuage à plusieurs niveaux (systèmes hétérogènes), l'allocation des ressources en fonction du temps de réponse, la mesure de la fonction d'utilité est proposée en considérant la CPU, mémoire et ressources de communication. Pour chaque niveau, les requêtes de l'application sont distribuées. Chaque serveur disponible est affecté à exactement l'un de ces niveaux d'application, à savoir le serveur peut uniquement servir les

demandes sur ce serveur spécifié. Chaque demande du client est envoyée au serveur à l'aide de la théorie de la mise en file d'attente et ce système répond aux exigences de SLA telles que temps de réponse et fonction utilitaire en fonction de son temps de réponse.

➢ Auction

Le mécanisme d'allocation des ressources en nuage par enchère est basé sur une enchère à offre scellée. Le service Cloud fournisseur collecte toutes les offres des utilisateurs et détermine le prix. La ressource est distribuée au k-ième premier enchérisseur sous le prix de la (k + 1) ème offre la plus élevée. Ce système simplifie l règle de décision du fournisseur de services et la règle d'allocation en réduisant le problème de ressources en ordre. Mais ce mécanisme ne garantit pas le profit de maximisation en raison de sa propriété de vérité sous contraintes. L'objectif de la stratégie d'allocation des ressources est de maximiser les bénéfices de l'agent client et de l'agent ressource dans un grand Datacenter en équilibrant la demande et l'offre dans le marché. Il est réalisé en utilisant des ressources basées sur la stratégie du marché d'allocation dans laquelle la théorie de l'équilibre est introduite.

❖ Organigramme d'affectation de ressource

Figure 15 Affectation des ressources [19]

➢ Request Identification:

C'est la première étape de l'affectation de ressources .Dans cette étape, diverses ressources seront identifiés par les fournisseurs de Cloud.

➢ Ressource Gathering

(Collecte de ressources / formation de ressources): après identification des ressources à l'étape 1, rassemblement ou la formation de ressources aura lieu. Cette étape identifiera les ressources disponibles. Elle peut également préparer à la personnalisation des ressources.

➢ Ressource Brokering:

Cette étape est la négociation de ressources avec les utilisateurs de Cloud pour s'assurer qu'ils sont disponibles selon les besoins.

➢ Ressource Discovery:

Cette étape permet logiquement de regrouper diverses ressources selon les exigences des clients.

- **Sélection des ressources**:

Cette étape consiste à choisir la meilleure ressource parmi les ressources disponibles pour les besoins fourni par les utilisateurs des systèmes Cloud.

- **Cartographie des ressources**:

Cette étape mappera les ressources avec des ressources physiques (comme le nœud, lien, etc.) fourni par les fournisseurs de Cloud.

- **Allocation de ressources**:

Cette étape allouera ou distribue des ressources aux utilisateurs des systèmes Cloud. Son objectif principal est de satisfaire les utilisateurs de leurs besoins et générer de revenus pour les fournisseurs Cloud.

III.2.3°) Techniques d'allocation de ressources dans le Cloud computing

L'allocation des ressources est un sujet qui a été abordé dans de nombreux domaines informatiques. Il repose sur un système de négociation efficace basé sur des contrats de service appelés Service Level Agreement ou SLA. En effet suite à la demande de ressources des utilisateurs au Cloud pour l'externalisation de leurs ressources, les fournisseurs doivent garantir certains de critère comme la qualité de service, la fiabilité, la disponibilité des ressources, la sécurité des données, etc. Cependant, les ressources disponibles des fournisseurs et les exigences des consommateurs en ressources sont à la fois variées dynamiquement. Par conséquent, définir un mécanisme d'allocation des ressources aux utilisateurs d'une manière souple, dynamique et fiable est l'un des principaux enjeux dans le Cloud Computing.

Ainsi de nombreuses méthodes ont été proposées pour gérer le phénomène d'allocation de ressources dans le système Cloud. Ses principaux constituants sont représentés dans le schéma ci-dessous :

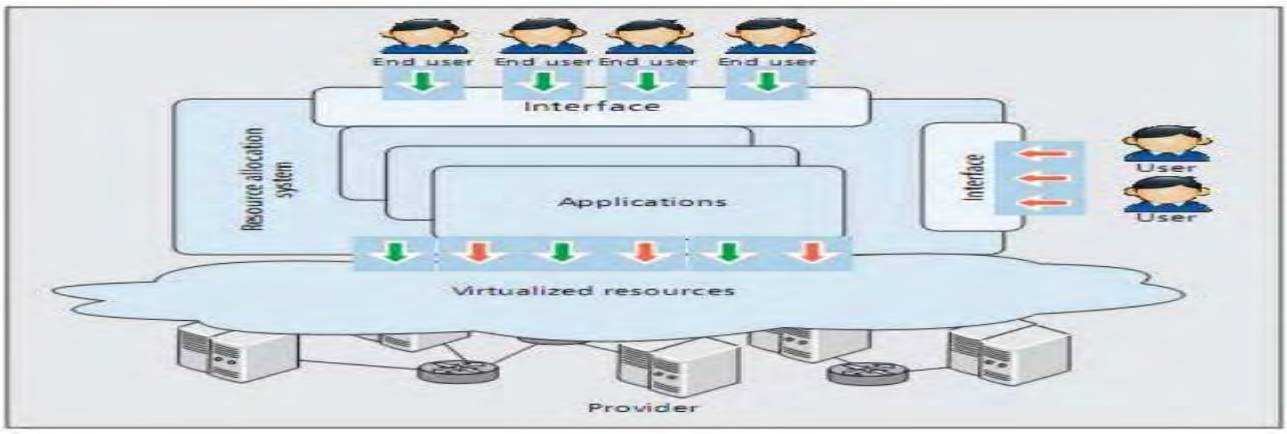


Figure 16 Les entités composant l'écosystème du Cloud computing [24]

Dans notre schéma le prestataire du Cloud fait face à deux types d'utilisateurs : utilisateur final et utilisateur du Cloud.

L'utilisateur de Cloud : il est situé au milieu, entre les utilisateurs finaux et le fournisseur de Cloud. Un utilisateur de Cloud peut être considéré comme un fournisseur de services, qui loue ressources/services offerts par le fournisseur afin d'héberger des applications qui seront consommés par les utilisateurs finaux [**5**].

L'utilisateur final : c'est le client d'une application qui utilise simplement les services de Cloud. Il est important de souligner que, dans certains scénarios (par exemple, calcul scientifique ou traitement par lots) les utilisateurs de Cloud computing peuvent se comporter comme des utilisateurs finaux dans le Cloud [**6**].

Le fournisseur de Cloud : c'est le propriétaire de l'infrastructure. C'est le responsable de la gestion des ressources physiques et virtuelles.

Les applications du Cloud : pratiquement sont illimitées, un système de Cloud computing pourrait exécuter tous les programmes qui peuvent fonctionner sur un ordinateur normal. Potentiellement, n'importe quelle application (logiciel de traitement de texte, programmes informatiques personnalisés conçus pour une société spécifique, etc.) pourrait fonctionner sur le Cloud computing. Les applications de Cloud computing peuvent être de différents types qui ont tous des besoins différents.

Les ressources virtualisées: le Cloud computing offre des ressources virtualisées pour les utilisateurs de Cloud computing. Il est basé sur la technologie de virtualisation qui permettre d'allouer les ressources des centres de données d'une manière dynamique pour les besoins des applications.

L'interface d'accès de l'utilisateur : elle définit le protocole de communication du Cloud avec l'utilisateur. L'Interface d'accès doit être équipée par les outils pertinents nécessaires pour une meilleure performance du système. En outre, l'interface d'accès peut être conçue comme des lignes de commande, basée sur des requêtes, basée sur console ou bien une interface de forme graphique. L'interface d'accès est très importante car si l'interface fournie à l'utilisateur n'est pas convivial, l'utilisateur ne peut pas utiliser facilement les services Cloud. Dans un autre scénario, on suppose qu'il y a deux fournisseurs de services Cloud computing qui fournissent les mêmes services, mais si on a une interface conviviale et la deuxième non, alors l'utilisateur serait préfère certainement celle avec une interface conviviale. Dans de tels scénarios, l'interface d'accès joue un rôle important et c'est pourquoi il est utilisé comme une fonction de comparaison dans le Cloud [**8**].

Le système d'allocation des ressources (RAS) : peut être considéré comme tous mécanismes qui visent à garantir que les exigences des applications sont prises en charge correctement par l'infrastructure du fournisseur. Ces mécanismes d'allocation des ressources devraient également examiner l'état actuel de chaque ressource dans l'environnement Cloud, afin d'appliquer des algorithmes pour mieux allouer les ressources physiques et/ou virtuels pour les applications des utilisateurs. Il est important de noter que les clients et les utilisateurs peuvent voir les ressources limitées comme illimité grâce au RAS.

❖ Modèles économiques d'allocation des ressources dans le Cloud

De plus en plus, l'implantation de fournisseur de services dans le Cloud connait un important développement et les solutions du Cloud s'élargissent progressivement pour devenir plus appropriés et plus attrayants pour toutes les entreprises et tous les particuliers. Cependant l'augmentation des sollicitations et des offres de ressources entraine une tournure plus complexe de l'allocation de ressources. Ce qui nous oblige à bien mettre en place des mécanismes d'allocations de ressources pour ne pas augmenter non seulement l'utilisation du CPU, mais aussi le temps de réponse. Il peut aussi créer un déséquilibre et une certaine iniquité dans le processus de partage des ressources et ainsi certains utilisateurs de Cloud pourraient être défavorisés par rapport à d'autres. Pour une bonne gestion d'allocation de ressource basée sur l'offre et la demande de ressources Cloud, le model économique est proposé. Il se base sur deux principales catégories :

- ✓ Des modèles basés sur la structure du marché.
- ✓ Et des modèles basés sur la vente aux enchères.

Pour les modèles basés sur la structure du marché, ils sont appliqués quand un grand nombre d'utilisateurs ne peuvent pas contrôler directement les prix de services. Les prestataires du Cloud appliquent des plans tarifaires pouvant varier d'un prestataire à l'autre. Quant aux modèles basés sur la vente aux enchères, ils sont adaptés aux situations où un petit nombre d'utilisateurs, cherchant à avoir un service spécifique, se mettent en concurrence [10].

❖ Stratégies d'allocation des ressources basées sur la structure du marché

Avec la stratégie d'allocation basée sur la structure du marché, les fournisseurs et les utilisateurs établissent un ensemble d'accord appelé les SLAs (Service Level Agreement). Les prestataires de Cloud ont besoin de mécanismes qui supportent la spécification de prix et augmentent l'utilisation du système. Tandis que les utilisateurs ont besoin de mécanismes qui garantissent que leurs objectifs soient atteints.

Modèle de marchandises : Dans ce type de modèle d'allocation de ressources, les prestataires de services spécifient leurs prix et facturent les services demandés par les utilisateurs selon la quantité de ressources qu'ils ont consommées. Dans ce cas, l'utilisateur est libre de choisir le fournisseur de service le plus approprié mais n'a aucun droit de changer le prix de service.

Le processus d'allocation de ressources est exécuté par des courtiers (brokers). Chaque courtier identifie plusieurs prestataires pour demander les prix et par la suite choisit un prestataire qui peut répondre aux exigences de l'utilisateur. La consommation du service est enregistrée et le paiement est fait comme convenu [**11**].

Modèle du prix affiché : La stratégie des prix affichés présente quelques offres spéciales pour motiver les clients à utiliser le service pendant les périodes libres. Les prestataires de services donnent le prix régulier, les offres bon marché et les conditions d'utilisation associées. Les courtiers observent le prix affiché et comparent s'il peut satisfaire l'exigence de l'utilisateur. Sinon, les courtiers appliquent la stratégie de marchandises.

Modèle de négociation : Dans la stratégie de négociation, le prix du service n'est pas donné par le prestataire de service uniquement mais aussi par le demandeur de service via la négociation. Dans ce scénario, un courtier ne compare pas tous les prix pour le même service, mais connecte avec un des

prestataires directement. Le prix offert par le prestataire pourrait être plus élevé que l'attente du client, donc le courtier propose un prix très bas. La négociation s'arrête quand un prix acceptable pour les deux parties est atteint ou quand un côté ne veut plus poursuivre la négociation. Dans ce cas, le courtier connectera avec d'autres fournisseurs et commencera à négocier de nouveau [**13**].

❖ Stratégies d'allocation de ressources basées sur la vente aux enchères

Différent du model d'allocation basé sur la structure de marché, le model basé sur la vente aux enchères met sur pied un ensemble de règles différents du prix. Parmi ces règles nous avons la manière dont on détermine prix de vente des services, le soumissionnaire gagnant, comment les ressources sont allouées, etc.

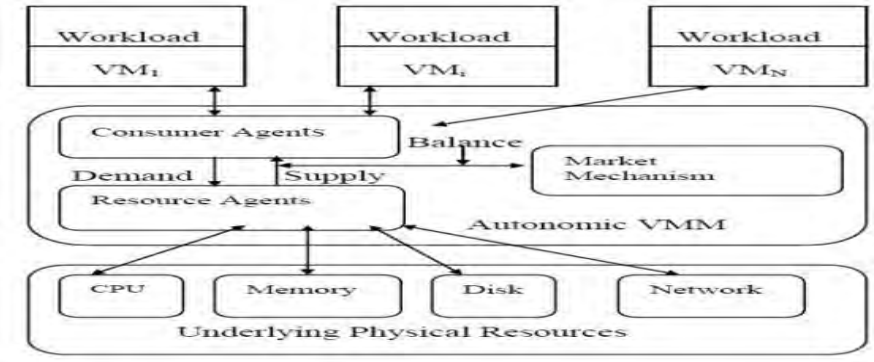


Figure 17 Une stratégie allocation des ressources basée sur le mécanisme de marché [24]

❖ Modèles d'allocation des ressources dans le Cloud basés sur la théorie des jeux

D'une manière générale la théorie des jeux constitue une approche mathématique de problèmes de stratégie tels qu'on en trouve en recherche opérationnelle et en économie. Elle étudie les situations où les acteurs du jeu, appelés joueurs, choisissent des stratégies dans le but de maximiser leurs profits dans un environnement où ils peuvent interagir. Elle consiste à chercher des états d'équilibre et le comportement optimal où les coûts et les gains de chaque état dépendent des choix des autres intervenants. C'est une approche largement utilisée pour la résolution de problèmes divers de l'ingénierie informatique, de gestion des ressources, d'intelligence artificielle, etc.

Ainsi, les éléments constituant un jeu sont les suivants :

- ✓ Le joueur est le candidat au jeu et peut être un individu, une compagnie, une nation, un équipement, etc. On définie un ensemble fini de joueurs J= {J1, J2…Jn}.
- ✓ La stratégie est la possibilité que peut prendre un joueur. Chaque joueur Ji de J a un espace fini de stratégies possibles, Si = {sk1, sk2… skm}. L'espace de stratégies est S = S1 x S2 x….Sn. Le résultat du jeu est une combinaison des n joueurs.

- ✓ Le profit est le gain obtenu par le joueur après l'achèvement du jeu. Dans le cas de l'allocation des ressources, le profit peut définir la quantité des ressources reçues.

Plusieurs modèles ont été définis pour la théorie de jeux, mais nous focalisons sur les deux modèles suivants : le model coopératif et le model non-coopératifs.

Un jeu est dit coopératif quand les différents joueurs peuvent communiquer librement entre eux et faire ainsi des compromis afin d'atteindre collectivement une décision ou une répartition optimale et satisfaisante pour tous les acteurs du jeu. Dans le cas des jeux non-coopératifs, les joueurs, qui ne communiquent pas entre eux, agissent selon le principe de rationalité économique où chacun d'entre eux essaye de prendre les meilleures décisions lui permettant de maximiser ses gains. Il existe aussi des modèles de jeux imparfaits, symétriques, asymétriques et séquentiels. Un jeu imparfait suppose qu'au moment de la demande des ressources, tous les joueurs soumettent leurs offres en même temps et que chaque joueur n'a connaissance que de la quantité disponible de ses propres ressources. Un jeu symétrique est un modèle où tous les joueurs ont les mêmes compétences de négociation contrairement au jeu asymétrique où les utilisateurs rivalisent tout en ayant des capacités financières et de négociation divergentes. Dans le cas des jeux séquentiels, les joueurs ne décident pas en même temps de leurs stratégies. Un joueur peut choisir sa stratégie en se basant sur les stratégies précédentes adoptées par les autres joueurs.

De nombreuses formes ont été utilisées pour représenter un jeu: la forme normale ou stratégique, la forme extensive et la forme caractéristique.

La forme normale : Un jeu est représenté sous forme normale, ou stratégique, quand l'ensemble des joueurs N = {1,..., n}, l'ensemble des stratégies pour chaque joueur et tous les profits correspondant à chacune des combinaisons possibles sont donnés. Si le jeu ne comporte que deux joueurs et un nombre fini et raisonnable de stratégies possibles, alors on peut représenter le jeu sous forme d'un tableau appelé matrice des gains.

		Joueur 2	
		Stratégie 1	**Stratégie 2**
Joueur 1	**Stratégie 1**	Résultat 1	Résultat 2
	Stratégie 2	Résultat 3	Résultat 4

Figure 18 Forme stratégie d'un jeu séquentiel

Forme extensive : Cette forme permet de représenter un jeu sous forme d'un arbre où chaque nœud non-terminal représente un joueur, chaque nœud terminal représente le résultat du jeu et chaque arc représente la stratégie adoptée par le joueur. Les jeux séquentiels sont généralement représentés sous forme extensive.

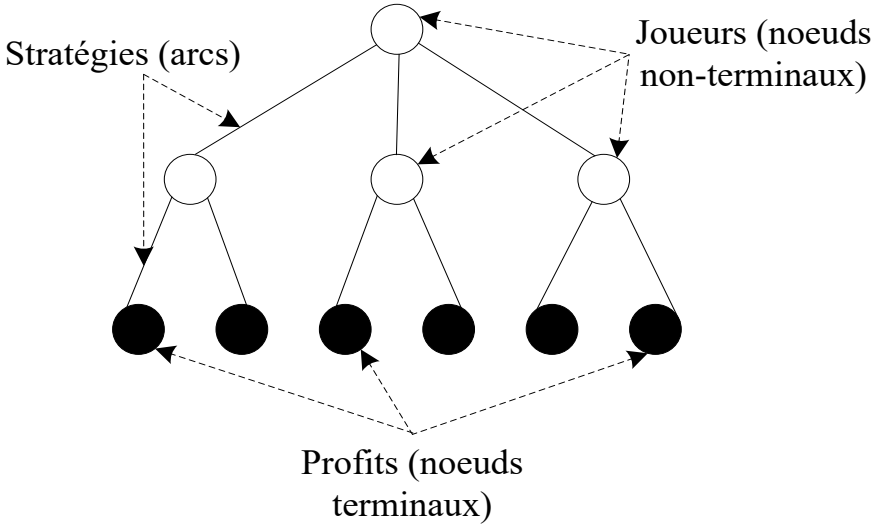


Figure 19 Forme extensive d'un jeu séquentiel [24]

Forme caractéristique : Cette forme est généralement utilisée pour représenter les jeux coopératifs. Sous cette forme un jeu est noté G= (N, v), où:

- ✓ N est l'ensemble des joueurs.
- ✓ v est la fonction caractéristique, elle associe à chaque sous-ensemble S de N (qui est une coalition) la valeur v(S), c'est-à-dire le gain obtenu par la coalition S.

  Elle s'écrit sous la forme suivante :
  max/ min(v)

sous contraintes:

contrainte 1,

contrainte 2,

contrainte 3 ….

❖ Modèles d'allocation des ressources dans le Cloud basés agents

L'allocation basé agent est subdivisé en deux : L'allocation adaptatif basé agent et allocation automatisé basé-agent avec l'utilisation de micro accords (agreements)

Pour la première, de nombreux facteurs sont pris en compte pour ce type de modèle d'allocation de ressource : la distance géographique entre le lieu de consommateur et les centres de données, et la charge de travail des centres de données.

Ainsi une architecture du nom de **testbed** a été élaborée et mise en œuvre pour mieux donner une vision sur le model d'allocation de ressource adaptative proposé. Dans l'architecture, les auteurs utilisent deux types d'agents : les consommateurs et les fournisseurs de service. Ils se comportent au nom de chaque consommateur et fournisseur. Le schéma du **testbed** est représenté ci-dessous :

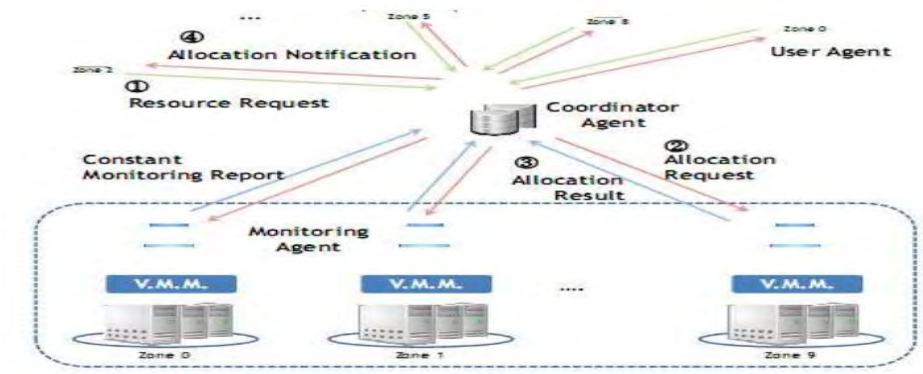


Figure 20 Vue d'ensemble sur le testbed d'allocation des ressources basé agent [24]

Le testbed est constitué par un ensemble de centre de donné et de consommateurs distribués dans diverses zones géographiques. Il possède également trois(3) types d'agents :

Agent Utilisateur: Il envoie un message de demande d'allocation pour le compte d'un consommateur à l'agent coordinateur. Puis, il attend de recevoir un message de résultat d'allocation de l'agent coordonnateur.

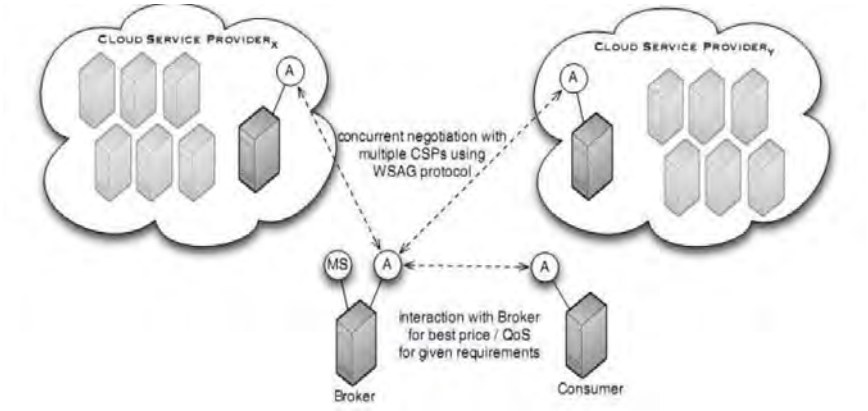
Agent Coordinateur: Un agent coordonnateur est chargé de coordonner l'allocation des ressources pour les participants (les consommateurs et les fournisseurs). Son rôle principal est de trouver un centre de données approprié pour une demande de consommateur.

Agent de contrôle: Un agent de contrôle est chargé de contrôler chaque centre de données distribuées. Cela signifie que l'agent est au même endroit avec le centre de donné.

Pour la deuxième, les auteurs ont proposés un service intelligent d'allocation des ressources Cloud (ICRAS) qui nécessite une architecture sous-jacente composé de trois éléments principaux : 1) un consommateur, 2) un fournisseur de service Cloud (CSP) et 3) un agent ICRAS. Ces éléments représentent les trois rôles sur le marché, qui peut contenir plusieurs instances de chaque un de ces derniers. En outre, cette architecture fournit les mécanismes et les protocoles qui permettent à ces éléments de communiquer entre eux et négocier les micro-SLA d'une manière autonome. Cette architecture est illustrée sur la **Figure 21 [17]**.

Le consommateur : Chaque utilisateur est en interaction avec un agent ICRAS et donne ses prétentions dans une offre SLA (besoins en hardware, software, les priorités, une plage d'options, et les dépendances entre ces exigences). Si les besoins en ressources sont changés, le consommateur doit informer un l'agent d'ICRAS par ces changements.

CSP (fournisseur de service Cloud) : le CSP fourni une interface qui prend en charge deux fonctions principales : la négociation et la migration pour participer à l'architecture ICRAS.



.

Figure 21 Une allocation des ressources Cloud basé-agent avec l'utilisation des micro-accords [24]

L'agent ICRAS : il a cinq(5) responsabilités: découvrir les offres de CSP, évaluer ces offres, négocier un contrat de service avec le CSP au nom du client, gérer l'approvisionnement des nouvelles ressources Cloud pour détecter les violations de SLA et gère la migration d'un CSP vers un autre CSP [18].

III.2.4°) Techniques d'allocation de ressources dans l'Edge computing

Il existe de nombreuses méthodes permettant de faire une allocation des ressources efficace dans les systèmes Edge Computing, parmi celles-ci nous avons :

- **Modèles An envy-free auction mechanism (Un mécanisme d'enchères sans envie) d'allocation des ressources dans l'edge computing**

Le principal défi du Mobile Edge Computing (MEC) et de Mobile Cloud Computing (MCC) est de voir comment allouer les ressources. La manière envisagée est d'allouer les ressources selon les modèles d'enchères, dans lequel les utilisateurs enchéris pour utiliser une certaine quantité de ressources. Dans cette partie, nous parlons les problèmes de l'allocation des ressources et de facturation dans un système Informatique de bord à deux niveaux. Nous considérons un système dans lequel les serveurs de différentes capacités sont situés dans le nuage ou au bord du réseau.

Les utilisateurs mobiles en compétition pour ces ressources et ont des exigences hétérogènes. Nous concevons un mécanisme basé sur les enchères qui alloue des ressources Cloud. Le mécanisme proposé est nouveau dans le sens où il gère l'allocation des ressources disponibles aux deux niveaux du système en combinant les caractéristiques de la position et des ventes aux enchères combinatoires [5].

- **Définition**

Malgré l'évolution de la technologie, les appareils mobiles sont limités en termes de ressources (stockage, puissance de calcul, la vie de la batterie…….).ces problèmes peuvent être réduit en exécutant des applications à distance sur une infrastructure statique qui ne souffre pas de ces limitations (MEC, MCC). Le MEC et le MCC produisent de nombreux avantages pour les utilisateurs mobiles. Il augmente la durée de vie de la batterie de dispositifs mobiles, permet la réduction du temps d'exécution d'applications, et améliore la capacité de stockage de données et la puissance de traitement. Cependant, dans le MCC et le MEC, les centres de données centralisés et décentralisés qui sont utilisés respectivement par les services de Cloud computing et de l'edge computing sont généralement loin d'être les utilisateurs finaux, et par conséquent, la

communication entre les appareils mobiles. Ainsi le MEC a été mis sur place pour assurer le traitement des données au bord de toute sorte de ressource informatique. Il permet de minimiser la consommation d'énergie, délai moyen, et le coût total. Par rapport aux centres de données en nuage, les systèmes de bord ont des ressources beaucoup plus limitées conduisant à une concurrence accrue entre les utilisateurs qui désirent acquérir des ressources au sein de leur proximité. Par conséquent, l'efficacité de l'allocation des ressources, les incitations et la monétisation sont des défis majeurs dans les systèmes [6].

Le mécanisme basé sur les enchères pour l'allocation des ressources et la tarification fait l'association caractéristique des positions et des ventes aux enchères combinatoires et administre les demandes de ressources de différentes natures des consommateurs. Il a besoin d'un temps d'exécution très réduit même s'il s'agit de plusieurs demandes de ressources avec des milliers d'utilisateurs.

➢ Fonctionnement

De nombreux moyens ont été mis sur place pour la conception d'un mécanisme d'allocation de ressource basé sur la tarification dans l'informatique.

Nous considérons un fournisseur de ressource à bord fournissant un certains nombre d'instance limité de machine virtuelle (vm) sur le périphérique et au niveau du Cloud. Vu que les ressources sont plus proche des utilisateurs au niveau edge computing, ces derniers optent l'exécution de leurs applications sur l'architecture décentralisée (edge).cependant, les consommateurs sont en concurrence pour obtenir des ressources sur les serveurs de bord du fait que les instances au niveau du bord sont limités par rapport celles du Cloud.

Sur ce fait nos ressources sont vendu au prestataire dans la base d'une vente à l'enchère dans lequel le prix des ressources ne sont pas fixé à l'avance mais plutôt défini en utilisant un mécanisme de vente aux enchères. Nous considérons qu'il y'a n utilisateurs qui se font concurrence pour des ressources situées à deux niveaux :

Bord (l=1) et nuage (l=2) , ou l indique le niveau. Les deux niveaux offrent leurs ressources aux utilisateurs en tant que m types d'instances de VM, ou chaque type d'instance de VM est caractérisé par trois types de ressources :vCPU (t=1), Mémoire (t=2) et Stockage (t=3).Une instance de machine virtuelle de type k fournit qkt unité de type t ,ou k=1 ,…… ,m et t=1,2,3 le nombre total d'instance de type k disponible au niveau l est noté C.Par exemple une « petite »instance de VM se compose 1vCPU, de 2Go de mémoire et de 20Go de Stockage. C'est-à-dire l'instance de type k=1 est caractérisée par q11=1, q12=2 et q13= 20.

L'utilisateur i demande un ensemble d'instance de machine virtuelle et soumet une offre pour le paquet la demande de l'utilisateur i est noté par O. (bi ri)= (b1,…….bim ; ri1,…….rim) ou bik est l'offre pour une instance de machine virtuelle de type k , et rik est le nombre de machine virtuelle d'instance de type k demandé par l'utilisateur i. Les demandes des utilisateurs sont soumises à un mécanisme qui détermine la répartition des Instances VM aux utilisateurs et le prix qu'ils doivent payer pour leurs allocations [6].

❖ Modèles G-ERAP mechanism (Greedy Edge Resource Allocation and Pricing) d'allocation des ressources dans l'edge computing

G-ERAP est utilisé périodiquement à des intervalles de temps d'une durée spécifiée. La répartition et le prix déterminés par le mécanisme est valide pour l'intervalle de temps actuel. L'entrée à G-ERAP est constituée

du vecteur de requêtes (ri) des utilisateurs, et le vecteur des capacités de la machine virtuelle (C). G-ERAP détermine Comment ces ressources sont assignées aux utilisateurs ?

Premièrement, le mécanisme détermine l'enchère moyenne par unité de ressource pour chaque utilisateur.

L'enchère moyenne pour chaque utilisateur i et défini comme suite [8]:

$$B_i = \frac{\sum_{k=1}^{m} b_{ik} r_{ik}}{\sum_{k=1}^{m} \sum_{t=1}^{3} w_t q_{kt} r_{ik}}$$

Cependant, il sélectionne dans un ordre non croissant les consommateurs de leurs enchères moyennes et attribut d'instances de machine virtuelle aux utilisateurs à partir du bord (premier niveau) Pour l'utilisateur actuel, il vérifie s'il y a suffisamment de ressources au niveau actuel. S'il y en a, il attribue la demande des instances de machine virtuelle à l'utilisateur et met à jour le bien-être et la capacité. S'il n'y en a pas assez de ressources pour allouer le paquet demandé au premier niveau, il augmente l'index du niveau de un (c'est-à-dire qu'il commence à allouer Instances VM au deuxième niveau) et stocke l'index de l'utilisateur en tant que premier utilisateur attribué au deuxième niveau.

Ainsi le G-ERAP s'arrête dès qu'il atteint un utilisateur pour lequel il n'ya pas assez de ressources pour satisfaire la demande au deuxième niveau.

Ensuite, G-ERAP détermine les paiements de base pour chaque unité des ressources au premier niveau et au deuxième niveau.

Supposons que l'utilisateur u est le dernier utilisateur de l'ordre trié qui est attribué au premier niveau. Par conséquent, l'utilisateur u + 1 est le premier utilisateur de la liste à attribuer au deuxième niveau [9].

L'algorithme de G-ERAP est représenté ci-dessous.