GESTION DES RESSOURCES DANS LES SYSTEMES CLOUD-EDGE COMPUTING

III.1 Introduction

L'allocation des ressources est l'un des principaux facteurs influents pour fournir un traitement efficace et économique des ressources dans l'infrastructure en tant que service Clouds. Bien qu'il existe de nombreux défis à relever pour fournir un répartiteur de ressources efficace, il est d'une grande importance d'optimiser l'utilisation des ressources physiques. Il existe plusieurs travaux axés sur l'optimisation de la sélection des machines virtuelles (VM) pour la migration, cependant, on accorde moins d'attention au placement des VM sélectionnées sur les machines physiques disponibles, en particulier pour le modèle de demande de réservation avancée.

Load-balancing est une technique qui garantit que le serveur d'une organisation n'est pas surchargé de trafic. Avec des mesures d'équilibrage de charge en place, les charges de travail et les demandes de trafic sont réparties entre les ressources du serveur pour offrir une résilience et une disponibilité accrues.

III.2 Définition et Type de ressources

L'allocation des ressources est le processus d'affectation des ressources qui sont disponibles pour les tierces parités par les fournisseurs de cloud l'environnement cloud en cas de besoin. Les allocations de ressources sont basées sur le schéma de tarification et le moment de l'attribution.

L'allocation des ressources comprend deux facteurs à éviter :

- **Sur-approvisionnement** des ressources - cette complexité survient lorsque les ressources vendues sont plus que les ressources disponibles.
- **Sous-approvisionnement** des ressources - ce problème se pose uniquement lorsque moins de ressources sont affectées au gens que la demande.

Passant aux types de ressources, la discussion est faite sur les ressources qui ont été fournies sur le cloud soumis au schéma de gestion des ressources.

- **Ressource informatique**

Qui comprend la collecte de mémoire, réseau, processeur, périphériques d'entrée / sortie dans l'environnement cloud. Ceux-ci sont appelés collectivement les machines physiques.

Selon les besoins de l'utilisateur, les ressources informatiques doivent être attribué ou acheté. Le concept de VM relève PM où PM crée un logiciel virtuel sur lequel l'utilisateur peut s'exécuter machine virtuelle dans différents OS, applications et plates-formes.

- **Ressources de mise en réseau**

Bande passante, stockage, communication, défis, trafics de tels problèmes arrivent du côté du réseau qui peuvent être pris soins en travaillant sur des protocoles pour améliorer la qualité de service nuage.

- **Ressources de stockage**

Lorsque l'évolutivité concerne le stockage (elle doit être atteinte en considérant la propriété ACID. Maintenant un jour le nuage le stockage est basé sur les technologies de stockage de données « NO SQL » sous certaines conditions fonctionnelles qui ont été incluses pour stockage de documents, valeur-clé.

- **Ressources énergétiques**

L'utilisation quotidienne de l'énergie par le système traite de l'énergie ressource. L'énergie consommée par le système pour fournir et allouer de la ressource est bien inférieure à l'énergie consommée par le système inactif, attendant qu'une ressource soit alloué. Cela a conduit à une autre technologie à savoir le cloud vert l'informatique.

III.3 Allocation des ressources dans les systèmes Cloud et Edge Computing

De plus en plus, on note une croissance importante de la popularité des systèmes qui offrent des ressources informatiques à la demande, basé sur la facturation à l'usage et l'équité des ressources partagées selon la demande des utilisateurs 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.

L'allocation de ressources (RA) consiste à attribuer les ressources disponibles aux applications nécessaires via Internet. Elle permet une répartition des ressources disponibles entre les utilisateurs de cloud et les applications d'une manière économique et efficace.

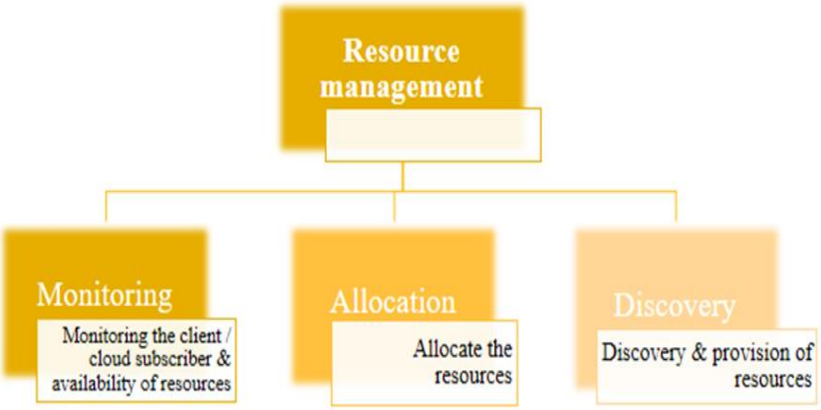

Figure 8 : Module de gestion de Ressource

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 des garanties de QoS aux utilisateurs cloud tel que la haute disponibilité des ressources, le partage des ressources. Elle permet de gérer les ressources physiques tels que les noyaux CPU, espace disque et la bande passante du réseau.

RA est un élément important dans l'allocation des ressources dans le cloud si celle-ci n'est gérée avec précision cela affaiblira les services du cloud d'où la nécessité d'adopter des stratégies.

III.3.1 Stratégies d'allocation des ressources (RAS)

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

Elle consiste à intégrer les activités des fournisseurs de cloud pour utiliser et allouer des ressources rares dans la limite de l'environnement cloud afin de répondre aux besoins de l'application cloud.

Il nécessite le type et la quantité de ressources nécessaires à chaque application pour effectuer le travail.

L'ordre et l'heure de l'affectation des ressources constituent également une entrée pour un RAS optimal.

- **Les Critère du RAS :**

Performance

La performance se définit par le plus petit temps de réponse pour l'exécution d'une tache résultante 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 dès 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.

III.3.2 Phases d'allocation de ressources

Nous avons deux (2) phases pour présenter un processus séquentiel complet de la gestion des ressources dans le cloud computing.

- **Phase 1 : L'affectation des ressources**

L'affectation initiale des ressources se réfère à la manière dont les ressources sont demandées dans le cloud computing pour la première fois.

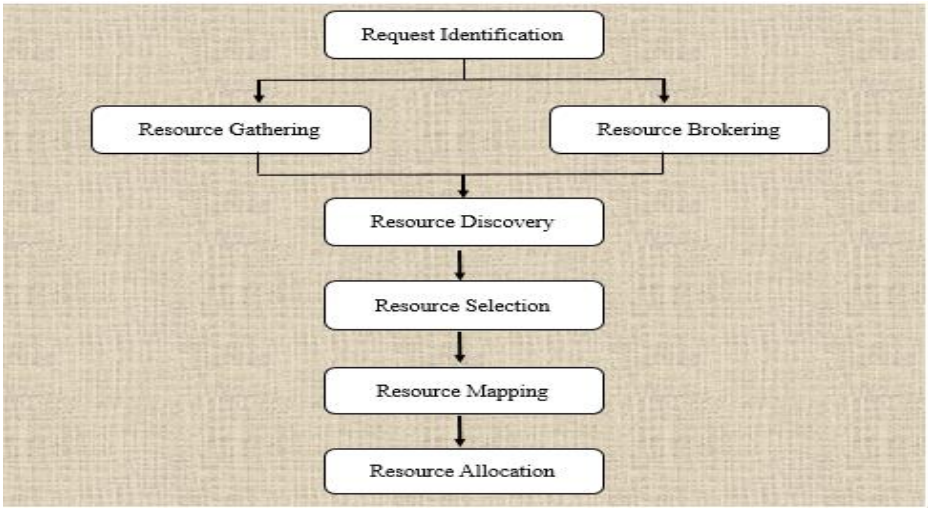

Figure 9: Organigramme d'affectation de ressources

Request Identification : C'est la première étape de l'affectation de ressources. Dans cette étape, diverses ressources seront identifiées 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 fournis par les utilisateurs de 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 de cloud. Son objectif principal est de satisfaire les utilisateurs de cloud de leurs besoins et générer de revenus pour les fournisseurs cloud.

- **Phase 2 : Optimisation périodique des ressources**

Comme son nom l'indique, il s'agit d'une phase où la gestion des ressources se fait à intervalles réguliers une fois la phase 1 est terminé. Ici, l'optimisation périodique des ressources est présentée comme un processus pour deux catégories différentes des ressources qui sont des ressources non virtualisées et ressources virtualisées.

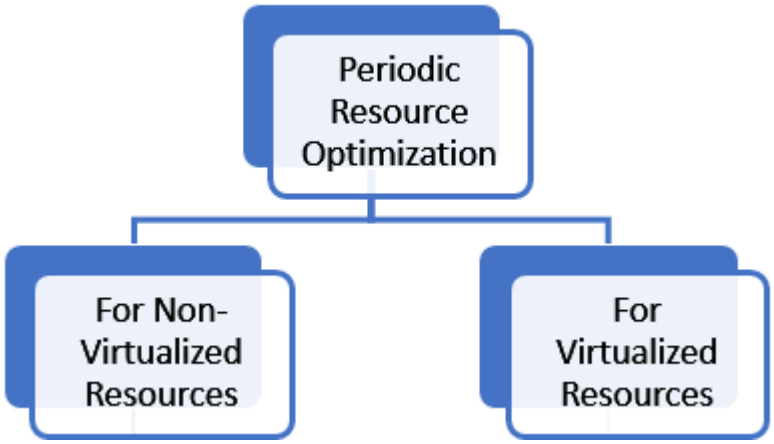

Figure 10 : Optimisation périodique des ressources

Les ressources non virtualisées sont également appelés ressources physiques. Pour les deux catégories de ressources, l'optimisation périodique des ressources contient des étapes similaires. La seule différence est que les ressources virtualisées peuvent être assemblées ensemble selon les besoins en ressources et peuvent être démonté.

- **Optimisation des ressources physiques**

Ressources Monitoring : surveillance des ressources est la première et cruciale étape de la période d'Optimisation des ressources. Diverses ressources cloud non virtualisées sont surveillées pour analyser l'utilisation des ressources. Cette étape permettra également de surveiller la disponibilité des ressources gratuites pour les besoins futurs. Le principal problème de la surveillance des ressources cloud est d'identifier et de définir des mesures.

Ressource Modelling - Ressources Prediction : permet de prédire les différentes ressources non virtualisées requises par les applications de cloud computing. C'est une des étapes complexes car les ressources en nuage ne sont pas de nature uniforme. En raison de cette non-uniformité, il est très difficile de prévoir les besoins en ressources pour les périodes de pointe et les périodes hors pointe.

Ressources Brokering : c'est la négociation des ressources non virtualisées avec les utilisateurs de cloud computing pour s'assurer qu'ils sont disponibles selon l'exigence.

Ressource Adaptation : selon les exigences des utilisateurs de cloud, les ressources non virtualisés cloud peuvent être augmentées ou réduites vers le bas. Cette étape peut augmenter le coût de perspective des fournisseurs de cloud.

Réallocation de ressources : permet de réaffecter ou redistribuer des ressources sur le cloud aux utilisateurs. Son objectif principal est de satisfaire les besoins des utilisateurs de cloud et la génération de revenus pour les fournisseurs de cloud.

Tarification des ressources ou Ressource Pricing : l'un des étapes les plus importantes de la perspective des fournisseurs de cloud et des clients de cloud computing. La tarification de l'utilisation des ressources cloud sera effectuée dans cette étape.

- **Optimisation des ressources virtuelles**

Les étapes Ressource Monitoring, Ressource Modelling /Ressource Prediction, Ressource Brokering, Ressource Adaptation, Ressource Réallocation et Ressource Pricing sont identiques à celui des étapes d'optimisation des ressources pour le non virtualisation des ressources.

Ressource Bundling : selon les besoins, diverses ressources non virtualisées peuvent être regroupées dans des ressources virtualisées.

Ressource Fragmentation : Diverses ressources virtualisées doivent être fragmentées pour libérer des ressources non virtualisées. Après cette étape, diverses ressources non virtualisées peuvent être regroupées dans des ressources virtualisées.

III.3.3 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.

Une allocation est sans envie si aucun utilisateur ne peut améliorer son utilité en échangeant des offres avec utilisateur avec la même demande d'instances de VM.

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

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

Le mécanisme proposé, appelé G-ERAP (Greedy Edge Allocation et tarification des ressources), est invoqué périodiquement à des intervalles de temps d'une durée spécifiée. L'allocation et le prix déterminés par le mécanisme est valide pour l'intervalle de temps actuel. L'entrée de G-ERAP est constitué du vecteur de requêtes (θi) des utilisateurs, et le vecteur des capacités VM (C). G-ERAP détermine comment ces ressources sont affectées aux utilisateurs. La sortie du mécanisme se compose de la protection sociale V, le prix pour chaque unité de ressources du premier niveau et du second level (π1, π2), et la matrice d'allocation X, où X = [xil], i = 1, ..., n et l = 1, 2. Premièrement, le mécanisme détermine l'offre moyenne par unité de ressource pour chaque utilisateur.

L'enchère moyenne de l'utilisateur i est définie comme suit,

Ensuite, il trie les utilisateurs dans l'ordre non croissant de leur enchères moyennes et attribue des instances de VM aux utilisateurs à partir du premier niveau (c'est-à-dire le niveau de bord), en conséquence.

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 VM à l'utilisateur et met à jour le social le bien-être et la capacité. S'il n'y en a pas assez ressources pour allouer le bundle demandé au premier niveau, il augmente l'indice du niveau d'un (c'est-à-dire qu'il commence à allouer des instances de VM au deuxième niveau) et stocke l'index de l'utilisateur en tant que premier utilisateur attribué au deuxième niveau, l'utilisateur noté u. de garantir la liberté d'envie et la rationalité individuelle, G-ERAP s'arrête une fois qu'il atteint un utilisateur pour lequel il n'y a pas assez de ressources pour satisfaire le bundle demandé au deuxième niveau.

Ensuite, G-ERAP détermine les paiements de base pour chaque unité de ressource au premier niveau et au deuxième niveau. Supposons que l'utilisateur u soit le dernier utilisateur dans 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 second

Algorithm 1 G-ERAP Mechanism

Input: Vector of requests; $\theta_i = (b_{i1}, \ldots, b_{im}; r_{i1}, \ldots, r_{im})$
Input: Vector of VMs' capacities at each level;
$C = \{C_{11}, \ldots, C_{1m}; C_{21}, \ldots, C_{2m}\}$

$V \leftarrow 0$
$X \leftarrow 0$
for $k = 1, \ldots, m$ **do**
  $\tilde{C}_{1k} \leftarrow C_{1k}$
  $\tilde{C}_{2k} \leftarrow C_{2k}$
for $i = 1, \ldots, n$ **do**
  $B_i \leftarrow \frac{\sum_{k=1}^{m} b_{ik} r_{ik}}{\sum_{k=1}^{m} \sum_{t=1}^{3} w_t q_{kt} r_{ik}}$
Sort users in non-increasing order of their B_i
$l \leftarrow 1$
$i \leftarrow 1$
while $i \leq n$ **do**
  $available \leftarrow true$
  for $k = 1, \ldots, m$ **do**
    if $\tilde{C}_{lk} < r_{ik}$ **then**
      $available \leftarrow false$
      break
  if $available$ **then**
    for $k = 1, \ldots, m$ **do**
      $\tilde{C}_{lk} \leftarrow \tilde{C}_{lk} - r_{ik}$
      $V \leftarrow V + \alpha_l b_{ik} r_{ik}$
    $x_{il} \leftarrow 1$
    $i \leftarrow i + 1$
  else
    if $l = 1$ **then**
      $u \leftarrow i - 1$
      $l \leftarrow l + 1$
    else
      break
if $i < n$ **then**
  $\pi_2 \leftarrow \alpha_2 B_i$
else
  $\pi_2 \leftarrow \alpha_2 (B_i - \epsilon)$
$\pi_1 \leftarrow \pi_2 + \frac{(\alpha_1 - \alpha_2)}{2}(B_u + B_{u+1})$

Output: $X = \{x_{11}, \ldots, x_{n1}; x_{12}, \ldots, x_{n2}\}$
Output: V
Output: (π_1, π_2)

III.3.4. Allocation Dynamique Vs. Allocation Statique

III.3.4.1 Allocation Statique

L'allocation statique permet d'attribuer des ressources fixes à l'utilisateur ou à l'application cloud. Dans l'allocation statique l'utilisateur du cloud doit connaître le nombre d'instances de ressources nécessaires pour l'application et quelles sont les ressources demandées.

Ces ressources devraient viser à confirmer les demandes de charge de pointe de l'application. Mais la limitation de l'allocation statique est généralement affectée par la surutilisation ou la sous-utilisation de ressources informatiques basées sur la charge de travail normale de l'application. Ce n'est pas rentable et est lié à une utilisation insuffisante de la ressource en période hors pointe.

III.3.4.2 Allocation Dynamique

L'allocation dynamique consiste à fournir des ressources cloud à la volée lorsque l'utilisateur ou l'application en demande, en particulier pour éviter la surutilisation et la sous-utilisation des ressources.

III.3.4.3 Allocation statique VS Allocation Dynamique

L'inconvénient de l'allocation dynamique est lorsque des ressources nécessaires sont demandées à la volée et qu'elles pourraient ne pas être accessibles. Ainsi, le fournisseur de services doit allouer les ressources de différents centres de données cloud.

III.4 Techniques de base d'allocation de ressources orientée Load-balancing pour le système Edge computing

III.4.1 Définition de load-balancing

Load-balancing est une technique qui garantit que le serveur d'une organisation n'est pas surchargé de trafic. Avec des mesures d'équilibrage de charge en place, les charges de travail et les demandes de trafic sont réparties entre les ressources du serveur pour offrir une résilience et une disponibilité accrues.

L'équilibrage de charge peut également être définie comme étant le processus de réaffectation de la charge totale sur les nœuds individuels du système collectif pour utiliser efficacement les ressources et améliorer le temps de réponse des travaux, en supprimant simultanément une condition dans laquelle certains des nœuds sont surchargés tandis que d'autres sont sous-chargés. Un défi clé sur ces applications est que les nuages doivent garder les performances identiques ou meilleures lors de l'explosion de données.

III.4.2 Fonctionnement

Une séquence d'équilibrage de charge typique fonctionne comme suit :

Le trafic arrive sur votre site : les visiteurs de votre site envoient de nombreuses requêtes simultanées à votre serveur via Internet.

Le trafic est réparti entre les ressources du serveur : le matériel ou le logiciel d'équilibrage de charge intercepte chaque requête et la dirige vers le nœud de serveur approprié.

Chaque serveur gère une charge de travail raisonnable : le nœud reçoit la demande et est capable d'accepter la demande et de répondre à l'équilibreur car il n'est pas surchargé avec trop de demandes.

Le serveur renvoie la requête : le processus est effectué dans l'ordre inverse pour renvoyer la réponse du serveur à l'utilisateur.

Cela peut sembler évident, mais il convient de noter que les étapes ci-dessus ne peuvent être effectuées que s'il existe plusieurs ressources (serveur, réseau ou virtuelles) déjà établies. Sinon, s'il n'y a qu'un seul serveur ou une seule instance de calcul, toutes les charges de travail sont distribuées au même endroit et l'équilibrage de charge n'est plus nécessaire.

III.4.3 Les techniques de base d'allocation de ressources orientée load-balancing

Round Robin

Round robin load balancer est le plus simple des algorithmes d'équilibrage de charge de machines virtuelles. Les tâches sont affectées aux machines virtuelles en faisant la rotation sur l'ensemble des machines virtuelles instanciées par le contrôleur du centre de données. Les requêtes de cet algorithme sont réparties entre des machines virtuelles, l'autre s'appuyant sur l'aide du contrôleur du data centre.

L'ordre d'allocation des tâches a lieu dans chaque machine virtuelle localement et indépendamment de l'autre machine distante, sur la base du nombre de tâches disponibles et du nombre de machines virtuelles.

Dans le round robin, un temps quantique fixe est attribué au travail. L'accent principale dans le round robin est mis sur l'équité et la limitation du temps.

Comme le nom du tournoi à la ronde l'indique, il fonctionne dans un motif circulaire. Chaque nœud est fixé avec un tranche de temps et exécute une tâche à l'heure indiquée à son tour. Il est moins complexe.

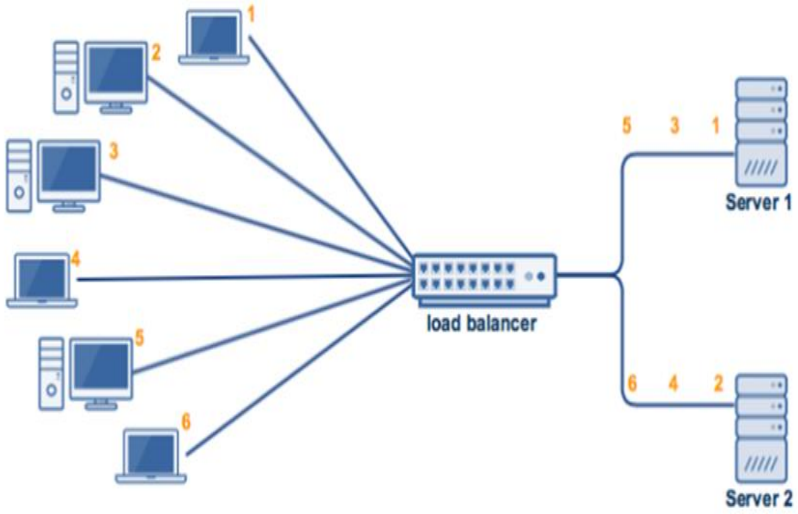

Figure 11 : Round Robin Algorithmes [W-22]

Étape 1 : les équilibreurs de charge Round Robin VM (RR load balancer) ont un index de VMs. Initialement, toutes les machines virtuelles n'ont aucune allocation.

Étape 2 :

a) le contrôleur du centre de données reçoit les requêtes des utilisateurs.

b) les demandes sont affectées aux machines virtuelles de manière circulaire.

c) l'équilibreur de charge Round Robin VM allouera le quantum de temps pour l'exécution de la demande de l'utilisateur.

Étape 3 : après l'exécution des requêtes, Les machines virtuelles sont désallouées par la charge de machine virtuelle Round Robin balancer.

Étape 4 : le contrôleur du centre de données vérifie les nouvelles demandes en attente dans la file d'attente.

Étape 5 : continuer à partir de l'étape 2.

Min-Min Load-balancing

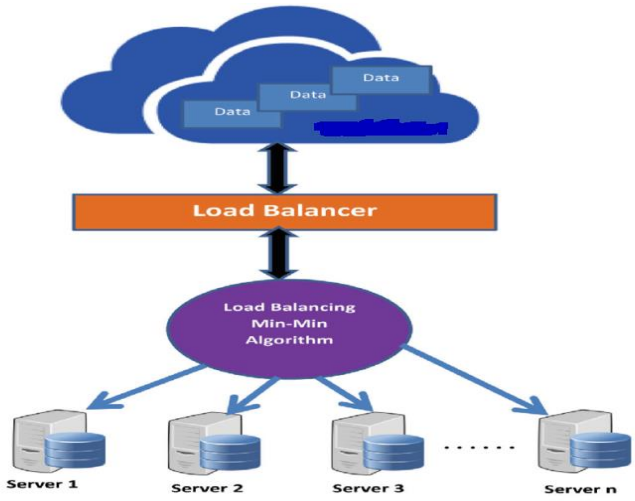

Figure 12 : Min – min Algorithmes [W-23]

Parmi toutes les tâches qui prennent le moins de temps, la recherche dans la première étape. La tâche consiste à organiser, selon que plus petite valeur de temps sur la machine. Le temps d'exécution pour d'autres tâches sont également mises à jour.

Min-min montre les meilleurs résultats quand il y a de petits tâches plus en nombre. La famine est un inconvénient majeur. La variation de la machine et des tâches ne peut être prédite grâce à cet algorithme.

Max-Min Load Balancing

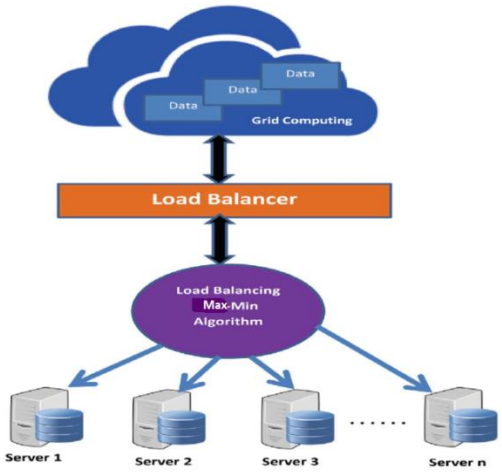

Figure 13 : Max-min Algorithmes

Max-min est identique à l'algorithme min-min. Mais Max-min choisit la tâche avec une valeur maximale et donne au machine respective. Après avoir attribué la tâche, la machine fonctionne selon les mises à jour. Ces tâches assignées sont supprimées de la liste. Le nœud et les tâches choisis s'organisent dans un les mises à jour du modèle concernant le temps de disponibilité sont données en combinant la durée d'exécution du travail.

III.5 Conclusion

La gestion des Ressources est un processus efficace et efficient qui gère les ressources ainsi fournie des garanties de QoS aux utilisateurs cloud tel que la haute disponibilité des ressources, le partage des ressources. Elle permet de gérer les ressources physiques tels que les noyaux CPU, espace disque et la bande passante du réseau.