



UNIVERSITE SIDI MOHAMED BEN ABDELLAH
FACULTE DES SCIENCES DHAR EL MAHRAZ-FES



Master Big Data Analytics & Smart System



Rapport Sur le thème :

Equilibrage de charge dans le Cloud Computing

Réaliser par :

ES-SAYEH Rabie

EL AZIZY Zouhair

SANGARE Boubacar Diam

Encadrer par :

Mr. Abdellatif EL ABDERRAHMANI

Année universitaire : 2023/2024

Table de matière

Introduction générale.....	4
Chapitre I . Introduction au Cloud Computing :	6
I. Introduction :	6
II. Contexte et définition du Cloud Computing :	6
II.1. Concept du Cloud Computing :	6
II.2. Caractéristiques et avantages du Cloud Computing :	7
II.2.a. Caractéristiques du Cloud Computing :	7
II.2.b. Avantages du Cloud Computing :	8
III. Modèles de service dans le Cloud Computing :	8
III.1. Infrastructure as a Service (IaaS) :	9
III.2. Platform as a Service (PaaS):	10
III.3. Software as a Service (SaaS):	10
IV. Les acteurs du Cloud Computing :	10
V. Modèles de déploiement dans le Cloud Computing :	11
V.1. Cloud public :	11
V.2. Cloud privé :	11
V.3. Cloud hybride :	12
V.4. Cloud communautaire :	12
VI. Architecture du Cloud Computing :	12
VI.1. La virtualisation:	13
VI.2. L'infrastructure :	13
VI.3. Le Datacenter :	13
VI.4. Cloud Computing et sécurité :	14
VII. Conclusion :	14
Chapitre II . Équilibrage de charge dans le cloud computing :	16
I. Introduction :	16
II. Définition et importance de l'équilibrage de charge :	16
II.1. Définition :	16
II.2. Importance de l'équilibrage de charge :	16
III. Types d'équilibrage de charge dans le Cloud Computing :	18
III.1. Équilibrage de charge de niveau réseau :	18
III.2. Équilibrage de charge de niveau application :	18
III.3. Équilibrage de charge de niveau global :	18
IV. Les Technologies d'équilibrage de charge :	18
IV.1. Hardware Load balancing :	19
IV.2. Software Load Balancing :	19

V.	Approches et algorithmes d'équilibrage de charge :	20
V.1.	Approches d'équilibrage de charge dans le cloud :	20
V.1.a.	Approche statique :	20
V.1.b.	Approche dynamique :	21
1.	Approche centralisée:	22
2.	Approche distribuée :	22
3.	Approche source-initiative vs receiver-initiative :	23
V.2.	Les algorithmes de Load balancing :	24
V.2.a.	Algorithmes statiques :	25
V.2.b.	Algorithme dynamique :	26
VI.	Migration des VM dans l'équilibrage de charge :	29
VII.	Algorithmes de l'apprentissage automatique utilisés dans le cloud computing :	31
VII.1.	Technique basée sur la régression, Random Forest et AdaBoost :	32
VII.2.	Support Vector Machines (SVM) et technique K-suggest :	32
VII.3.	Approche du réseau de neurones artificiels propagés (BPANN) :	32
VII.4.	ANN et technique d'évolution différentielle auto-adaptative (SaDE) :	33
VII.5.	Technique de régression d'apprentissage en profondeur (Deep Learning):	33
VIII.	Les métriques d'évaluation de l'équilibrage de charge:	35
IX.	Conclusion :	36
Chapitre III .	Études de cas AWS (Amazon Web Services) :	37
I.	Introduction :	37
II.	Les solutions et services d'AWS liés à l'équilibrage de charge :	37
II.1.	Solution d'équilibrage de charge avec Elastic Load Balancer (ELB) :	38
II.1.a.	Classic Load Balancer (CLB) :	38
II.1.b.	Application Load Balancer (ALB) :	38
II.1.c.	Network Load Balancer (NLB) :	39
II.2.	Solution d'équilibrage de charge avec Auto Scaling :	40
III.	Conclusion :	40
	Conclusion générale :	41
	Bibliographie:	42

Introduction générale

Le Cloud Computing (appelé en français l'informatique en nuage) apparaît comme une technologie à la croissance très rapide dans le monde des technologies de l'information (IT), qui se concentre sur la fourniture de services informatiques et de ressources informatiques, et à travers le monde à ses utilisateurs sur Internet. L'infrastructure et les services de cloud Computing sous-jacents sont généralement détenus et gérés par un tiers, connu sous le nom de fournisseur de services cloud.

Le principal avantage du Cloud Computing par rapport aux technologies informatiques existantes est le libre-service à la demande, les services de réseau étendus, l'élasticité rapide, la mise en commun des ressources et le service mesuré. La croissance du service cloud peut entraîner un ralentissement du débit, l'utilisation des ressources informatiques et, en fin de compte, réduire l'efficacité du système cloud.

Le cloud Computing est la technologie émergente dans un environnement distribué composé de plusieurs centres de données, serveurs, machines virtuelles, équilibres de charge, qui sont connectés intelligemment. De plus, le nuage traite de nombreuses choses comme le stockage et la récupération de documents, le partage de contenu multimédia et le calcul scientifique.

La gestion des ressources est un problème complexe de l'informatique distribuée, mais elles n'en sont qu'à leurs débuts malgré des recherches exhaustives ces dernières années. L'équilibrage de charge dans l'environnement cloud computing a un impact important sur les performances du cloud. Un bon équilibrage de la charge rend le cloud Computing plus efficace et améliore la satisfaction des utilisateurs.

En plus d'une introduction et d'une conclusion, ce rapport est composé de deux chapitres.

Dans le premier chapitre nous allons présenter l'exploitation des ressources dans le cloud Computing, et donner ses caractéristiques, ses modèles de service, ses modèles de déploiement.

Dans le deuxième chapitre, nous allons aborder la présentation des différentes méthodes d'équilibrage de charge, ainsi que les algorithmes correspondants, qui sont classés en deux approches : statique et dynamique. De plus, nous explorerons les divers algorithmes d'apprentissage automatique utilisés dans le domaine de l'équilibrage de charge.

Dans le troisième chapitre, nous débiterons une étude de cas sur AWS (Amazon Web Services).

Chapitre I . Introduction au Cloud Computing :

I. Introduction :

Depuis quelques années, un nouveau paradigme nommé cloud computing révolutionne la façon dont les entreprises et les particuliers accèdent à des ressources informatiques telles que la puissance de calcul, la capacité de stockage...etc.

Parmi les domaines de recherche qui ont prêté une attention particulière au problème d'allocation de ressources, se trouve l'économie, la recherche opérationnelle et l'informatique. De plus le problème d'allocation de ressources est pertinent à une large gamme d'applications, tel que le commerce électronique, la chaîne d'approvisionnement, les réseaux de capteurs, la composition de service grid/web, le flux de travail, et l'intégration d'application d'entreprise.

Dans ce chapitre, nous donnons une présentation générale sur le Cloud Computing, ensuite nous parlons à l'exploitation des ressources dans le Cloud Computing.

II. Contexte et définition du Cloud Computing :

II.1. Concept du Cloud Computing :

Le concept du cloud computing n'est pas nouveau, ni même excessivement compliqué venant des ressources technologiques ainsi que du point de vue de l'interconnexion. Ce qui est nouveau, c'est le développement et aussi la maturité des méthodes de cloud computing, ainsi que des approches qui permettent les objectifs d'agilité opérationnelle. Rappelant, l'expression clé "calcul de l'énergie" n'a vraiment pas envôûter ou faire le bruit dans le domaine des détails comme l'expression "cloud computing" invite les années en cours. Néanmoins, le respect des sources d'appel commodément est arrivé là et les composants pratiques ou même de service sont ce qui va au centre de l'externalisation de l'accès aux ressources infotech et aussi aux entreprises. Dans cette optique, l'infonuagique illustre une

plateforme de distribution adaptable, économique et testée pour l'entreprise, ainsi que des solutions de détails individuelles sur Internet.¹

II.2. Caractéristiques et avantages du Cloud Computing :

II.2.a. Caractéristiques du Cloud Computing :

Le cloud a 5 caractéristiques essentielles qui sont illustrées dans la figure 1. Si une de ces caractéristiques est absente alors ce n'est pas le cloud :



Figure 1 : Caractéristiques du cloud computing

- ✚ On-demand self-service : libre-service à la demande, l'utilisateur peut s'approvisionner des services, tel que serveur de calcul ou stockage réseau, au besoin automatiquement sans aucune interaction humaine avec le fournisseur du service.
- ✚ Broad network access : Accessible à sur l'ensemble du réseau, les ressources/services sont disponible sur le réseau (local/internet) grâce à des mécanismes standards.
- ✚ Elastic ressource pooling : ressources mutualisées, les ressources sont partagées pour permettre la fourniture des services en parallèle à plusieurs utilisateurs (multitenant model). Ces ressources virtuelles/physiques sont assignées dynamiquement suivant le besoin de l'utilisateur.

¹ (Infrastructural Constraints of Cloud Computing, December 24, 2020)

- ✚ Scalable and Elasticity : Les ressources sont rapidement approvisionnées ou libérées en diverses quantités afin que les systèmes puissent être mis à l'échelle selon les besoins. Pour le consommateur, les ressources semblent être illimitées
- ✚ Measured service : Service mesurable, L'utilisation des ressources peut être surveillée, contrôlée et signalée, ce qui assure la transparence pour le fournisseur et le consommateur du service utilisé.²

II.2.b. Avantages du Cloud Computing :

- ✚ Pas d'investissement initial : Avec le cloud, il est inutile d'investir dans une Infrastructure qui serait très onéreuse à l'achat. L'abonnement étant mensuel vous maîtrisez mieux votre budget et vous ne payez que ce que vous consommez.
- ✚ Gain de temps sur la maintenance : Vous n'avez plus à vous soucier des mises à jour à effectuer, des problématiques de stockages et de performances. Grâce au cloud tout ceci est géré par votre prestataire.
- ✚ L'accessibilité: Les applications et services que vous utilisez dans le Cloud sont accessibles où que vous soyez à partir du moment où vous disposez d'un terminal et d'une connexion internet.
- ✚ Flexibilité : Si vos besoins évoluent il est possible d'adapter votre offre rapidement et simplement. L'utilisateur peut personnaliser les services présents dans son interface.

III. Modèles de service dans le Cloud Computing :

Il existe trois modèles principaux de cloud computing, chaque modèle représente une partie différente de la pile du cloud computing :

² (Cloud Computing : services informatiques dynamiques basés sur le Web - Concepts et notions de base -, 17-Dec-2020)

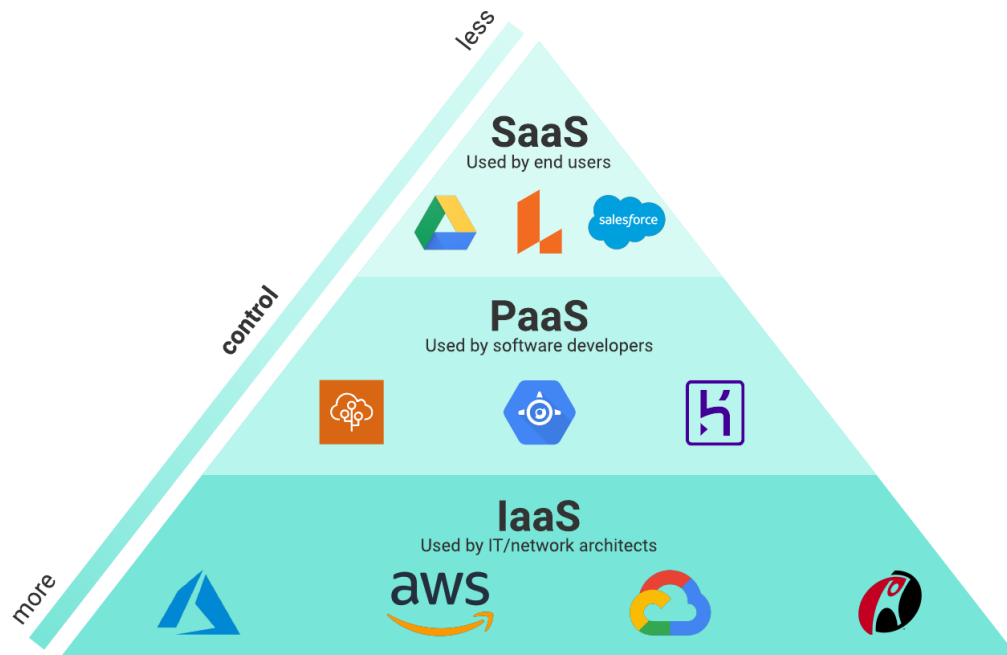


Figure 2 : Les modèles de services de Cloud

III.1. Infrastructure as a Service (IaaS) :

L'utilisateur loue des moyens de calcul et de stockage, des capacités réseau et d'autres ressources indispensables (partage de charge, pare-feu, cache), il a la possibilité de déployer n'importe quel type de logiciel incluant les systèmes d'exploitation. L'utilisateur ne gère pas ou ne contrôle pas l'infrastructure Cloud sous-jacente mais il a le contrôle sur les systèmes d'exploitation, le stockage et les applications. Il peut aussi choisir les caractéristiques principales des équipements réseau comme le partage de charge, les pare-feux, etc.

L'exemple emblématique de ce type de service est Amazon Web Services qui fournit du calcul (EC2), du stockage (S3, EBS), des bases de données en ligne (SimpleDB) et quantité d'autres services de base. Il est maintenant imité par de très nombreux fournisseurs.³

³ (Equilibrage de charge dans les environnement cloud computing., 2019)

III.2. Platform as a Service (PaaS):

C'est une plateforme d'exécution, de déploiement et de développement des applications. La plateforme PaaS regroupe la partie développeur (client) et système (fournisseur) du Cloud Computing. Elle propose des fonctions qui privent le développeur de la gestion des utilisateurs ou des questions de disponibilité par exemple. Le développeur a ainsi uniquement besoin d'héberger son application pour qu'elle soit disponible en SaaS.⁴

III.3. Software as a Service (SaaS):


Ce modèle de service est caractérisé par l'utilisation d'une application partagée qui fonctionne sur une infrastructure Cloud. L'utilisateur accède à l'application par le réseau au travers de divers types de terminaux (souvent via un navigateur web). L'administrateur de l'application ne gère pas et ne contrôle pas l'infrastructure sous-jacente (réseaux, serveurs, applications, stockage). Il ne contrôle pas les fonctions de l'application à l'exception d'un paramétrage de quelques fonctions utilisateurs limitées.

De bons exemples de SaaS sont les logiciels de messagerie au travers d'un navigateur comme Gmail ou Yahoo mail. Ces infrastructures fournissent le service de messagerie à des centaines de millions d'utilisateurs et à des dizaines de millions d'entreprises.⁵

IV. Les acteurs du Cloud Computing :

Le modèle conceptuel du Cloud Computing permet de présenter la cartographie des normes de l'usage du Cloud. Il indique le niveau de maturité et présente une catégorisation des normes concernant les aspects de sécurité, d'interopérabilité et de portabilité des données sur le Cloud.

L'écosystème de la technologie du Cloud est composé principalement de cinq acteurs :

 **Cloud consumer :** Il regroupe l'ensemble des utilisateurs des ressources du Cloud. Ces derniers peuvent être soit des développeurs ou bien des utilisateurs

⁴ (Equilibrage de charge dans les environnement cloud computing., 2019)

⁵ (Equilibrage de charge dans les environnement cloud computing., 2019)

finaux. Ils peuvent être des personnes, des groupes de personnes, des PME, des gouvernements ou des multinationales.

- ✚ **Cloud provider:** Le fournisseur est responsable de fournir les services Cloud, en respectant les caractéristiques. Son rôle est d'allouer les services tout en assurant le niveau de sécurité.
- ✚ **Cloud carrier:** Il est l'acteur principal pour assurer la connectivité des ressources et la liaison entre les autres acteurs. Il est chargé de l'acheminement des données et d'offrir les fonctionnalités avancées dans le réseau.
- ✚ **Cloud broker:** Il joue le rôle d'un intermédiaire qui négocie les relations entre les autres acteurs. Il permet également d'assurer l'arbitrage des services Cloud.
- ✚ **Cloud auditor:** Le rôle de l'auditeur est de s'occuper de la vérification des services, il se charge de l'évaluation des services proposés par le Cloud provider, carrier ou broker. L'objectif est de contrôler la performance et la sécurité des données sur le Cloud, afin de vérifier si les fournisseurs respectent les normes de la charte. La figure suivante présente le diagramme conceptuel de référence.⁶

V. Modèles de déploiement dans le Cloud Computing :

Il existe quatre modèles de déploiement du Cloud communément utilisé : privé, public, et hybride, un modèle additionnel est le Cloud de communauté.

V.1. Cloud public :

L'infrastructure Cloud est ouverte au public ou à de grands groupes industriels, cette infrastructure est possédée par une organisation qui vend des services Cloud, c'est le cas le plus courant, c'est celui de la plate-forme Amazon Web services déjà citée.⁷

V.2. Cloud privé :

Un Cloud privé est un ensemble des services et des ressources disponible pour un seul client par exemple une entreprise ou groupement d'entreprise (appelé organisation), il peut

⁶ (L'émergence du Cloud Computing au service de la transparence des collaborations inter-organisationnelles et de la confiance numérique: revue de littérature., Avril 2022)

⁷ (Equilibrage de charge dans les environnement cloud computing., 2019)

être géré par l'entreprise elle-même, ou ses branches, dans ce cas il s'appelle "Le Cloud privé Interne", en d'autre façons il peut être géré par un prestataire externe loué par l'entreprise, dans ce cas s'appelle "Le Cloud privé Externe", il est accessible via des réseaux sécurisés de type VPN (Virtual Private Network). L'avantage de ce type de Cloud par rapport au Cloud public réside dans l'aspect de la sécurité et la protection des données.⁸

V.3. Cloud hybride :

L'infrastructure cloud hybride est une composition de deux ou plusieurs infrastructures cloud distinctes (privées, communautaires ou publiques) qui restent des entités uniques, mais sont liées par une technologie standardisée ou propriétaire qui permet la portabilité des données et des applications (par exemple, l'éclatement du cloud entre nuages).⁹

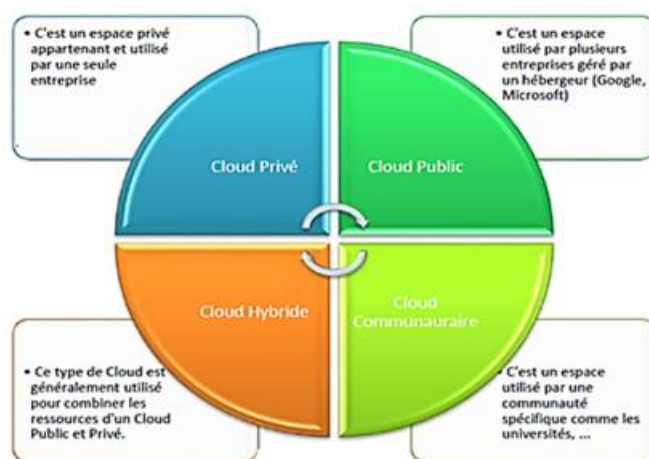


Figure 3 : Les modèles de services de Cloud de déploiement.

V.4. Cloud communautaire :

L'infrastructure Cloud est partagée par plusieurs organisations pour les besoins d'une communauté qui souhaite mettre en commun des moyens (sécurité, conformité, etc..), elle peut être gérée par les organisations ou par une tierce partie et peut être placée dans les locaux ou à l'extérieur.¹⁰

VI. Architecture du Cloud Computing :

⁸ (Equilibrage de charge dans les environnement cloud computing., 2019)

⁹ (Equilibrage de charge dans les environnement cloud computing., 2019)

¹⁰ (Equilibrage de charge dans les environnement cloud computing., 2019)

Les éléments pouvant constituer le système Cloud sont les suivants :

VI.1. La virtualisation:

La virtualisation est la principale technologie dans le Cloud, elle permet une gestion optimisée des ressources matérielles en disposant de plusieurs machines virtuelles sur une machine physique, c'est une technologie qui permet une plus grande modularité dans la répartition des charges et la reconfiguration des serveurs en cas d'évolution ou de défaillance momentanée. Le principe de virtualisation permet d'intégrer les différents serveurs de façons plus flexibles pour faciliter l'utilisation, le but de la virtualisation est de faire la transparence d'utilisation et l'efficacité d'exploitation des ressources, d'assurer le fonctionnement des différents services et la séparation entre de multiple locataire (utilisateurs) impliqués dans un matériel physique.¹¹

VI.2. L'infrastructure :

L'infrastructure informatique du Cloud est un assemblage de serveurs, d'espaces de stockage et de composants réseaux organisés de manière à permettre une croissance incrémentale supérieur à celle que l'on obtient avec les infrastructures classiques. Ces composants doivent être sélectionnés pour leur capacité à répondre aux exigences d'extensibilité, d'efficacité, de robustesse et de sécurité, les serveurs d'entreprise classique ne disposent pas des capacités réseau, de la fiabilité ni des autres qualités nécessaires pour satisfaire efficacement et de manière sécurisé les accords de niveau de service SLA (Service Level Agreement).¹²

VI.3. Le Datacenter :

Un centre de traitement de données, en anglais 'datacenter ' est un site physique sur lequel sont regroupés des équipements constituant le système d'information de l'entreprise (mainframes, serveurs, baies de stockage, équipements réseaux et de la télécommunication, etc.). Il peut être interne ou externe à l'entreprise, exploité ou non avec le soutien de prestataires. Il comprend en général un contrôle sur l'environnement (climatisation, système

¹¹ (Equilibrage de charge dans les environnement cloud computing., 2019)

¹² (Equilibrage de charge dans les environnement cloud computing., 2019)

de prévention contre l'incendie, etc.), une alimentation d'urgence et redondante, ainsi qu'une sécurité physique élevée, des particuliers ou des entreprises peuvent venir y stocker leurs données suivant des modalités bien définies.¹³

VI.4. Cloud Computing et sécurité :

Il y a grandes préoccupations des utilisateurs sur le Cloud Computing est sa sécurité. Dans les Centres de Données Internet (IDC), les fournisseurs de services offrent les grilles et les réseaux seulement, et les appareils restants doivent être préparés par les utilisateurs eux-mêmes, y compris les serveurs, le pare-feu, les logiciels, les périphériques de stockage, etc, la sécurité des utilisateurs peut être réfléchié dans les règles suivantes :

- ✚ La confidentialité des données de stockage des utilisateurs : Le stockage des données d'utilisateur ne peut pas être lues ou modifiées par d'autres personnes (y compris l'opérateur).
- ✚ La confidentialité des données d'utilisateur lors de l'exécution : Les données d'utilisateur ne peuvent pas être lues ou modifiées par d'autres personnes lors de l'exécution (c.à.d. chargée dans le mémoire système).
- ✚ Le secret des données privées d'utilisateur lors du transfert à travers le réseau: Il comprend la sécurité de transfert des données Cloud Computing Internet. Ils ne peuvent pas être affichées ou modifiées par d'autres personnes.
- ✚ Authentification et autorisation nécessaire pour les utilisateurs d'accéder à leurs données: Les utilisateur peuvent accéder efficacement à leurs données et peuvent autoriser d'autres utilisateurs d'y accéder.¹⁴

VII. Conclusion :

Le Cloud Computing est une nouvelle technologie d'utilisation des services informatique, nous pouvons être beaucoup plus flexibles et productif dans l'utilisation des ressources allouées dynamiquement. Le Cloud Computing va continuer à évoluer comme le

¹³ (Equilibrage de charge dans les environnement cloud computing., 2019)

¹⁴ (Equilibrage de charge dans les environnement cloud computing., 2019)

fondement de l'Internet du futur, ou nous seront interconnectés dans un réseau de contenus et des services.

Chapitre II . Équilibrage de charge dans le cloud computing :

I. Introduction :

L'équilibrage de charge joue un rôle crucial dans les systèmes informatiques modernes, permettant de distribuer efficacement la charge de travail entre les différentes ressources disponibles. Dans ce chapitre, nous aborderons en détail les méthodes d'équilibrage de charge, ainsi que les algorithmes qui les accompagnent.

II. Définition et importance de l'équilibrage de charge :

II.1. Définition :

L'équilibrage de la charge de travail consiste à améliorer les performances du système en déplaçant la charge de travail entre les processeurs. La charge de travail d'une machine correspond au temps de traitement total nécessaire pour exécuter toutes les tâches assignées à la machine. Équilibrer la charge des machines virtuelles de manière uniforme signifie qu'aucune des machines disponibles n'est inactive ou partiellement chargée tandis que d'autres sont fortement chargées. L'équilibrage de la charge est l'un des facteurs importants pour améliorer les performances de travail du fournisseur de services cloud.¹⁵

II.2. Importance de l'équilibrage de charge :

Dans les environnements à fort trafic, l'équilibrage de charge est ce qui permet aux demandes des utilisateurs de se dérouler de manière fluide et précise. Ils épargnent aux utilisateurs la frustration de se quereller avec des applications et des ressources qui ne répondent pas.

¹⁵ (Load Balancing in cloud computing, 2019)

L'équilibrage de charge joue également un rôle clé dans la prévention des temps d'arrêt et la simplification de la sécurité, réduisant ainsi la probabilité de perte de productivité et de profits.

Les autres avantages de l'équilibrage de charge sont les suivants :

- **Flexibilité:** En plus de diriger le trafic pour maximiser l'efficacité, l'équilibrage de charge offre la flexibilité nécessaire pour ajouter et supprimer des serveurs en fonction de la demande. Il permet également d'effectuer la maintenance du serveur sans causer de perturbation pour les utilisateurs puisque le trafic est redirigé vers d'autres serveurs pendant la maintenance.
- **Évolutivité :** Au fur et à mesure que l'utilisation d'une application ou d'un site Web augmente, l'augmentation du trafic peut nuire à ses performances si elle n'est pas gérée correctement. Avec l'équilibrage de charge, vous avez la possibilité d'ajouter un serveur physique ou virtuel pour répondre à la demande sans provoquer d'interruption de service. Au fur et à mesure que de nouveaux serveurs sont mis en ligne, l'équilibreur de charge les reconnaît et les inclut de manière transparente dans le processus. Cette approche est préférable au déplacement d'un site Web d'un serveur surchargé vers un nouveau, ce qui nécessite souvent un certain temps d'arrêt.
- **Redondance:** Lors de la répartition du trafic sur un groupe de serveurs, l'équilibrage de charge fournit une redondance intégrée. Si un serveur tombe en panne, vous pouvez rediriger automatiquement la charge vers des serveurs en état de marche afin de minimiser l'impact sur les utilisateurs.¹⁶
- **Utilisation efficace des ressources :** En distribuant la charge de manière équilibrée, l'équilibrage de charge assure une utilisation optimale des ressources disponibles. Plutôt que de surcharger certains serveurs tout en laissant d'autres sous-utilisés, l'équilibrage de charge permet d'exploiter pleinement les capacités de chaque serveur, ce qui maximise l'efficacité globale du système.

¹⁶ (IBM)

III. Types d'équilibrage de charge dans le Cloud Computing :

On peut classer l'équilibrage de charge en trois catégories principales en fonctions de ce que l'équilibreur de charge vérifie dans la demande du client pour rediriger le trafic.

III.1. Équilibrage de charge de niveau réseau :

Les Network Load Balancers examinent les adresses IP et d'autres informations réseau pour rediriger le trafic vers les ressources de manière optimale. Ils suivent la source du trafic des applications et peuvent attribuer une adresse IP statique à plusieurs serveurs.¹⁷

III.2. Équilibrage de charge de niveau application :

Les applications modernes complexes disposent de plusieurs batteries de serveurs avec plusieurs serveurs dédiés à une seule fonction applicative. Les Application Load Balancers examinent le contenu de la demande, tel que les en-têtes HTTP ou les ID de session SSL, pour rediriger le trafic.

III.3. Équilibrage de charge de niveau global :

L'équilibrage de charge global du serveur s'effectue sur plusieurs serveurs répartis géographiquement. Par exemple, les entreprises peuvent avoir des serveurs dans plusieurs centres de données, dans différents pays et chez des fournisseurs de cloud tiers dans le monde entier. Dans ce cas, les équilibreurs de charge locaux gèrent la charge des applications au sein d'une région ou d'une zone. Ils essaient de rediriger le trafic vers une destination de serveur géographiquement plus proche du client. Ils peuvent rediriger le trafic vers des serveurs situés en dehors de la zone géographique du client uniquement en cas de défaillance du serveur.

IV. Les Technologies d'équilibrage de charge :

¹⁷ (IBM, s.d.)

Avant de parler des différentes technologies d'équilibrage de charge nous devons définir c'est quoi un équilibreur de charge.

Un équilibreur de charge dans le cloud également sous le nom de « Load Balancer » en anglais est un composant essentiel dans les infrastructures cloud qui permet de distribuer le trafic entrant de manière équilibré entre plusieurs serveurs, instances ou ressources.

Les équilibreurs de charges sont de deux types qui sont Hardware communément appelé méthode traditionnel et Software.

IV.1. Hardware Load balancing :

Les équilibreurs de charge matériels sont des appareils physiques, tels qu'un appareil spécifique. Ils dirigent le trafic vers les serveurs en se basant sur des critères tels que le nombre de connexions existantes vers un serveur, l'utilisation du processeur et les performances du serveur.

Les équilibreurs de charge matériels comprennent un micrologiciel propriétaire qui nécessite une maintenance et des mises à jour à mesure que de nouvelles versions et correctifs de sécurité sont publiés. Les équilibreurs de charge matériels offrent généralement de meilleures performances et un meilleur contrôle, tout en proposant une gamme complète de fonctionnalités. Cependant, ils nécessitent un certain niveau de compétence pour une gestion et une maintenance appropriée. Étant donné qu'ils sont basés sur du matériel, ces équilibreurs de charge sont moins flexibles et moins évolutifs, ce qui conduit souvent à une sur provision des équilibreurs de charge matériels.

IV.2. Software Load Balancing :

Les équilibreurs de charge logiciels sont généralement plus faciles à déployer que les versions matérielles. Ils sont également plus rentables et flexibles, et sont utilisés en conjonction avec des environnements de développement logiciel. L'approche logicielle vous donne la flexibilité de configurer l'équilibreur de charge en fonction des besoins spécifiques de votre environnement. Cependant, cette flexibilité accrue peut nécessiter un peu plus de travail pour mettre en place l'équilibreur de charge. Comparés aux versions matérielles qui offrent une approche plus fermée, les équilibreurs de charge logiciels vous donnent plus de liberté pour effectuer des changements et des mises à niveau.

Les équilibreurs de charge logiciels peuvent prendre la forme de machines virtuelles (VM) préconfigurées. Les VM vous épargnent une partie du travail de configuration, mais elles peuvent ne pas offrir toutes les fonctionnalités disponibles avec les versions matérielles.

Les équilibreurs de charge logiciels sont disponibles soit sous forme de solutions installables nécessitant une configuration et une gestion, soit sous forme de service cloud - **Load Balancer as a Service (LBaaS)**. En choisissant cette dernière option, vous êtes dispensé de la maintenance, de la gestion et de la mise à niveau régulières des serveurs installés localement ; le fournisseur de services cloud s'occupe de ces tâches.¹⁸

V. Approches et algorithmes d'équilibrage de charge :

V.1. Approches d'équilibrage de charge dans le cloud :

Dans l'équilibrage de charge nous avons principalement deux approches : **Statique et Dynamique**.

V.1.a. Approche statique :

Dans l'approche de répartition de charge statique, la décision de déplacer la charge ne dépend pas de l'état actuel du système. Elle nécessite une connaissance des applications et des ressources du système.

Les algorithmes de répartition de charge statique ne sont pas préemptifs, ce qui signifie que chaque machine a au moins une tâche lui étant assignée. L'objectif est de minimiser le temps d'exécution de la tâche et de limiter les surcharges et les retards de communication. Cependant, cet algorithme présente un inconvénient majeur : la tâche est assignée aux processeurs ou aux machines uniquement après sa création et cette tâche ne peut pas être déplacée vers une autre machine une fois qu'elle lui a été attribuée.

¹⁸ (IBM, s.d.)

Cela signifie que l'algorithme de répartition de charge statique manque de flexibilité, ce qui peut entraîner une répartition non optimale des charges et une utilisation inefficace des ressources.

Les algorithmes de répartition de charge statique reposent généralement sur une connaissance préalable du système, telle que les caractéristiques de performance des applications et des ressources, pour prendre des décisions de répartition de charge. L'orchestrateur, qui agit comme coordinateur central, attribue les tâches aux processeurs en fonction de leurs capacités de performance. L'objectif est de minimiser le temps d'exécution des tâches et de réduire les surcharges de communication et les retards.

L'un des inconvénients des algorithmes de répartition de charge statique est leur caractère non préemptif. Une fois qu'une tâche est attribuée à une machine, elle ne peut pas être déplacée ou migrée vers une autre machine, même si cela devient nécessaire en raison de l'évolution de la charge de travail ou de la disponibilité des ressources. Ce manque d'adaptabilité peut entraîner des déséquilibres de charge, où certaines machines sont sous-utilisées tandis que d'autres sont surchargées.¹⁹²⁰

V.1.b. Approche dynamique :

Dans ce type d'algorithme de répartition de charge, l'état actuel du système est utilisé pour prendre des décisions de répartition de charge, ce qui signifie que le déplacement de la charge dépend de l'état actuel du système. Cela permet aux processus de passer dynamiquement d'une machine surchargée à une machine sous-utilisée, ce qui accélère leur exécution. Cela signifie qu'il permet la préemption des processus, ce qui n'est pas pris en charge dans l'approche de répartition de charge statique.

Un avantage important de cette approche est que sa décision de répartition de charge est basée sur l'état actuel du système, ce qui contribue à améliorer les performances globales du système en migrant dynamiquement la charge. En d'autres termes, le système est capable de détecter les machines surchargées et les machines sous-utilisées en temps réel, et de déplacer

¹⁹ (Yeluri, Load balancing in cloud computing, 2019)

²⁰ (Slimane, 2019)

les processus vers les machines sous-utilisées pour rééquilibrer la charge de manière dynamique.

Prenons exemple, une machine virtuelle surchargée et a du mal à traiter toutes les tâches qui lui sont assignées, l'algorithme de répartition de charge dynamique peut décider de changer la place de certains processus vers une autre machine moins sollicitée. Ainsi on pourra répartir les charges de travail de manière plus équilibrée et d'améliorer les performances globales du système.

Les algorithmes de répartition de charge dynamique peuvent être **centralisés** ou **distribués**, en fonction de la question de savoir si la responsabilité de la planification globale²¹ de la dynamique des tâches doit résider dans un processeur unique (centralisé) ou si le travail nécessaire pour prendre des décisions doit être réparti physiquement entre les processeurs.

1. Approche centralisée:

Dans une approche centralisée, un composant centralisé, tel qu'un gestionnaire de ressources ou un Load balancer central, est responsable de la répartition de la charge dans le système cloud. Ce composant central collecte des informations sur l'état et la charge des ressources disponibles, analyse ces informations et prend des décisions de répartition de charge pour équilibrer la charge entre les ressources. Il peut utiliser des algorithmes de répartition de charge dynamique pour déplacer les tâches ou les requêtes vers les ressources moins sollicitées et optimiser les performances globales du système.²²

2. Approche distribuée :

Dans l'approche distribuée, chaque nœud construit individuellement son propre vecteur de charge, qui rassemble les données de charge des autres nœuds. Les décisions de répartition de charge sont prises localement en utilisant ces vecteurs de charge locaux. Cette approche est plus adaptée aux systèmes généralement distribués, tels que le cloud computing.

²¹ (Misra, 2022)

²² (Dr Syeda Gauhar Fatima, 2019)

Alors l'approche distribuée peut elle aussi être divisé en deux types : coopératives et non coopératives.²³

a. Coopératives :

Dans l'approche coopérative, les nœuds ou les processus travaillent ensemble de manière collaborative pour prendre des décisions de répartition de charge. Ils échangent des informations sur leur charge de travail, leurs performances, la disponibilité des ressources, etc. afin de parvenir à un consensus sur la meilleure répartition de charge. Les nœuds peuvent coordonner leurs actions et ajuster leur charge de travail en conséquence pour atteindre un équilibre global.

Cette approche coopérative permet une prise de décision plus globale et plus équilibrée, car elle prend en compte les informations de plusieurs nœuds pour parvenir à un consensus. Elle favorise la répartition équitable de la charge et peut conduire à une meilleure utilisation des ressources. Des protocoles de communication et de coordination spécifiques sont utilisés pour faciliter cette collaboration entre les nœuds.

b. Non-Coopératives :

Dans l'approche non-coopérative, les nœuds ou les processus prennent des décisions de manière concurrentielle, sans nécessairement coopérer ou échanger d'informations entre eux. Chaque nœud prend des décisions de répartition de charge en fonction de ses propres critères locaux, tels que sa charge de travail, ses performances, etc. Il peut s'agir d'une approche plus individualiste où chaque nœud cherche à optimiser ses propres objectifs locaux sans tenir compte de l'effet global sur le système.

Cette approche non coopérative peut conduire à des décisions de répartition de charge plus locales et potentiellement moins équilibrées. Elle peut être utilisée dans des environnements où la coordination entre les nœuds est difficile ou coûteuse, ou lorsque les nœuds ont des intérêts divergents.²⁴

3. Approche source-initiative vs receiver-initiative :

²³ (Misra, 2022)

²⁴ (A Review on Different Load Balancing Techniques in Cloud Computing Environments, 2015)

L'approche source-initiative et l'approche receveur-initiative sont deux stratégies différentes dans le contexte de l'équilibrage de charge dans le cloud. Elles décrivent la façon dont la décision de répartition de la charge est prise, que ce soit par le client (source) ou par le système d'équilibrage de charge (receveur).

Dans l'approche source-initiative, comme expliqué précédemment, c'est le client ou la source de la demande qui prend l'initiative de choisir le serveur sur lequel il enverra sa requête. Le client évalue les différentes options disponibles en fonction de ses critères (latence, charge, disponibilité, etc.) et prend une décision autonome sur le serveur à utiliser. Cela permet au client d'avoir un certain contrôle sur la répartition de la charge, mais cela implique également une plus grande responsabilité et complexité du côté du client.

En revanche, dans l'approche receveur-initiative, c'est le système d'équilibrage de charge qui prend l'initiative de choisir le serveur approprié pour traiter la demande. Le système d'équilibrage de charge évalue les différentes métriques et critères de performance des serveurs disponibles, tels que la charge actuelle, la disponibilité des ressources, la latence, etc. Il sélectionne ensuite le serveur le mieux adapté pour traiter la demande et redirige la demande vers ce serveur. Cela décharge la responsabilité de la décision de répartition de la charge du client vers le système d'équilibrage de charge.

L'approche receiver-initiative offre une plus grande simplicité car il n'a pas à prendre de décisions complexes concernant la répartition de la charge. Cependant, cela signifie également que le système d'équilibrage de charge doit disposer d'une bonne connaissance des métriques et des performances des serveurs pour prendre des décisions éclairées.

V.2. Les algorithmes de Load balancing :

L'équilibrage de charge est un problème important et complexe dans le cloud computing. L'équilibrage de charge dans le cloud computing contribue à une utilisation efficace des ressources, à des temps de réponse réduits, à une répartition équitable de la charge et à une consommation d'énergie réduite, permettant ainsi aux services d'atteindre une utilisation

complète des ressources. Plusieurs algorithmes ont été proposés pour l'équilibrage de charge dans le cloud computing, dont certains sont dynamique et d'autres statiques.²⁵

V.2.a. Algorithmes statiques :

- **Round Robin :** Il s'agit d'un algorithme d'équilibrage de charge fixe qui ne tient pas compte de la charge précédente d'un nœud au moment de l'attribution des tâches. Il utilise l'algorithme d'ordonnancement de type round robin pour l'attribution des tâches. Il choisit le premier nœud de manière aléatoire, puis attribue les tâches à tous les autres nœuds de manière circulaire. Cet algorithme ne conviendra pas à l'informatique en nuage car certains nœuds peuvent être fortement chargés tandis que d'autres ne le sont pas. Étant donné que le temps d'exécution de tout processus n'est pas connu avant l'exécution, il est possible que certains nœuds soient profondément chargés.²⁶
- **Weighted Round Robin :** L'algorithme du round-robin pondéré a été proposé pour résoudre ce problème. Dans cet algorithme, chaque nœud se voit attribuer un poids spécifique. En fonction du poids attribué à chaque nœud, il recevra un nombre approprié de requêtes. Si le poids attribué à tous les nœuds est égal, alors chaque nœud recevra un trafic équivalent. Dans un système informatique en nuage, il n'est pas possible de prédire précisément le temps d'exécution d'une tâche. Par conséquent, cet algorithme n'est pas préféré.

L'algorithme du round-robin pondéré permet de répartir la charge de manière plus équilibrée entre les nœuds en fonction de leur poids respectif. Les nœuds ayant un poids plus élevé recevront une part plus importante du trafic, tandis que les nœuds ayant un poids plus faible recevront une part moins importante. Cela permet de mieux gérer les différences de performances entre les nœuds et d'optimiser l'utilisation des ressources disponibles.

Cependant, dans les systèmes de cloud computing où la prédiction précise du temps d'exécution n'est pas possible, cet algorithme peut ne pas être préféré. Étant donné

²⁵ (cloudfare, s.d.)

²⁶ (Semwal, 2021)

que le temps d'exécution des tâches n'est pas connu à l'avance, attribuer un poids approprié à chaque nœud peut être difficile. De plus, les charges de travail peuvent varier dynamiquement, ce qui rend la répartition de la charge basée uniquement sur les poids des nœuds moins efficace.

- **Min-Min** : L'algorithme Min-Min [28, 29] commence avec un ensemble de toutes les tâches non attribuées. Tout d'abord, le temps de complétion minimum pour toutes les tâches est calculé. La tâche avec le plus court temps de complétion est sélectionnée. Ensuite, le nœud qui a le plus court temps de complétion pour toutes les tâches est choisi. Enfin, le nœud sélectionné et la tâche sélectionnée sont assignés. Le temps de disponibilité du nœud est mis à jour. Ce processus est répété jusqu'à ce que toutes les tâches non attribuées soient assignées. L'avantage de cet algorithme est que la tâche avec le plus court temps d'exécution est exécutée en priorité.

Cependant, l'inconvénient de cet algorithme est que certaines tâches peuvent souffrir, c'est-à-dire qu'elles peuvent être constamment reléguées en arrière-plan et ne pas être exécutées pendant longtemps. Cela peut se produire si les tâches avec de plus longs temps de complétion sont constamment sélectionnées en dernier, ce qui peut entraîner un déséquilibre et une utilisation inefficace des ressources. Les tâches avec des temps d'exécution plus longs peuvent devoir attendre pendant une période prolongée, ce qui peut entraîner des retards globaux dans l'exécution des tâches.

- **Hachage IP** : Combine les adresses IP source et destination du trafic entrant et utilise une fonction mathématique pour les convertir en un hachage. Sur la base de ce hachage, la connexion est attribuée à un serveur spécifique.²⁷

V.2.b. Algorithme dynamique

- **Least connection** : La répartition de charge "Least Connection" est un algorithme dynamique de répartition de charge où les requêtes des clients sont réparties vers le serveur d'application ayant le moins de connexions actives au moment où la

²⁷ (Misra, 2022)

requête du client est reçue. Dans les cas où les serveurs d'application ont des spécifications similaires, un serveur peut être surchargé en raison de connexions plus longues ; cet algorithme prend en compte la charge des connexions actives. Cette technique est plus appropriée pour les requêtes entrantes qui ont des durées de connexion variables et un ensemble de serveurs relativement similaires en termes de puissance de traitement et de ressources disponibles.

Pour maintenir un équilibre de charge continu, il est recommandé de mettre en place des mécanismes de rééquilibrage périodique. Cela peut inclure la mise à jour des poids des serveurs en fonction de leurs performances actuelles, la redistribution des connexions entre les serveurs si la charge devient déséquilibrée, ou d'autres stratégies visant à optimiser la répartition de la charge.

- **Weighted Least connection** : Donne aux administrateurs la possibilité d'attribuer des poids différents à chaque serveur, en partant du principe que certains serveurs peuvent gérer plus de connexions que d'autres.
- **Temps de réponse pondéré** : Il calcule la moyenne du temps de réponse de chaque serveur et la combine avec le nombre de connexions ouvertes sur chaque serveur pour déterminer où envoyer le trafic. En envoyant le trafic vers les serveurs dont le temps de réponse est le plus rapide, l'algorithme garantit un service plus rapide aux utilisateurs.²⁸

Les étapes de son fonctionnement :

1. Collecte des temps de réponses : pour chaque requête ou utilisateur, le temps de réponse est enregistré, celle-ci représente la durée écoulée entre l'envoi de la requête et la réception de la réponse correspondante.
2. Attribution des poids : Pour chaque requête ou utilisateur, le temps de réponse est multiplié par son poids correspondant pour obtenir le temps de réponse pondéré. Par exemple, si une requête a un poids de 2 et un temps

²⁸ (cloudfare, s.d.)

de réponse de 5 secondes, le temps de réponse pondéré pour cette requête serait de 10 secondes.

3. **Sélection du serveur** : L'algorithme d'équilibrage de charge utilise les temps de réponse pondérés pour prendre des décisions sur la répartition des requêtes entre les serveurs disponibles. Il peut sélectionner le serveur ayant le temps de réponse pondéré le plus faible pour la prochaine requête à traiter.
 4. **Rééquilibrage périodique** : Pour maintenir un équilibre de charge optimal, il est souvent nécessaire de réaliser un rééquilibrage périodique. Cela peut inclure le recalcul des poids en fonction de la charge actuelle des serveurs, la redistribution des requêtes entre les serveurs si la charge devient déséquilibrée, ou d'autres stratégies pour optimiser la répartition de la charge.
- **Throttled Load Balancing** : c'est une méthode d'équilibrage de charge qui vise à limiter le débit ou la quantité de trafic redirigé vers un serveur ou une ressource spécifique afin de prévenir la surcharge ou de garantir une répartition équilibrée de la charge. L'idée principale derrière le throttled load balancing est de contrôler le flux de requêtes ou de données qui sont dirigées vers un serveur ou une ressource en particulier, en fonction de sa capacité à les traiter. Cela permet de prévenir les problèmes de performances, les temps de réponse lents ou les pannes dues à une surcharge excessive.²⁹
1. **Surveillance de la charge** : Le système de throttled load balancing surveille en permanence la charge sur les serveurs ou les ressources afin de détecter les niveaux de saturation ou de surcharge potentiels. Cela peut se faire en collectant des métriques telles que l'utilisation du processeur, la mémoire, la bande passante réseau, etc.
 2. **Limite de débit** : Lorsque la charge sur un serveur ou une ressource dépasse un seuil prédéfini, le throttled load balancing limite le débit de nouvelles requêtes ou

²⁹ (Amrutanshu Panigrahi)

de données qui lui sont redirigées. Cela peut être réalisé en ralentissant ou en rejetant certaines requêtes, en ajustant les priorités ou en mettant en file d'attente les demandes pour une exécution ultérieure.

3. **Stratégies de throttling** : Différentes stratégies peuvent être mises en place pour gérer le throttling de la charge. Par exemple, on peut décider de ralentir progressivement le débit lorsque la charge atteint un certain seuil, de rediriger les requêtes vers d'autres serveurs moins chargés, ou d'utiliser des mécanismes de mise en attente et de rééquilibrage de la charge.
4. **Rétablissement progressif** : Une fois que la charge sur un serveur ou une ressource diminue en dessous d'un certain seuil, le throttled load balancing peut rétablir progressivement le débit normal des requêtes ou des données. Cela permet au serveur ou à la ressource de récupérer et de traiter les demandes en temps opportun.

VI. Migration des VM dans l'équilibrage de charge

Dans les environnements de cloud computing, l'équilibrage de charge joue un rôle crucial pour garantir une utilisation optimale des ressources et une distribution équitable des charges de travail sur les serveurs physiques. L'une des techniques clés utilisées dans l'équilibrage de charge est la migration des machines virtuelles (VM). La migration des VM permet de déplacer dynamiquement les instances de VM d'un serveur physique à un autre afin de répartir efficacement la charge et d'optimiser les performances du système.

La migration des VM offre de nombreux avantages dans le contexte de l'équilibrage de charge. Elle permet de réduire les goulots d'étranglement et de prévenir la surcharge des serveurs en redistribuant les charges de travail de manière proactive. Lorsqu'un serveur devient surchargé, les VM peuvent être migrées vers des serveurs moins sollicités, permettant ainsi une meilleure utilisation des ressources disponibles.

L'objectif de la migration des VM dans l'équilibrage de charge est de garantir une répartition équilibrée des charges de travail, d'améliorer les performances globales du système, de réduire les temps de réponse et d'optimiser l'utilisation des ressources. Cette approche dynamique permet au système de s'adapter aux variations de charge en temps réel,

assurant ainsi une meilleure expérience utilisateur et une utilisation plus efficace des ressources du cloud.

Au cours de cette migration, plusieurs aspects doivent être pris en compte, tels que la minimisation des interruptions de service, la gestion des ressources réseau, la synchronisation de l'état des VM migrées, et l'optimisation des coûts liés à la migration. Différentes techniques et algorithmes sont utilisés pour déterminer le moment opportun pour migrer une VM, le serveur de destination approprié, ainsi que les méthodes de transfert des données de la VM migrée.

La migration des machines virtuelles peut être réalisée de différentes manières selon les techniques et les outils disponibles dans l'environnement de cloud computing. Voici quelques approches couramment utilisées :

- **Migration en temps réel (live migration) :** Cette technique permet de déplacer une VM d'un serveur vers un autre sans interruption de service pour les applications en cours d'exécution. Elle nécessite généralement une synchronisation en temps réel des états mémoire entre les serveurs source et de destination pour assurer la continuité de l'exécution.
- **Migration à froid (cold migration):** Contrairement à la migration en temps réel, cette technique implique l'arrêt temporaire de la VM pendant le processus de migration. Les états mémoire de la VM sont transférés vers le serveur de destination, puis la VM est redémarrée sur ce dernier. Cette technique peut être utilisée lorsque des interruptions temporaires de service sont acceptables.
- **Migration prédictive :** Cette approche utilise des techniques d'apprentissage automatique et d'analyse prédictive pour anticiper les pics de charge et planifier la migration des VM en conséquence. Les modèles prédictifs sont utilisés pour estimer les besoins en ressources et décider des migrations à effectuer afin d'optimiser les performances du système.

VII. Algorithmes de l'apprentissage automatique utiliser dans le cloud computing :

Les techniques d'équilibrage de charge sont classées différemment selon les différentes caractéristiques. Ils jouent un rôle majeur dans l'utilisation des ressources du serveur. Les équilibreurs de charge sont construits sur certains aspects de l'environnement cloud comme le ressource CPU et mémoire du serveur, accords de niveau de service (SLA), prévision de la congestion du réseau, qualité de service (QoS), estimation du temps de réponse du service et demande de stockage dans le cloud.

L'apprentissage automatique fait partie de l'intelligence artificielle qui se concentre sur la formation de systèmes pour effectuer de nouvelles tâches sans être explicitement programmés. Les données historiques et les techniques statistiques sont combinées par un processus appelé formation pour construire des modèles qui peuvent prévoir de nouvelles valeurs invisibles. L'apprentissage en profondeur est un sous-ensemble de l'apprentissage automatique qui utilise des variations de réseaux de neurones avec des réseaux plus profonds et de grands ensembles de données. L'apprentissage en profondeur combine l'extraction de caractéristiques et la prédiction dans un réseau profond au sein de couches cachées. Il réussit mieux performance que les problèmes d'apprentissage automatique traditionnels.

Pour la collecte des données nécessaires pour l'entraînement des modèles une approche est utilisée s'appelant **Active Monitoring Load Balancing (AMLB)**.

L'AMLB (Active Monitoring Load Balancing) marche en installant des agents sur chaque nœud du système dans le cloud. Ces agents sont des composants matériels qui collectent les données sur le fonctionnement du nœud, ses performances, son utilisation de ressources, etc...

Les agents ou sondes peuvent être configurés pour collecter les données de manière continue ou à intervalles réguliers. Ils enregistrent ces données localement sur chaque nœud ou les transmettent à un système centralisé de collecte et d'analyse des données.

L'avantage de cette approche est qu'elle permet de collecter des données spécifiques à chaque nœud, ce qui offre une vue détaillée du fonctionnement et des performances

individuelles. Cela permet également de détecter rapidement les problèmes potentiels, tels que des goulots d'étranglement, des surcharges ou des erreurs sur un nœud particulier.

VII.1. Technique basée sur la régression, Random Forest et AdaBoost :

Le modèle de distribution de charge basé sur l'apprentissage automatique était composé de plusieurs modèles, à savoir la régression linéaire multiple (MLR), la forêt aléatoire (RF) et AdaBoost (Ada) ont été utilisés pour déterminer où chaque requête doit être traitée en fonction du temps d'exécution du CPU et GPU. Cette technique a abordé l'hétérogénéité architecturale en tenant compte de la différence entre les unités de traitement et leurs caractéristiques de performance associées. Son objectif principal était la distribution des transactions des systèmes de gestion de bases de données distribuées.³⁰

VII.2. Support Vector Machines (SVM) et technique K-suggest :

Lilhore et d'autres ont proposé une solution d'équilibrage de charge basée sur plusieurs algorithmes d'apprentissage automatique tels que SVM et l'outil de clustering K-suggest. Le clustering est utilisé pour établir des groupes de machines virtuelles qui sont dérivés de l'utilisation du CPU et de la mémoire principale (RAM). Cette technique a partagé les actifs avec différents groupes et les machines virtuelles également. Ensuite, il a utilisé le mappage d'aide dynamique pour affecter les charges à leurs groupes de machines virtuelles appropriés en fonction de leurs tailles, c'est-à-dire : machines virtuelles normales, inactives, sous-chargées et surchargées. Le mappage des ressources impliquait le mappage des tâches groupées avec le groupe de machines virtuelles approprié. Cette méthode a amélioré la qualité de service et réduit le temps d'attente ou de rejet global.³¹

VII.3. Approche du réseau de neurones artificiels propagés (BPANN) :

BPANN a été utilisé sur un équilibreur de charge basé sur un agent dynamique qui a été proposé par Prakash et Lakshmi sur le réseau défini par logiciel (SDN). Le SDN est un

³⁰ (Machine Learning Load Balancing Techniques in, 2022)

³¹ (Machine Learning Load Balancing Techniques in, 2022)

composant de l'architecture cloud qui est visible à l'échelle mondiale. Dans le cadre de l'allégement de la charge de travail au sein de la charge, ils sont chargés de migrer les machines virtuelles au sein du centre de données.

L'algorithme BPANN a été formé sur les données de charge et de migration des machines virtuelles. Le modèle résultant a été utilisé pour prédire la charge de la machine virtuelle. La charge projetée a ensuite été utilisée pour déterminer la migration de la machine virtuelle. Une migration efficace des machines virtuelles améliore l'efficacité du réseau et le taux de migration des données. La vitesse de traitement a été considérablement réduite car les charges lourdes sont adaptées aux machines virtuelles sous-utilisées.³²

VII.4. ANN et technique d'évolution différentielle auto-adaptative (SaDE) :

Cette technique a été développée par Kumar et d'autres pour prédire la charge de travail au sein du centre de données cloud. Cette approche combine le réseau de neurones artificiels (ANN) et l'évolution différentielle auto-adaptative (SaDE). Les demandes des utilisateurs ont été regroupées dans des unités de temps qui ont été utilisées comme données historiques. La partie ANN a été formée avec les charges de travail réelles et les données historiques. Le modèle résultant a été utilisé pour prévoir les travaux à venir dans le centre de données. Le modèle a été formé sur des ensembles de données provenant de serveurs de la NASA et de la Saskatchewan.³³

VII.5. Technique de régression d'apprentissage en profondeur (Deep Learning):

La régression basée sur l'apprentissage en profondeur a été utilisée pour prédire le calendrier continu des tâches à partir du temps et du coût de calcul par Kaur et d'autres. Le réseau d'apprentissage en profondeur a été conçu pour avoir 3 couches cachées de réseaux de neurones convolutifs, une couche de mutualisation et la couche d'activation constituée de la fonction ReLU. Les données de formation étaient composées des données des paramètres de temps et de coût de flux de travail plus importants.³⁴

³² (Machine Learning Load Balancing Techniques in, 2022)

³³ (Machine Learning Load Balancing Techniques in, 2022)

³⁴ (Machine Learning Load Balancing Techniques in, 2022)

Modèle de Machine/ Deep Learning	Données utilisées	Problème d'équilibrage de charge résolu
Technique basée sur la régression, Random Forest et AdaBoost	Données de requêtes de base de données	Cloud Hétérogénéité du CPU et du GPU
Support Vector Machines (SVM) et technique K-suggest	Données d'utilisation de la RAM et du processeur	Utilisation des ressources des machines virtuelles et réduction du temps d'exécution
Approche du réseau de neurones artificiels propagés (BPANN)	Journaux de trafic réseau	Migration de VM, migration de données
ANN et technique d'évolution différentielle auto-adaptative (SaDE)	Requête client regroupée en unités de temps	Répartition des charges de travail
Technique de régression d'apprentissage en profondeur	Données de workflow de tâches	Utilisation et débit des ressources de qualité de service (QoS)

VIII. Les métriques d'évaluation de l'équilibrage de charge

Les métriques qualitatives ou paramètres considérés importants pour l'équilibrage de charge dans le cloud computing sont les suivants :

- **Débit** : Le nombre total de tâches qui ont terminé leur exécution est appelé débit. Un débit élevé est nécessaire pour une meilleure performance du système.
- **Surcharge associée** : La quantité de surcharge générée par l'exécution de l'algorithme d'équilibrage de charge. Une surcharge minimale est attendue pour une mise en œuvre réussie de l'algorithme.
- **Tolérance aux pannes** : Il s'agit de la capacité de l'algorithme à fonctionner correctement et uniformément même en cas de défaillance d'un nœud arbitraire du système.
- **Temps de migration** : Le temps nécessaire pour la migration ou le transfert d'une tâche d'une machine à une autre dans le système. Ce temps doit être réduit au minimum pour améliorer les performances du système.
- **Temps de réponse** : Il s'agit du temps minimum qu'un système distribué exécutant un algorithme d'équilibrage de charge met pour répondre.
- **Utilisation des ressources** : Il s'agit du degré d'utilisation des ressources du système. Un bon algorithme d'équilibrage de charge permet une utilisation maximale des ressources.
- **Scalabilité** : Il détermine la capacité du système à effectuer un algorithme d'équilibrage de charge avec un nombre limité de processeurs ou de machines.
- **Performance** : Elle représente l'efficacité du système après l'exécution de l'équilibrage de charge. Si tous les paramètres ci-dessus sont optimisés, cela améliorera grandement les performances du système.

IX. Conclusion :

À la fin de ce chapitre, nous aurons acquis une compréhension approfondie des principes fondamentaux des méthodes d'équilibrage de charge, ainsi que des approches statique et dynamique qui les sous-tendent. De plus, nous aurons exploré en détail les différents algorithmes d'apprentissage automatique utilisés pour améliorer l'efficacité de la répartition de charge dans les systèmes informatiques.

Dans le prochain chapitre, nous aborderons une étude de cas spécifique sur les services d'Amazon Web Services (AWS). Cette étude de cas nous permettra d'appliquer les concepts et les techniques présentés précédemment à un environnement réel, en mettant en évidence les solutions mises en place par AWS pour gérer efficacement la répartition de charge dans leurs infrastructures.

Chapitre III . Études de cas AWS (Amazon Web Services) :

I. Introduction :

Dans ce chapitre, nous nous concentrerons sur une étude de cas spécifique portant sur les services d'Amazon Web Services (AWS). Nous explorerons comment AWS aborde et résout les défis liés à la répartition de charge dans ses infrastructures cloud. Cette étude de cas nous permettra d'appliquer les connaissances acquises précédemment sur les méthodes d'équilibrage de charge et les algorithmes d'apprentissage automatique à un environnement réel. Nous examinerons les solutions et les stratégies mises en place par AWS pour garantir une distribution efficace de la charge de travail et optimiser les performances des applications hébergées sur leur plateforme cloud.

II. Les solutions et services d'AWS liés à l'équilibrage de charge :

AWS (Amazon Web Services) est un leader mondial dans le domaine du cloud computing et propose une gamme de solutions avancées spécifiquement conçues pour l'équilibrage de charge dans le cloud. Ces solutions offrent aux utilisateurs une manière flexible et évolutive de distribuer la charge sur leurs ressources cloud, assurant ainsi une répartition équilibrée du trafic et une expérience utilisateur optimale. Dans cette section, nous explorerons en détail les solutions proposées par AWS pour l'équilibrage de charge, allant des services de répartition de charge spécialisés aux mécanismes de mise à l'échelle automatique, qui permettent aux utilisateurs de tirer pleinement parti des avantages du cloud en termes de performances et de disponibilité.

Parmi ces solutions et services nous avons :

II.1. Solution d'équilibrage de charge avec Elastic Load Balancer (ELB)

L'Elastic Load Balancer (ELB) d'AWS est un service de répartition de charge entièrement géré qui permet de distribuer le trafic entrant de manière équilibrée sur plusieurs instances.

Il surveille la santé de ses cibles enregistrées et achemine le trafic uniquement vers les personnes en bonne santé Cibles. Elastic Load Balancing met à l'échelle votre équilibreur de charge en fonction de l'évolution de votre trafic entrant au fil du temps. Il peut Évolutivité automatique à la grande majorité des charges de travail.

Il est conçu pour améliorer la disponibilité, la résilience et les performances des applications en dirigeant le trafic vers des ressources cloud appropriées.

ELB offre plusieurs types de Load Balancers :

II.1.a. Classic Load Balancer (CLB)

Classic Load Balancer fournit un équilibrage de charge de base sur plusieurs instances Amazon EC2 et fonctionne à la fois au niveau de la demande et au niveau de la connexion. L'équilibreur de charge classique est destiné aux applications créées au sein du réseau EC2-Classic.

Cet équilibreur de charge sert de point de contact unique pour les clients. Cela augmente la disponibilité de votre application. Vous pouvez ajouter et supprimer des instances de votre équilibreur de charge au fur et à mesure que vos besoins changent, sans perturber le flux global de à votre application. Elastic Load Balancing met à l'échelle votre équilibreur de charge en tant que trafic vers votre L'application change au fil du temps. Elastic Load Balancing peut s'adapter à la grande majorité des charges de travail automatiquement.

II.1.b. Application Load Balancer (ALB)

Application Load Balancer fonctionne au niveau de la demande (couche 7), acheminant le trafic vers les cibles (instances EC2, conteneurs, adresses IP et fonctions Lambda) en fonction du contenu de la demande. Idéal pour l'équilibrage avancé de charge du trafic HTTP et HTTPS, Application Load Balancer fournit un routage avancé des demandes destiné à la fourniture d'architectures d'applications modernes, y compris des micro services

et des applications basées sur des conteneurs. Application Load Balancer simplifie et améliore la sécurité des applications, en garantissant que les derniers chiffrements et protocoles SSL/TLS sont utilisés à tout moment.

Un listener recherche les demandes de connexion des clients à l'aide de l'icône Protocole et port que vous configurez. Les règles que vous définissez pour un écouteur déterminent Comment l'équilibreur de charge achemine les demandes vers ses cibles enregistrées. Chaque règle consiste en une priorité, une ou plusieurs actions et une ou plusieurs conditions. Lorsque les conditions d'un sont respectées, puis ses actions sont effectuées. Vous devez définir une règle par défaut pour chaque, et vous pouvez éventuellement définir des règles supplémentaires.

Chaque groupe achemine les demandes vers un ou plusieurs cibles à l'aide de protocole et du numéro de port que vous spécifiez. Des contrôles de santé sont effectués sur toutes les cibles enregistrées dans un Groupe cible spécifié dans une règle d'écoute pour votre équilibreur de charge.

II.1.c. Network Load Balancer (NLB)

Un Network Load Balancer fonctionne à la quatrième couche du modèle OSI (Open Systems Interconnection). Il peut gérer des millions de requêtes par seconde. Une fois que l'équilibreur de charge a reçu une demande de connexion, il sélectionne une cible dans le groupe cible pour la règle par défaut. Il tente d'ouvrir une connexion TCP à la cible sélectionnée sur le port spécifié dans le Configuration de l'écouteur.

Lorsque vous activez une zone de disponibilité pour l'équilibreur de charge, Elastic Load Balancing crée une charge nœud d'équilibrage dans la zone de disponibilité. Par défaut, chaque nœud d'équilibrage de charge distribue trafic sur les cibles enregistrées dans sa zone de disponibilité uniquement. Si vous activez équilibrage de charge interzone, chaque nœud d'équilibrage de charge répartit le trafic sur les cibles enregistrées dans toutes les zones de disponibilité activées. Pour augmenter la tolérance aux pannes de vos applications, vous pouvez activer plusieurs Zones de disponibilité pour votre équilibreur de charge et assurez-vous que chaque groupe cible dispose d'au moins une cible dans chaque zone de disponibilité activée.

II.2. Solution d'équilibrage de charge avec Auto Scaling

Amazon EC2 Auto Scaling vous aide à vous assurer que vous disposez du nombre correct d'instances Amazon EC2 disponibles pour Gérer la charge de votre application. Vous créez des collections d'instances EC2, appelées groupes Auto Scaling. Vous pouvez spécifier le nombre minimal d'instances dans chaque groupe Auto Scaling et Amazon EC2 Auto Scaling garantissent que votre groupe ne descend jamais en dessous de cette taille. Vous pouvez spécifiez le nombre maximal d'instances dans chaque groupe Auto Scaling, et Amazon EC2 Auto Scaling garantit que votre ne dépasse jamais cette taille. Si vous spécifiez la capacité souhaitée, soit lorsque vous créer le groupe ou à tout moment par la suite, Amazon EC2 Auto Scaling garantit que votre groupe en dispose d'autant Cas. Si vous spécifiez des stratégies de mise à l'échelle, Amazon EC2 Auto Scaling peut lancer ou résilier des instances. À mesure que la demande sur votre application augmente ou diminue.

III. Conclusion :

En conclusion de cette étude de cas portant sur les services d'Amazon Web Services (AWS), nous avons examiné de près les solutions proposées par AWS en matière d'équilibrage de charge. À travers une analyse approfondie des fonctionnalités et des stratégies mises en place, nous avons pu constater l'expertise d'AWS dans la gestion efficace de la répartition de charge dans leurs infrastructures cloud.

Conclusion générale

D'après la recherche et le travail qu'on a mené à propos du cloud computing on peut sans hésitation conclure que ce dernier est une révolution dans la manière d'organiser, de gérer et de distribuer des ressources informatiques. Sa définition opérationnelle annonce un modèle informatique qui permet un accès facile et à la demande, par le réseau, à un ensemble partagé de ressources informatiques configurables et depuis n'importe quel appareil disposant d'un navigateur et d'une connexion internet. Il peut s'agir de serveurs, de stockage, d'applications ou de services, rapidement provisionnés et libérés par un minimum d'efforts de gestion.

Bibliographie

- A Review on Different Load Balancing Techniques in Cloud Computing Environments. (2015). *International Journal for Scientific Research & Development*.
- Amrutanshu Panigrahi, B. S. (s.d.). M-Throttled: Dynamic Load Balancing Algorithm for Cloud Computing. *M-Throttled: Dynamic Load Balancing Algorithm for Cloud Computing*.
- cloudflare. (s.d.). *Type d'équilibrage de charge*. Récupéré sur cloudflare:
<https://www.cloudflare.com/fr-fr/learning/performance/types-of-load-balancing-algorithms/>
- Dr Syeda Gauhar Fatima, S. K. (2019). CLOUD COMPUTING AND LOAD BALANCING. *International Journal of Advanced Research in Engineering and Technology (IJARET)*.
- Haryani, N. (2014). Dynamic Method for Load Balancing in Cloud Computing. *IOSR Journal of Computer Engineering (IOSR-JCE)*.
- Hicham, R. (17-Dec-2020). *Cloud Computing : services informatiques dynamiques basés sur le Web - Concepts et notions de base -*.
- IBM. (s.d.). IBM. Récupéré sur What is load balancing: <https://www.ibm.com/topics/load-balancing>
- Juliet Gathoni Muchori, P. M. (2022). Machine Learning Load Balancing Techniques in. *International Journal of Computer Applications Technology and Research*.
- Misra, D. S. (2022). LOAD BALANCING IN CLOUD COMPUTING. *International Research Journal of Engineering and Technology*.
- Najwa, S. L. (Avril 2022). *L'émergence du Cloud Computing au service de la transparence des collaborations inter-organisationnelles et de la confiance numérique: revue de littérature*.
- Roopha, K. e. (December 24, 2020). Infrastructural Constraints of Cloud Computing. *International Journal of Management, Technology And Engineering*.
- Semwal, P. (2021). A Review of Load Balancing Algorithms in Cloud Computing. *International Journal of Creative Research Thoughts*.
- Slimane, y. (2019). *Equilibrage de charge dans les environnement cloud computing*.
- Yeluri, G. (2019). *Load balancing in cloud computing*.
- Yeluri, G. (2019). *Load Balancing in cloud computing*.