Gérez le passage à l'échelle de votre site

Dans ce cours, vous avez appris à mettre en place de nombreux services et je suis sûr qu'ils seront bientôt très populaires auprès de vos utilisateurs. Attention toutefois à ne pas être victime de votre succès et à anticiper la croissance de votre infrastructure. Dans ce chapitre, je vais vous parler de la problématique de la croissance et des différents moyens d'y faire face.

Quand on parle de croissance, il faut déjà savoir de quoi on parle.

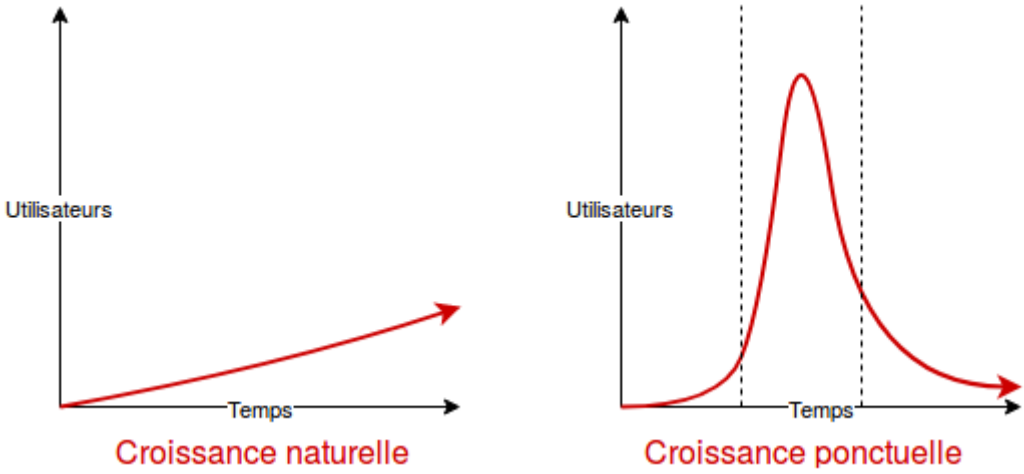
Comprenez ce qu'est la croissance

Quand on parle de croissance d'une infrastructure ou d'un site, on parle de l'augmentation de certains **indicateurs clés**. Des indicateurs classiques peuvent être :

- le nombre de visiteurs
- le nombre de requêtes
- le nombre de pages vues
- la bande passante consommée
- l'espace de stockage nécessaire
- la puissance de calcul nécessaire
- etc.

Cette liste pourrait encore être complétée par beaucoup d'autres indicateurs. Évidemment, certains indicateurs sont liés les uns aux autres. Par exemple, plus de visiteurs impliquera généralement plus de... tout le reste 

Le contexte dans lequel cette croissance arrive est aussi important. Un service dont les utilisateurs sont satisfaits a tendance à se développer et à voir son audience augmenter petit à petit de manière régulière. On est alors dans une forme de **croissance naturelle**. Par contre, quand vous faites brusquement face à des pics très importants de l'activité, on parle de **croissance ponctuelle**. Cette croissance ponctuelle peut être **prévue** : par exemple deux fois par an, les soldes provoquent un afflux de clients dans les boutiques. Elle peut aussi être **imprévue**, par exemple si un média important communique sur un service.



Croissance naturelle contre croissance ponctuelle

> Il y a quelques années, dans les milieux informatiques, on parlait de “Slashdot Effect” pour parler de certains effets d’une croissance ponctuelle imprévue. En effet, **Slashdot est un site d’informations spécialisé dans les nouvelles technologies**. Régulièrement, quand un nouvel article faisait mention d’un site web, le site en question se retrouvait tellement submergé de visiteurs qu’il restait indisponible pendant plusieurs jours.

La croissance, c’est comme la célébrité, tout le monde en rêve mais tout le monde ne sait pas bien la gérer. Depuis les débuts de l’informatique, en fonction des moyens techniques disponibles, les stratégies pour faire face à la croissance ont largement évolué.

De la croissance verticale à la croissance horizontale

Dans les débuts de l’informatique, les services étaient centralisés sur de gros ordinateurs très puissants (pour l’époque) : les mainframes. À l’époque, le seul moyen de faire face à la croissance d’un service était d’augmenter la puissance du serveur en rajoutant de la mémoire, du CPU, des disques, etc. On parle alors de **croissance verticale** pour évoquer l’augmentation des capacités physiques d’une architecture existante.

Ce modèle de croissance a un avantage principal et beaucoup d’inconvénients.

L’avantage principal :

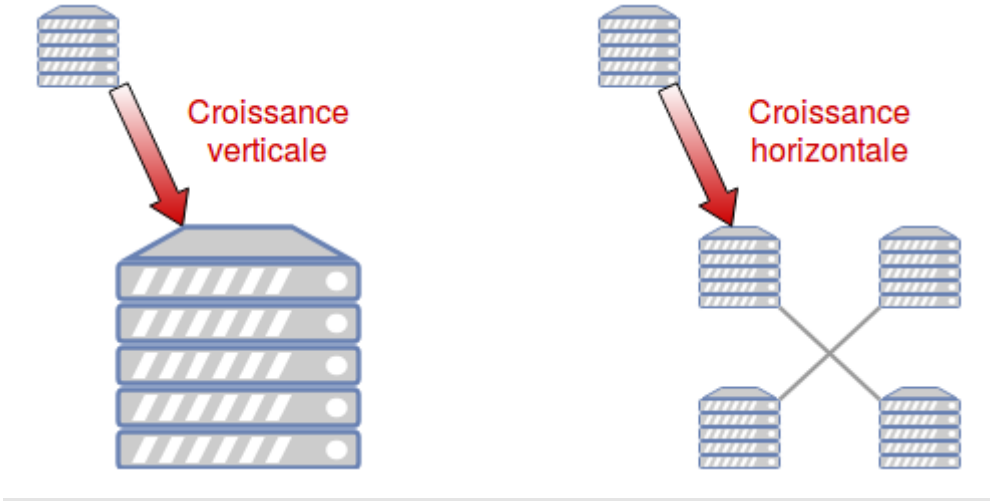
- un modèle centralisé est plus simple à concevoir. Pour le faire grossir de manière verticale, vous ne changez rien à votre architecture. Vous avez juste à payer les nouveaux composants et vous êtes reparti pour un tour.

Inconvénients :

- ajouter des composants nécessite souvent d’interrompre le service
- il faut anticiper un minimum la croissance pour prévoir dès le départ du matériel évolutif
- ça peut coûter cher

- surtout, ce modèle a des limites : on ne peut pas rajouter indéfiniment de la mémoire et des CPU sur une machine unique.

Avec le développement du modèle client-serveur, les ordinateurs ont commencé à travailler de plus en plus en réseau. Il a alors été possible d'envisager de faire croître une architecture, non pas en faisant grossir chaque machine, mais en rajoutant des machines. On parle alors de **croissance horizontale** pour faire référence à la croissance d'un service par l'augmentation du nombre de machines.



Croissance verticale contre croissance horizontale

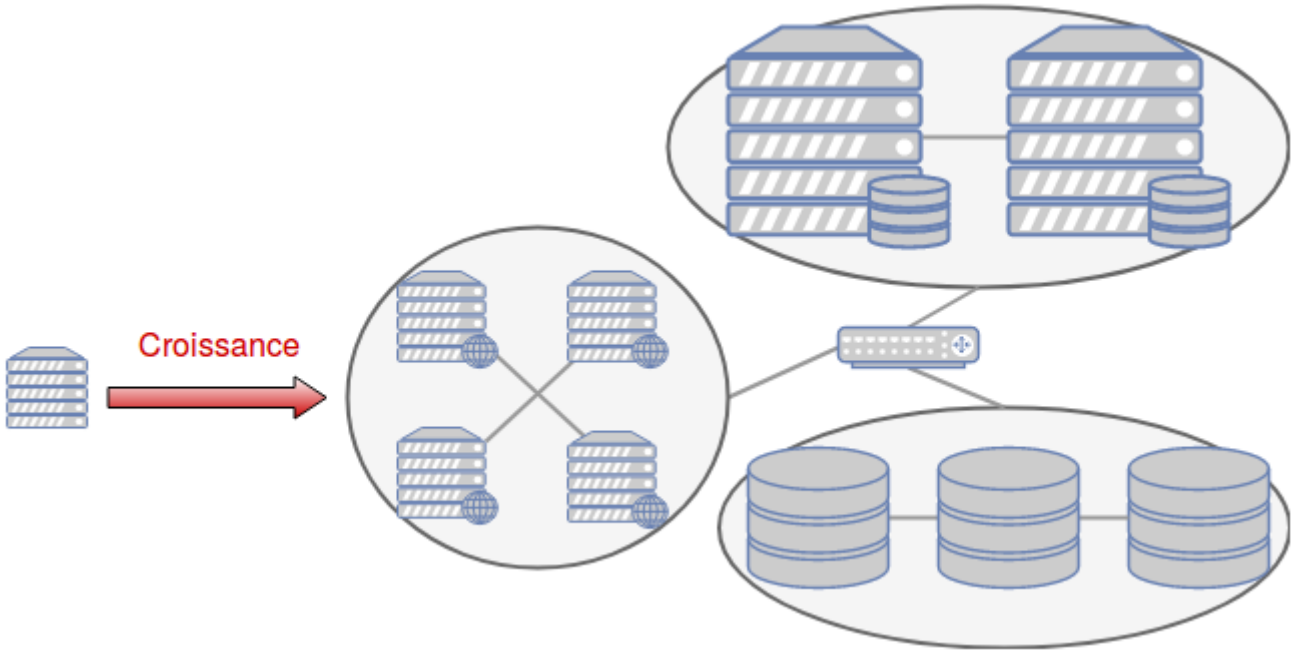
Ce modèle de croissance présente beaucoup d'avantages :

- avec la standardisation du matériel, on s'est aperçu qu'il était souvent bien moins cher d'avoir un grand nombre de machines à bas coût plutôt qu'une seule machine très puissante
- l'augmentation du nombre de machines peut souvent se faire sans interruption de service. Globalement, la répartition d'un service sur plusieurs machines augmente largement la disponibilité.
- la croissance horizontale est potentiellement illimitée (ou en tout cas les limites sont bien plus hautes).

Dans ces architectures distribuées, on peut adopter plusieurs stratégies :

- répartir les différents composants logiques (serveur web, base de données, etc.) sur des machines différentes
- augmenter le nombre de machines pour chacun des composants
- augmenter la puissance de chacune des machines individuellement (croissance verticale)

Généralement, la stratégie globale de croissance est un savant mélange de tout ça.



Stratégie de croissance mixant plusieurs statégies

Malgré la puissance théorique quasiment illimitée du modèle distribué et de la puissance horizontale, ce modèle présente des limites pratiques. Ce modèle est bien adapté à la croissance naturelle mais son manque de souplesse rend encore difficile la gestion des pics ponctuels d'activité.

C'est la problématique à laquelle était confronté le site de vente en ligne Amazon il y a quelques années. Amazon réalisait alors plus de la moitié de son chiffre d'affaire annuel sur le seul mois de décembre. Pour faire face à cet énorme pic d'activité, il a investi dans une architecture informatique gigantesque. Ainsi, il a effectivement pu faire face à la charge du mois de décembre mais son infrastructure était clairement sur-dimensionnée pour les 11 mois de l'année restants. Il a alors eu l'idée géniale de louer ses ressources informatiques non-utilisée et a créé un nouveau modèle : **le Cloud**.

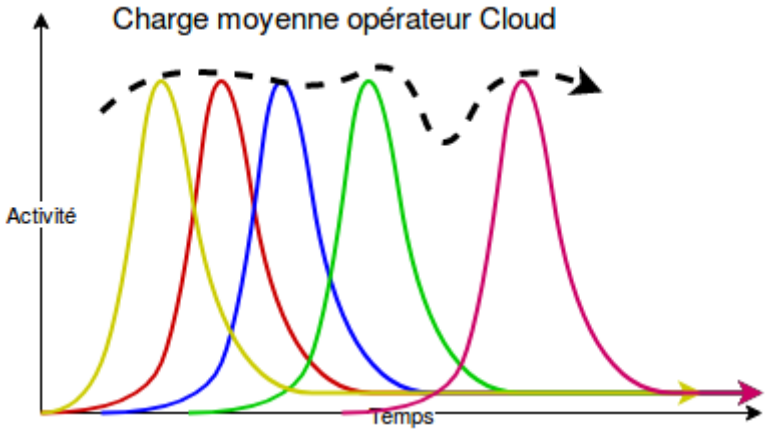
L'apport du Cloud à la gestion de la croissance

Le problème du modèle distribué en auto-hébergement (**On Premise** disent les anglophones), c'est que pour pouvoir faire face à un pic d'activité 1000 fois plus important que votre charge quotidienne, vous devez investir dans une architecture 1000 fois trop grande pour vous qui sera inutilisée le reste de l'année.

> *i* Pour parler de la capacité d'une infrastructure à grossir vite ET à réduire vite, on parle d'**élasticité**. C'est un des avantages offerts par le Cloud.

L'idée du Cloud, c'est que tout le monde ne connaît pas ses pics d'activité en même temps : les sites de commerce plutôt en fin d'année, les cabinets comptables plutôt au printemps, etc. S'il arrive à mutualiser les ressources de suffisamment de monde, un opérateur Cloud peut arriver à

une charge globale beaucoup plus constante tout en permettant à chacun d'avoir des pics ponctuels importants.



Un opérateur Cloud joue sur le volume et la diversité de ses clients pour arriver à une charge plus constante

Cette mutualisation des ressources doit aussi beaucoup aux possibilités des systèmes modernes de virtualisation qui permettent de réallouer des ressources inutilisées là on en a besoin.

Le modèle du Cloud évite des investissements coûteux à ses utilisateurs car ils ne paient que pour les ressources réellement consommées. Il fournit le matériel et le réseau comme des commodités et permet d'automatiser complètement la création ou la destruction d'une plate-forme.

Dans un service de Cloud, vous paierez le même prix pour utiliser 1 serveur pendant 1000 heures ou 1000 serveurs pendant 1 heure. Cerise sur le gâteau, une fois votre service configuré, vous pourrez créer et utiliser vos 1000 serveurs en quelques minutes.

> Aujourd'hui tous les grands studios d'animation utilisent des plate-formes de Cloud pour réaliser les calculs de leurs films. L'un d'eux avait déclaré aux médias qu'ils pourraient attendre quelques dizaines d'années pour faire les calculs sur un petit nombre de machines mais qu'ils préféraient louer un grand nombre de machine pour faire le travail en quelques heures car ils n'avaient pas la patience d'attendre

Malgré les possibilités alléchantes du Cloud, il n'est pas toujours si facile de concevoir une architecture scalable et élastique. Pour pouvoir profiter de tous les bénéfices de ce modèle, il faut parfois adapter les outils utilisés et faire appel à des technologies spécifiques. Pour en savoir plus sur le cloud et ses outils, je vous recommande de suivre le cours "Découvrez le cloud avec Amazon Web Services".

En résumé

- on parle de croissance pour faire référence à l'augmentation d'indicateurs clés tel que le nombre de visiteurs

- la croissance peut être naturelle ou ponctuelle
- pour faire face à la croissance, vous pouvez faire grossir votre infrastructure de manière verticale ou de manière horizontale
- la croissance horizontale est moins limitée que la croissance verticale
- le cloud apporte l'élasticité qui manquait aux architectures distribuées On Premise

Le professeur

Etienne Lavanant

Étienne Lavanant, Ingénieur Systèmes diplômé de Télécom Sud-Paris, travaille en tant que Freelance sur Paris