De l'archivage du cyber espace-temps

Objectif :

Concevoir un robot capable de parcourir une partition du cyber espace-temps et d'en conserver une succession d'images.

I. Théorie

II. Structure applicative

III. Performances

IV. Conclusion

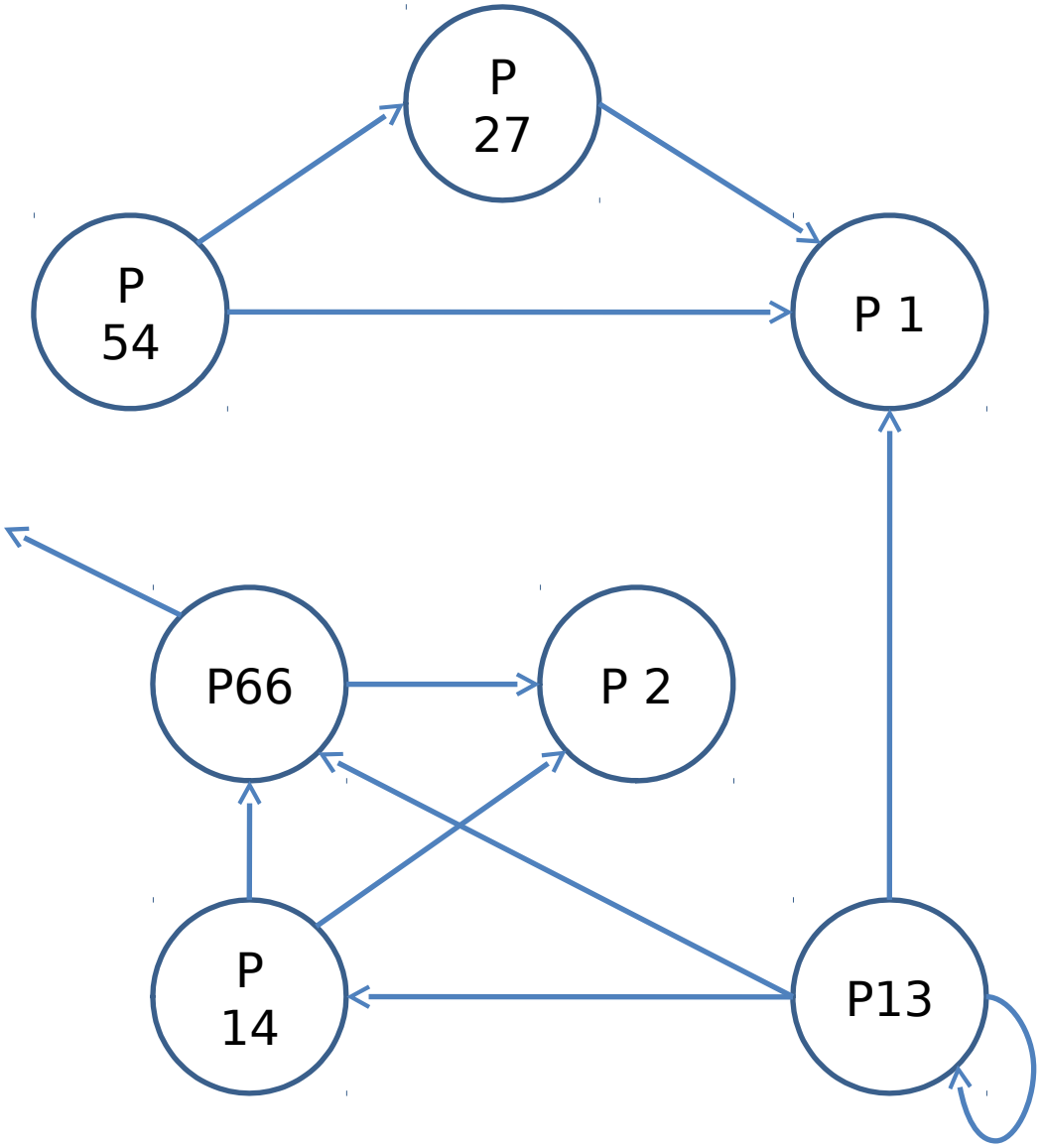
Internet

Internet : 1960
Web : 1992, Tim Berners-Lee

Aujourd'hui :

- 10^9 sites
- 4 % du PIB français
- 20 % de la croissance mondiale

Structure du Web



Graphe à 7 nœuds

HTML

<balise> : ouverture
</balise> : fermeture

```
<html>
    <head>
        <title> Une page html </title>
    </head>
    <body>
        <p>
            Un paragraphe ……
            ……..
            <a href ''https://site1.fr''>
Un lien </a>
            Fin du paragraphe
        </p>
    </body>
</html>
```

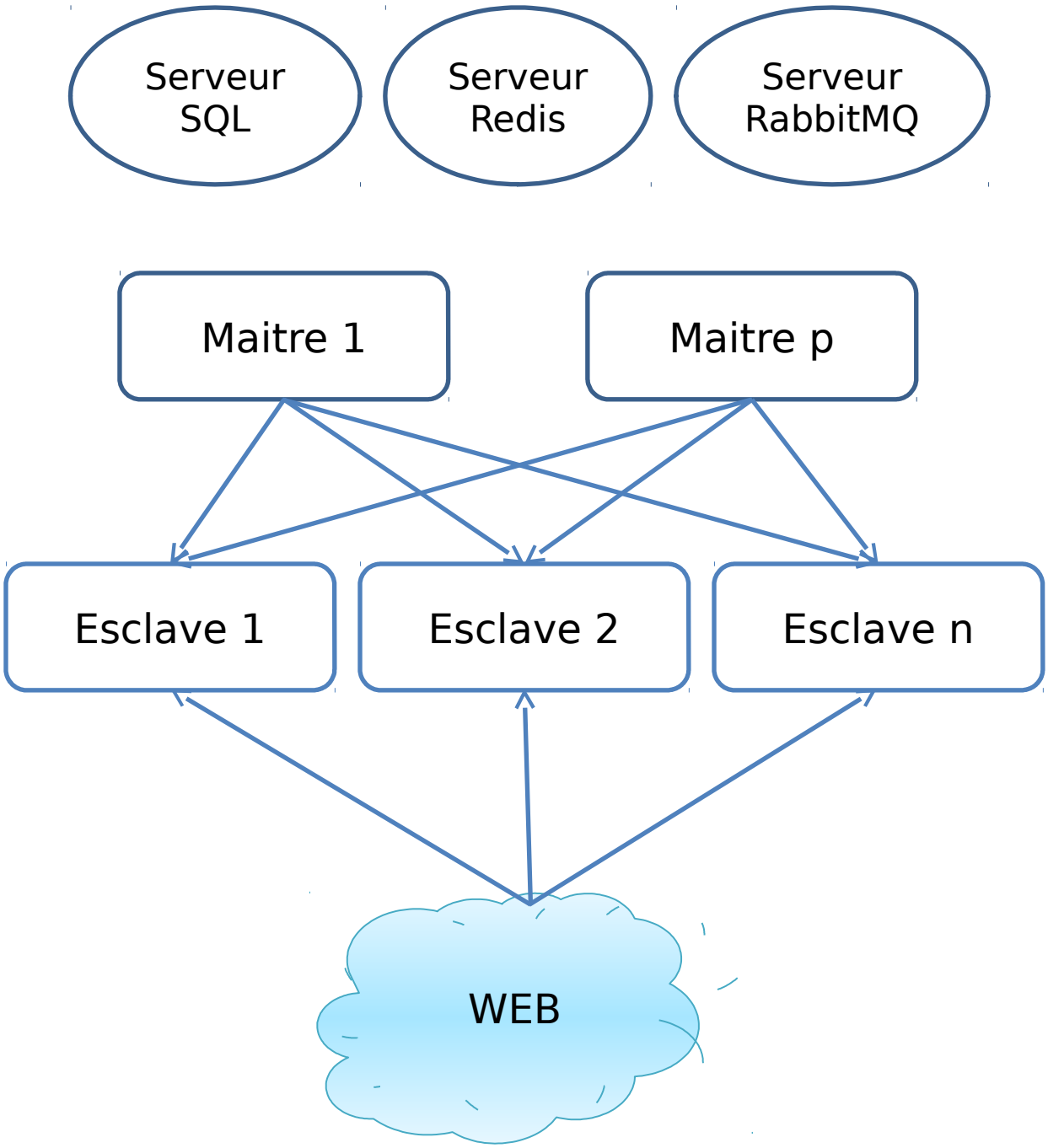
URL

http: //www.google.fr /query? q=1

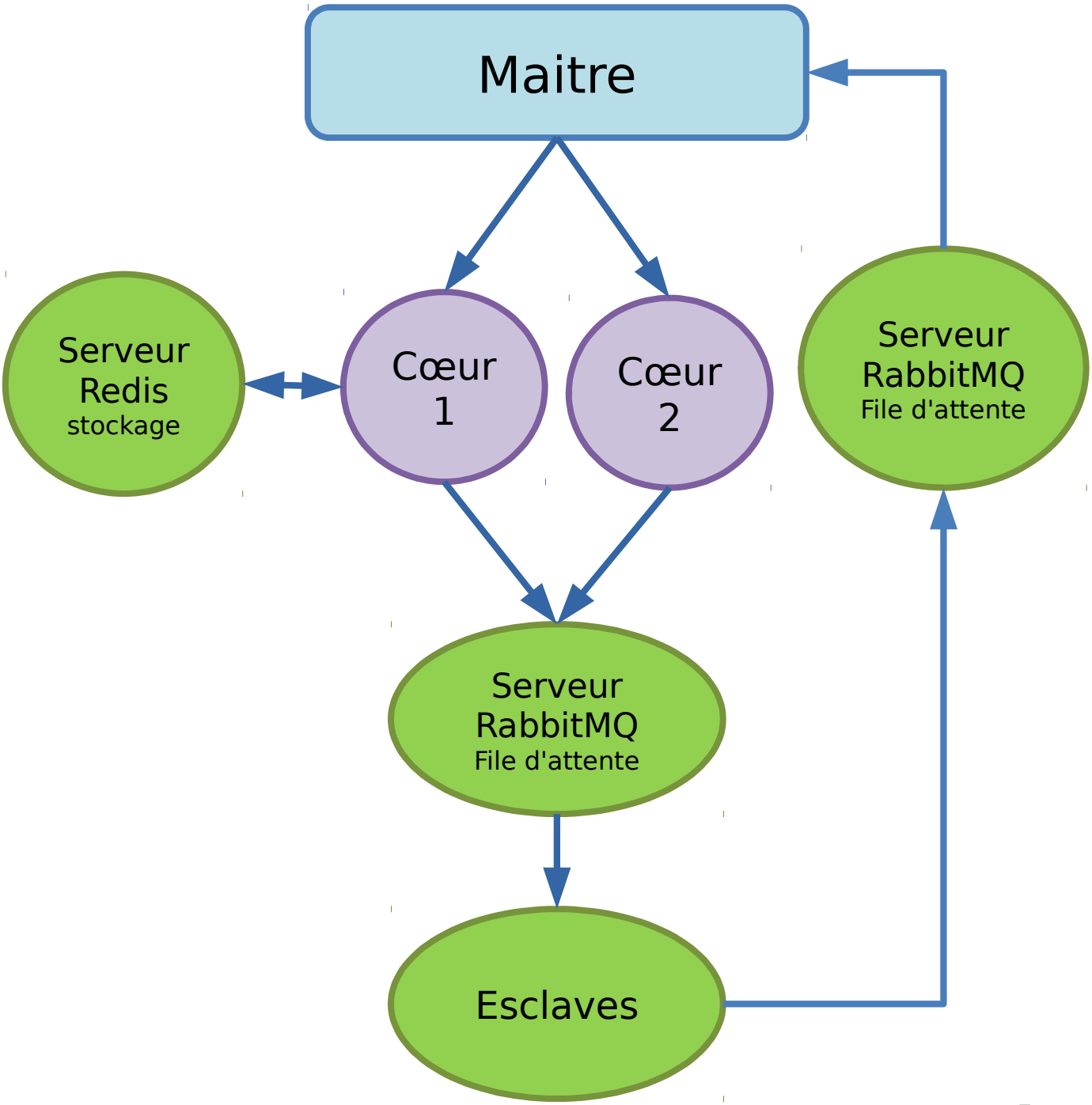
Metadonnées utilisées

Taille d’une page
Type
Hash

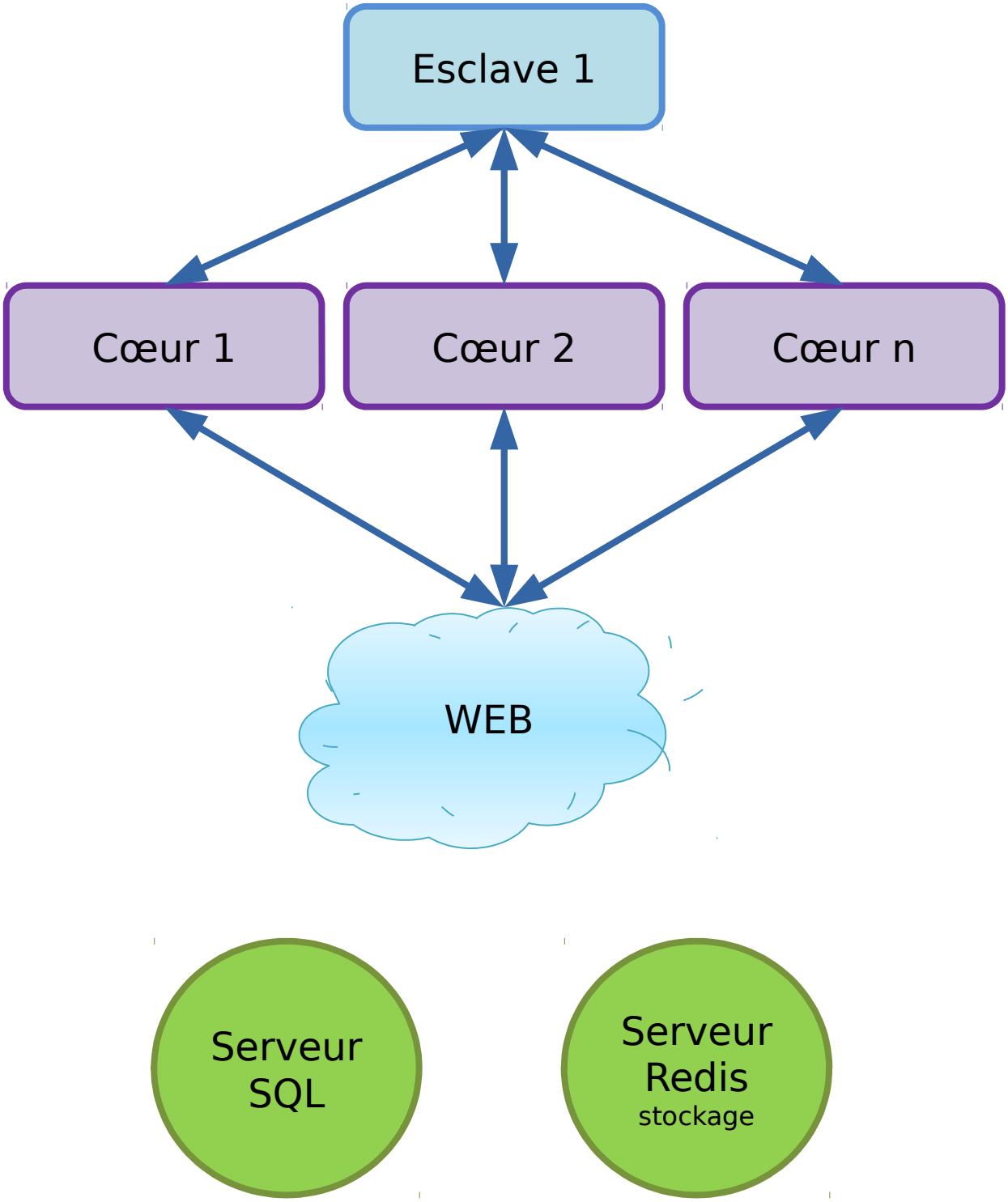
Structure applicative



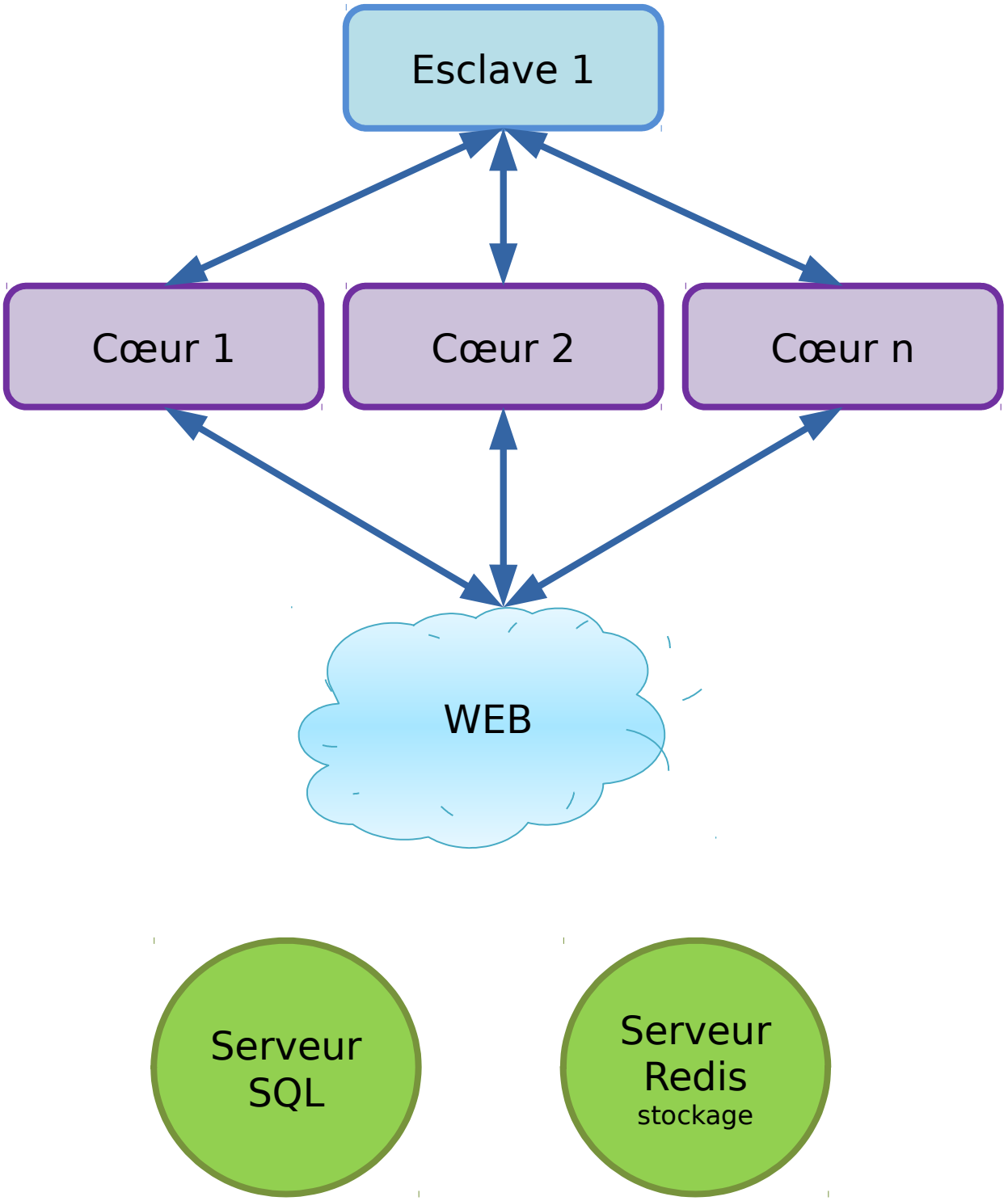
Structure d’un serveur maitre



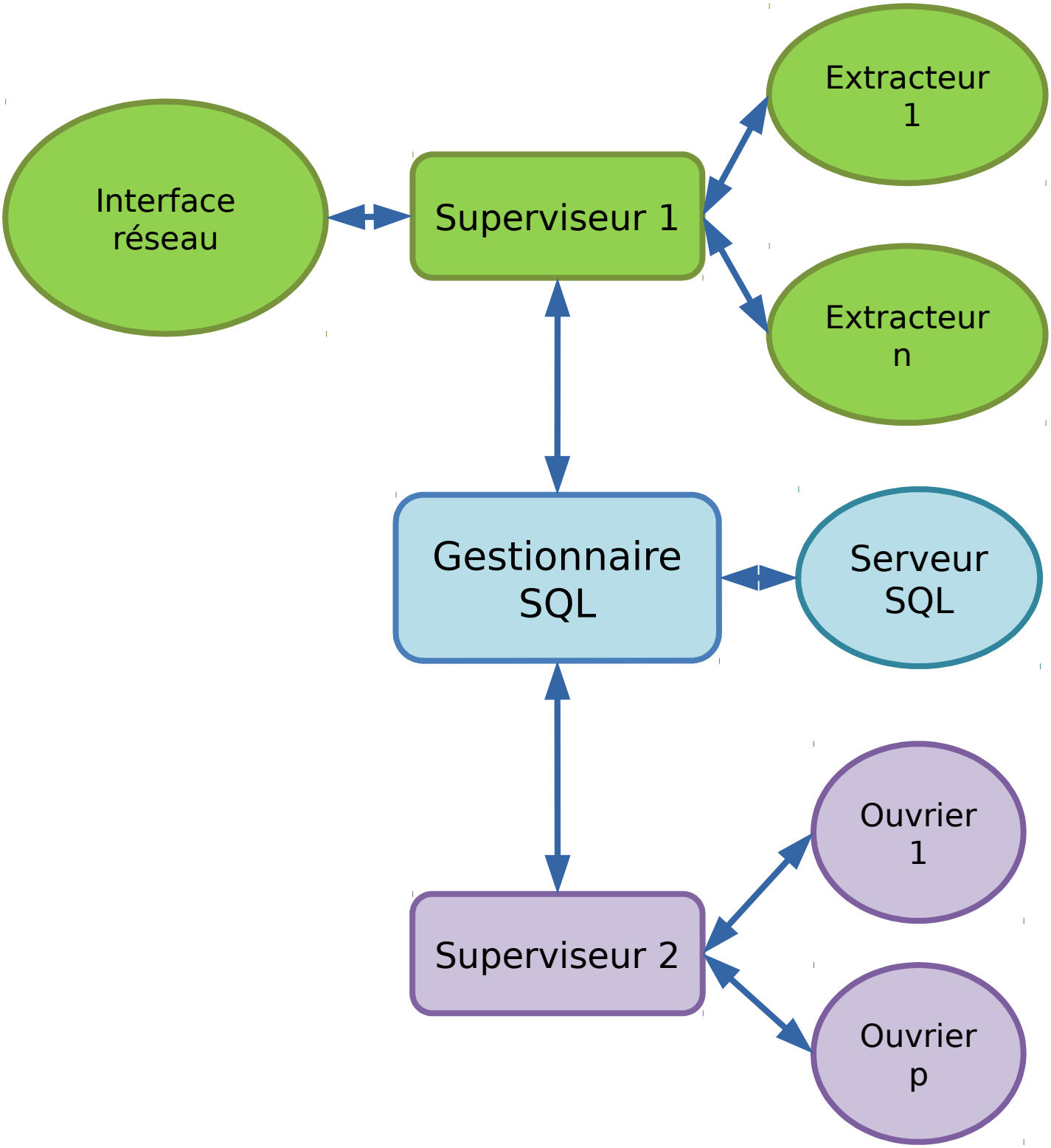
Structure d’un serveur esclave



Structure d’un serveur esclave



Structure d’un cœur esclave



Application

~ 6000 lignes de codes

Langages :

- Python
- C++

API et langages associés :

- SQL
- Redis
- amqp
- Bash

Protocole de test

Objectifs :

Étude de la performance d’un esclave isolé

Étude de l’apport de la parallélisation pour un esclave isolé

Étude de la scalabilité réseau

Méthodologie :

Site d'étude : fr-wikipedia.org
Fichiers : html uniquement
Statistiques : métadonnées
Vitesse : pages/seconde

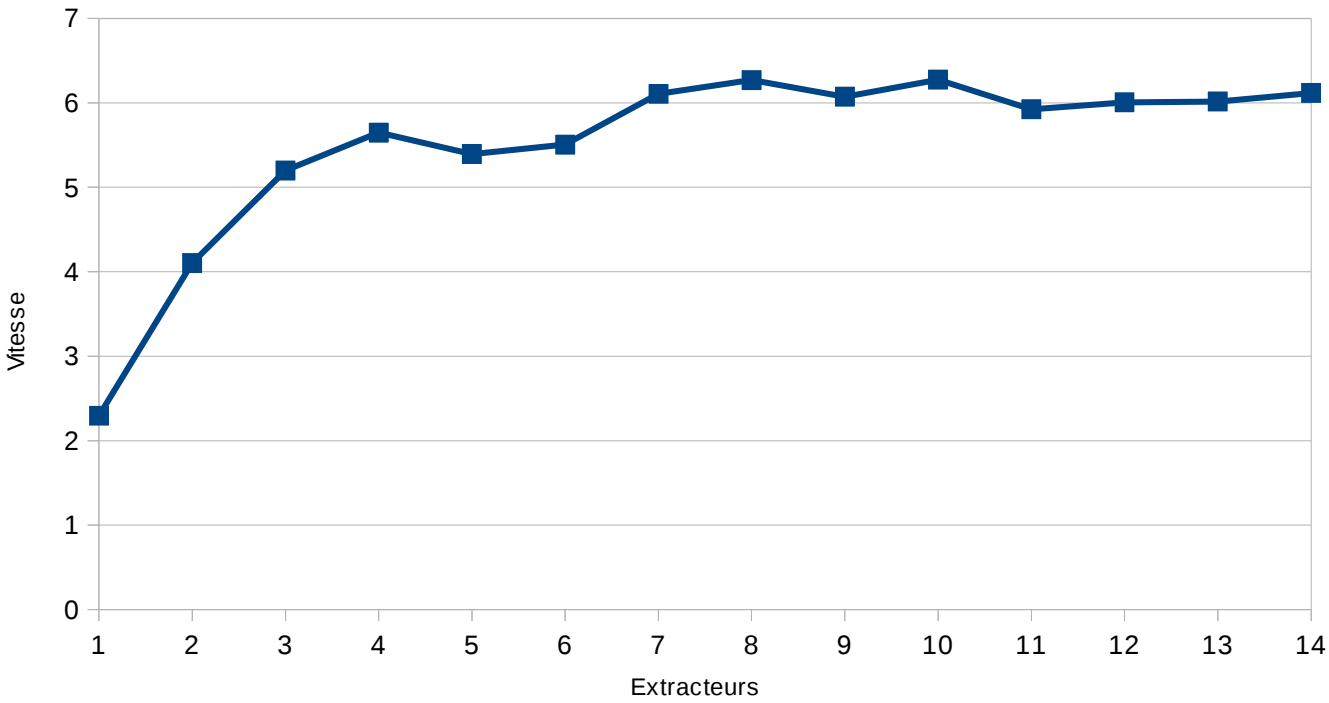
Machines utilisées :

Instances Runabove (Cloud public d'OVH)

Étude de la performance d’un esclave isolé

Machines utilisées :

Sandbox M(1VCore)

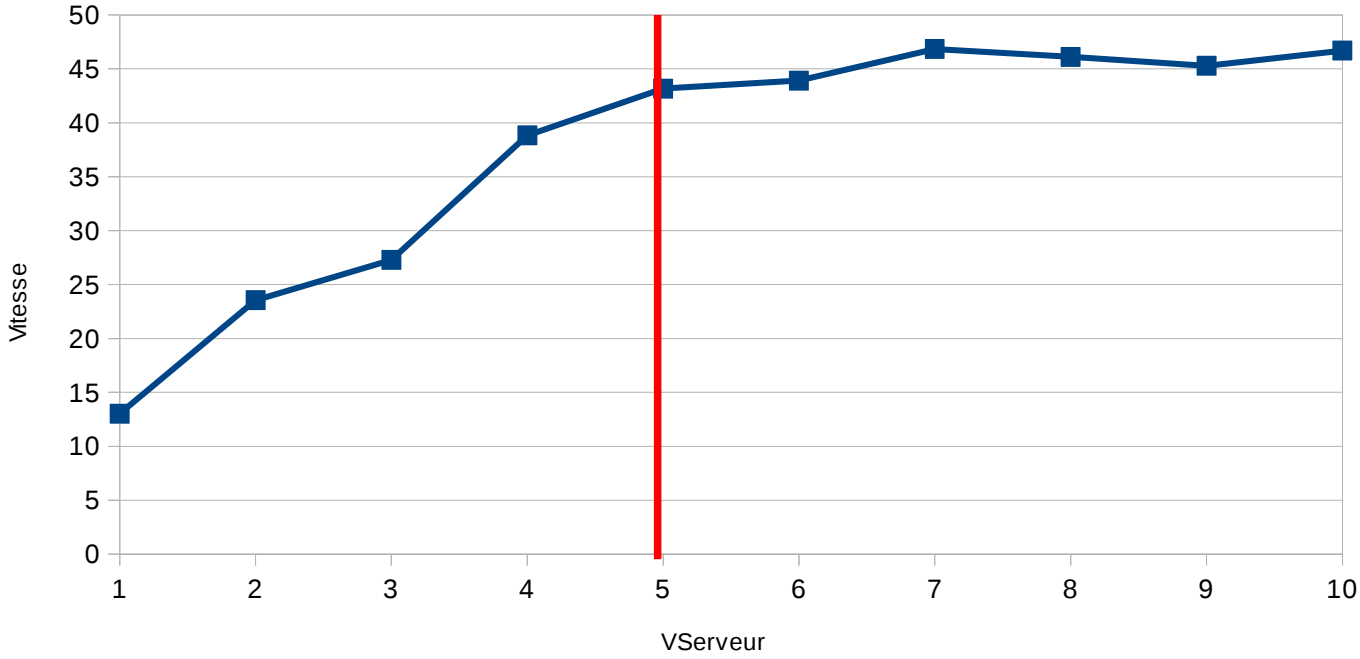


Étude de l’apport de la parallélisation pour un esclave isolé

Machines utilisées :

Slave : XL4(5Vcores)

Autre: Sandbox M(1Vcore)

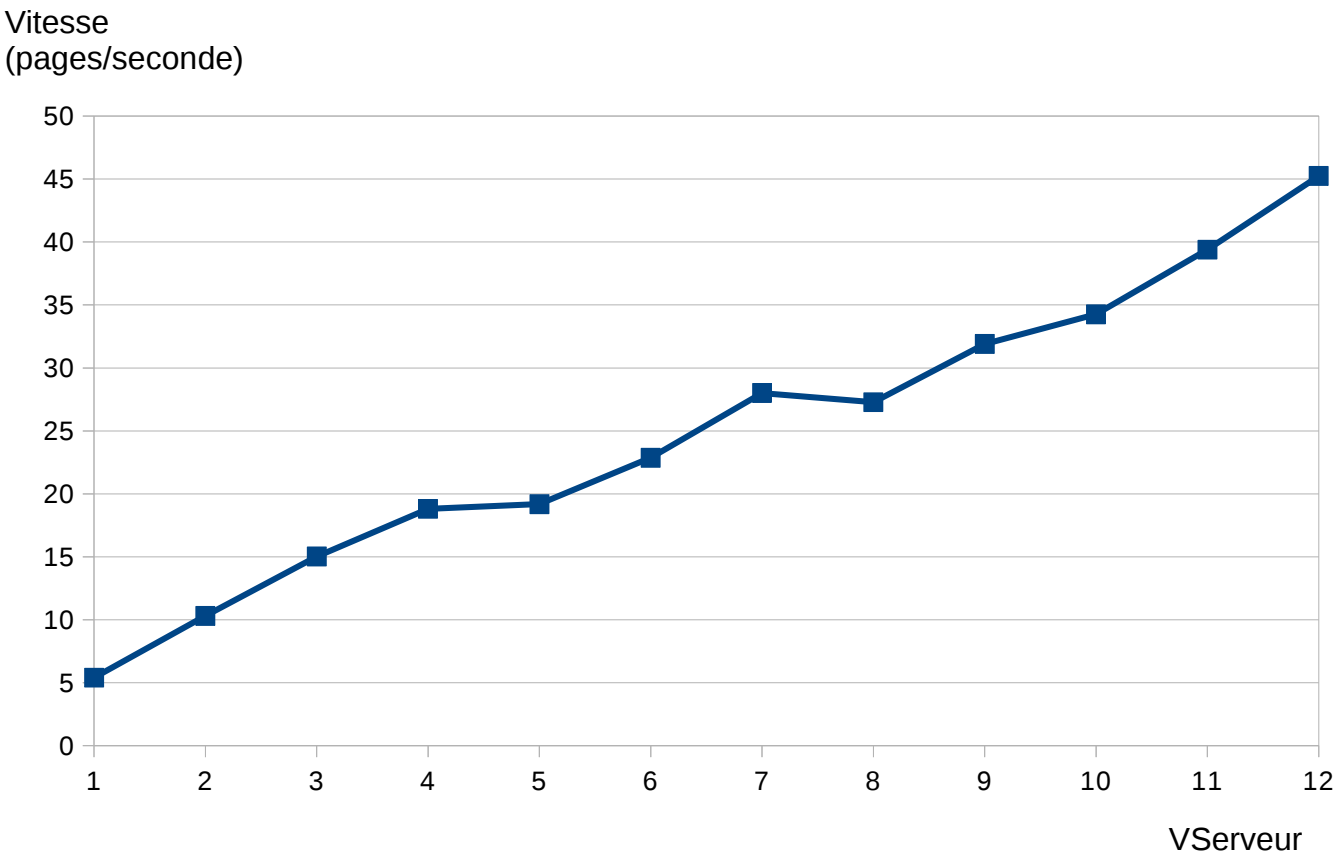


Etude de l’apport de la parallélisation pour un esclave isolé

Machines utilisées :

Slave : Power8 XL2 (24Vcores)

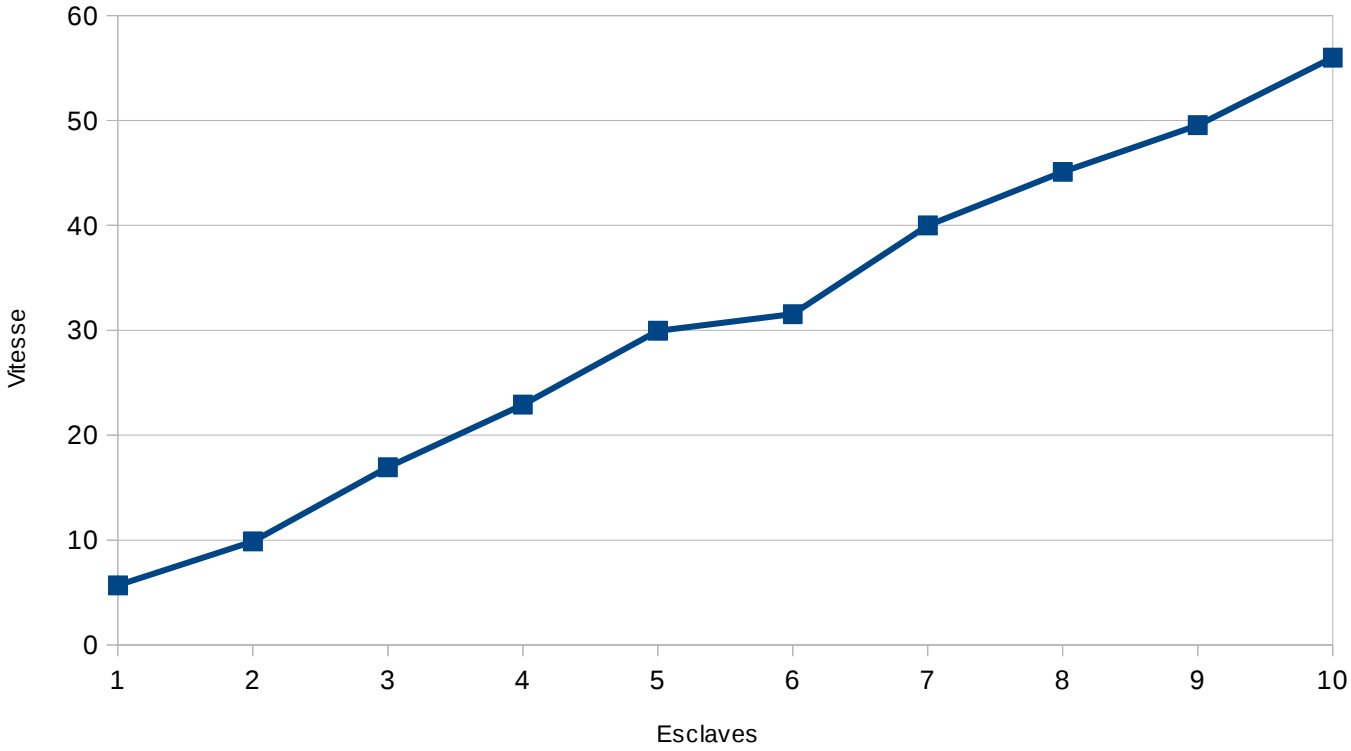
Autre: L(2Vcores)



Etude de la scalabilité réseau

Machines utilisées :

Sandbox M (1Vcore)



Limites

Tolérance aux pannes :

Stockage local

Sécurité :

Pas d'analyse avant archivage

Performances :

Structure pseudo-centralisée
Consommation RAM

Améliorations

Système de fichiers distribué
Supprimer les réplicats
Crawler : hash-base partitioning

Conclusion

Capacités maximales atteintes :

4 Esclaves (XL4)
1 Broker RabbitMQ (S)
1 Serveur MySQL (S)
1 Serveur Redis (S)
1 Maitres (XL4)

Vitesse : 504 000 pages/heure
140 pages/seconde
Volumétrie : 50 Go/heure

Estimation Wikipedia français :

Pages : ~$6*10^6$
Temps : ~12H
Taille : ~600Go

Conclusion

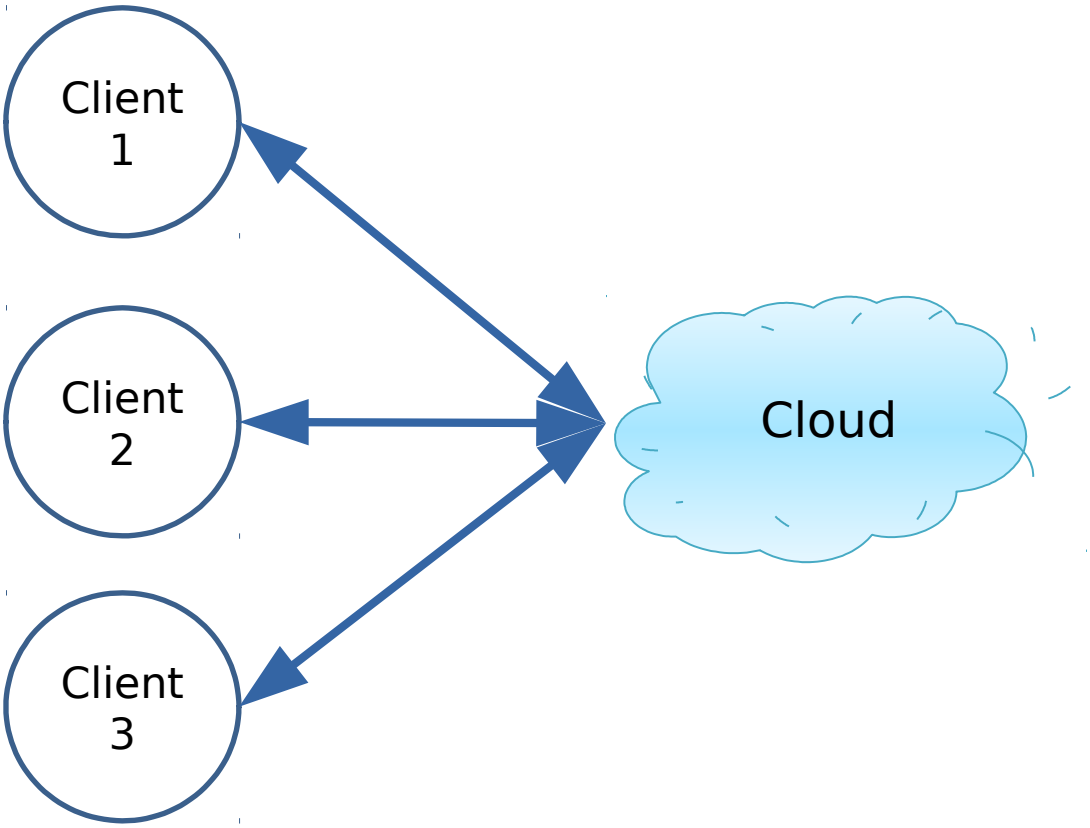
Usages :

- Moteurs de recherches
- Bibliothèque national de France
- Archive.org

Extensions :

- Github : https://github.com/athena-project
- Developements en cours :
  - Hash-base partitionning
  - Repartition de charge (serveur maître)

Le Cloud Public



Sandbox M : 1Vcore, 2GB Ram, 1Gbs
S : 1Vcore, 2GB Ram, 1Gbs, SLA
XL4 : 5Vcores, 24GB Ram, 10Gbs, SLA
Power8 XL2 : 24Vcores, 48GB Ram, 10Gbs, SLA

Cyber espace-temps

https://fr.wikipedia.org/wiki/France