

Cassandra

Haute disponibilité et élasticité avec Cassandra

17/06/2011

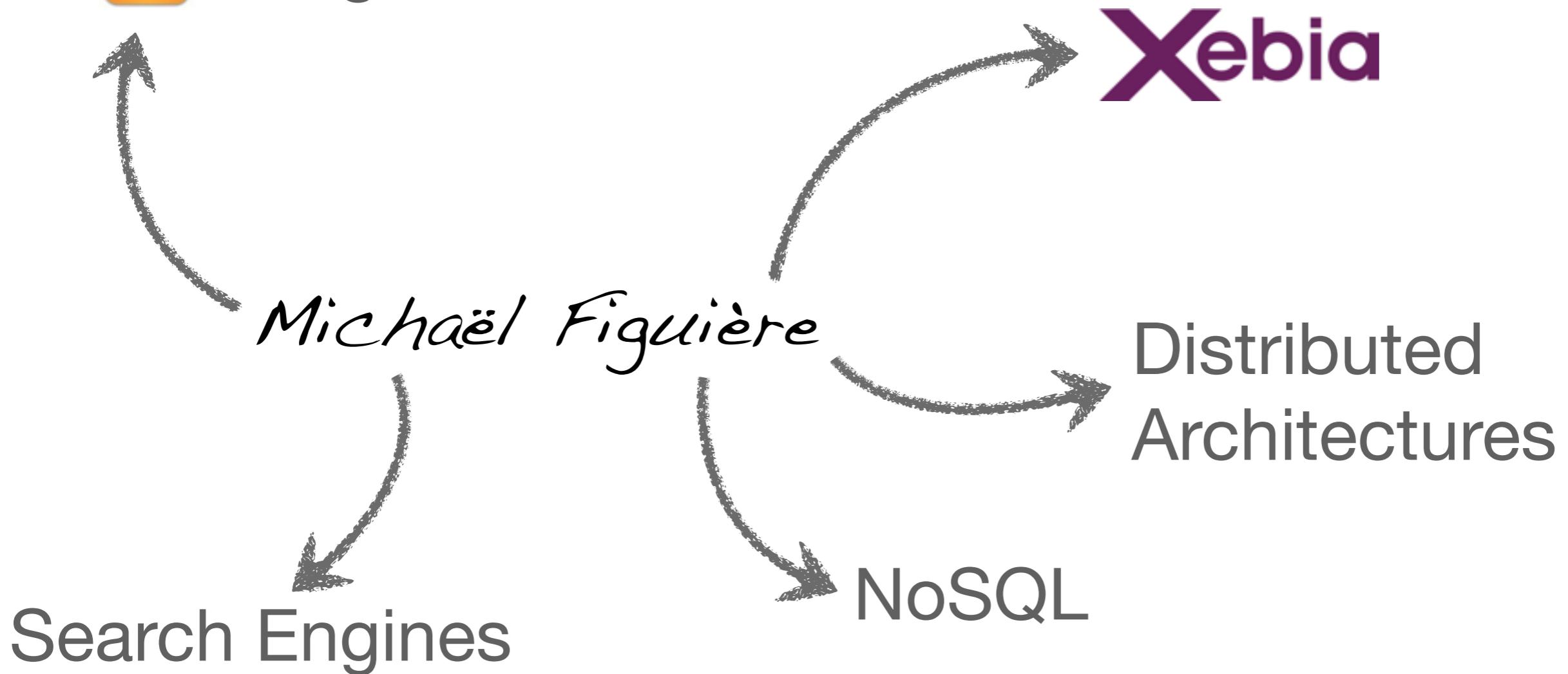
Speaker



@mfiguiere



blog.xebia.fr



Cassandra en quelques mots

- Projet Apache, base de données NoSQL, peer to peer, hautement disponible
↳ Possibilité de stocker plusieurs To
- Encore en version 0.8
↳ Mais déjà déployée en production !
- Approche très différente des bases de données relationnelles
↳ Nécessite un peu d'apprentissage...

Dynamo et Cassandra

Objectifs similaires :

- Faible latence
- Très haute disponibilité
- Scalabilité massive



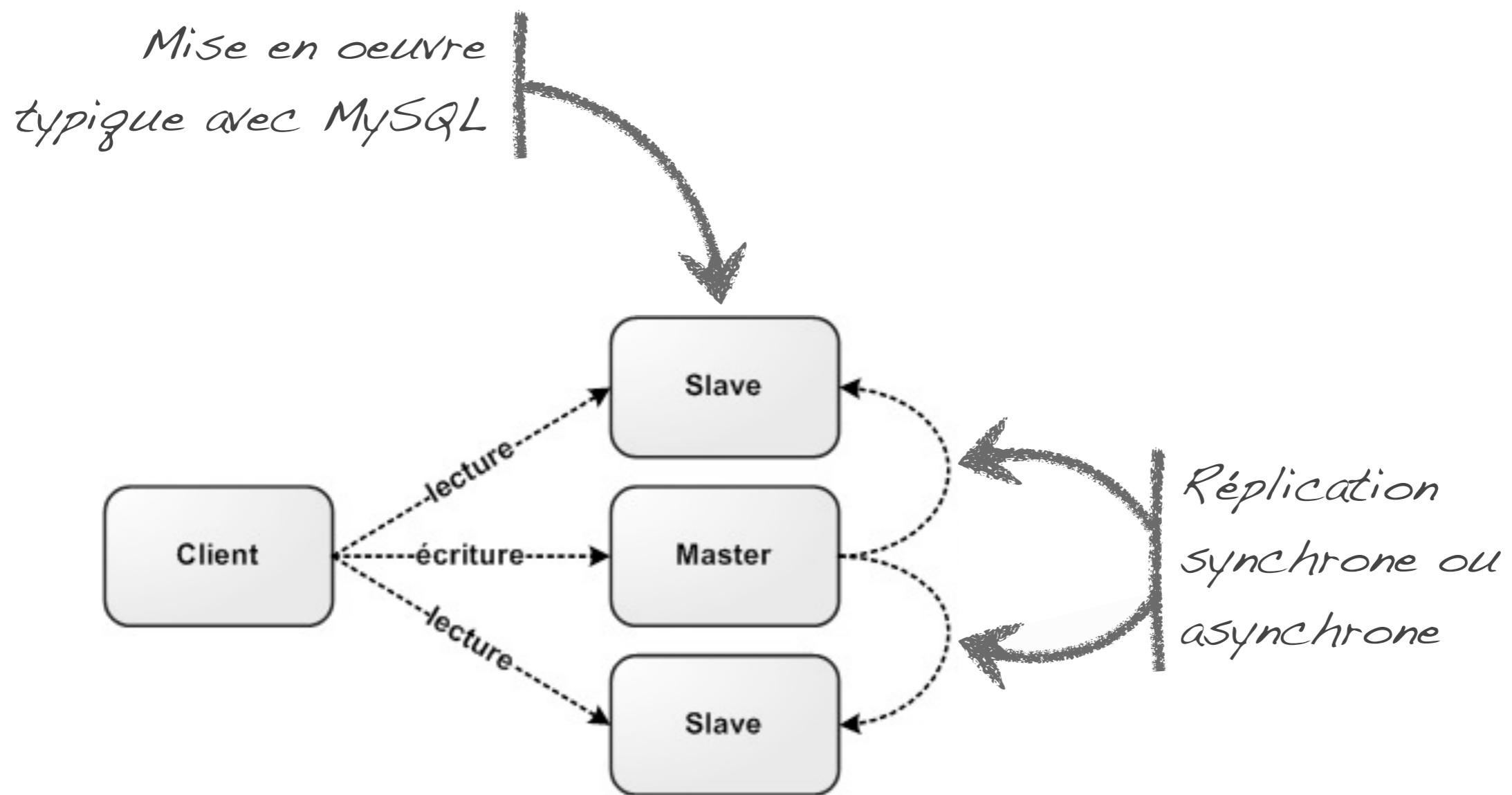
- Création de Dynamo
- Dernier incident majeur en 2004
- < 40 min d'indisponibilité par an



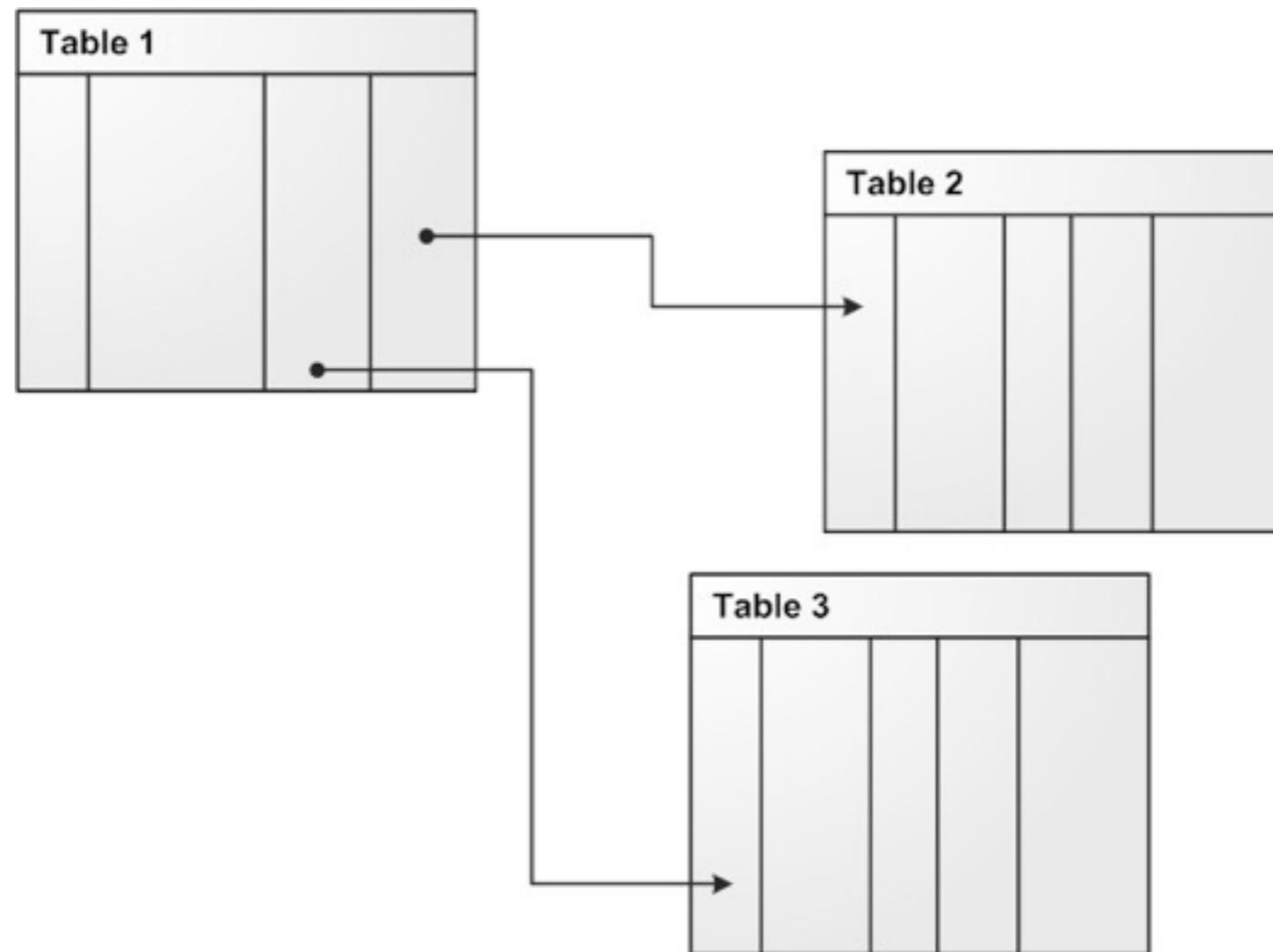
- Création de Cassandra
- Recherche dans les messages
- 500 millions d'utilisateurs

SGBDR et scalabilité

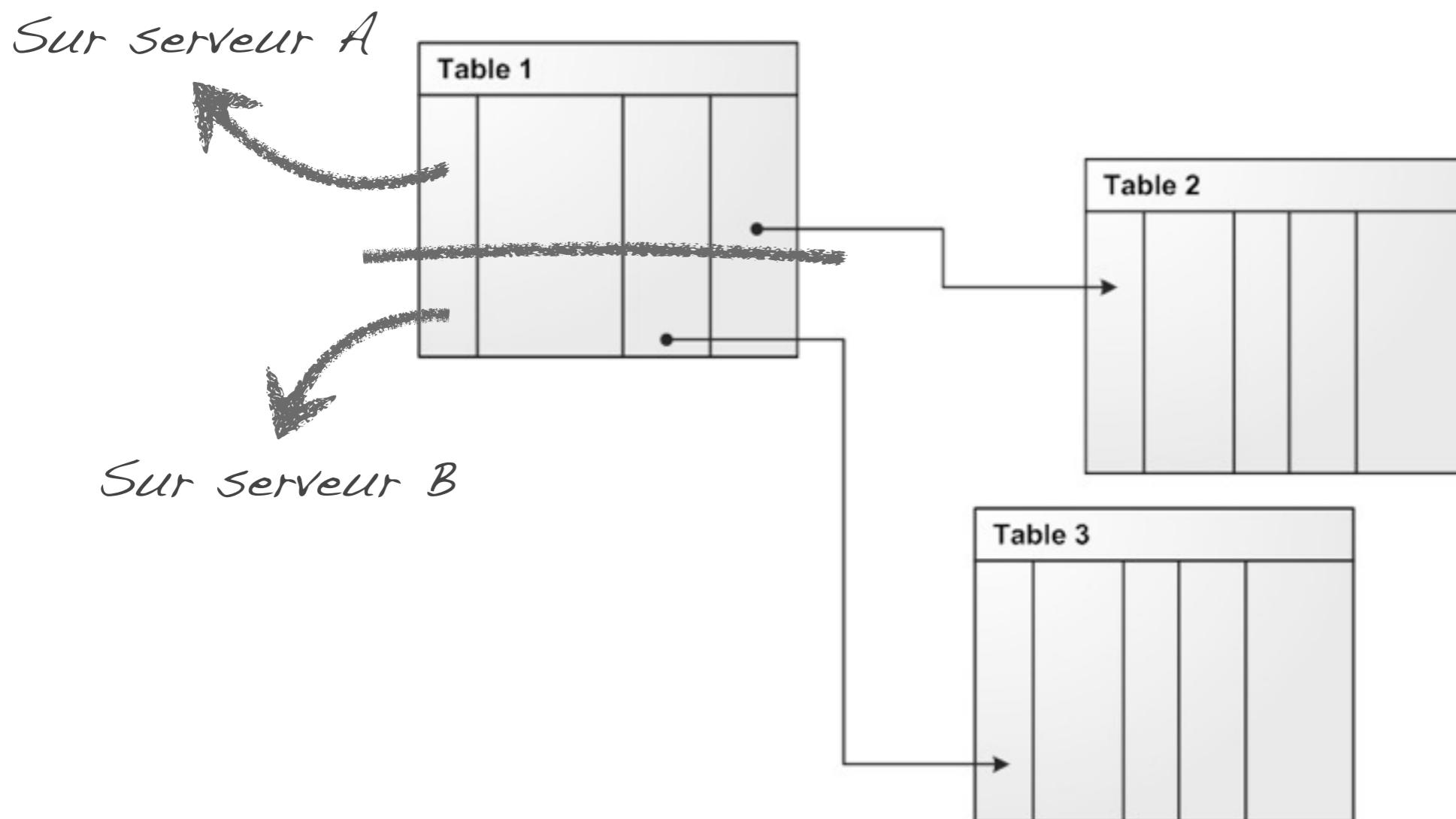
Comment assurer la scalabilité avec un SGBDR ?



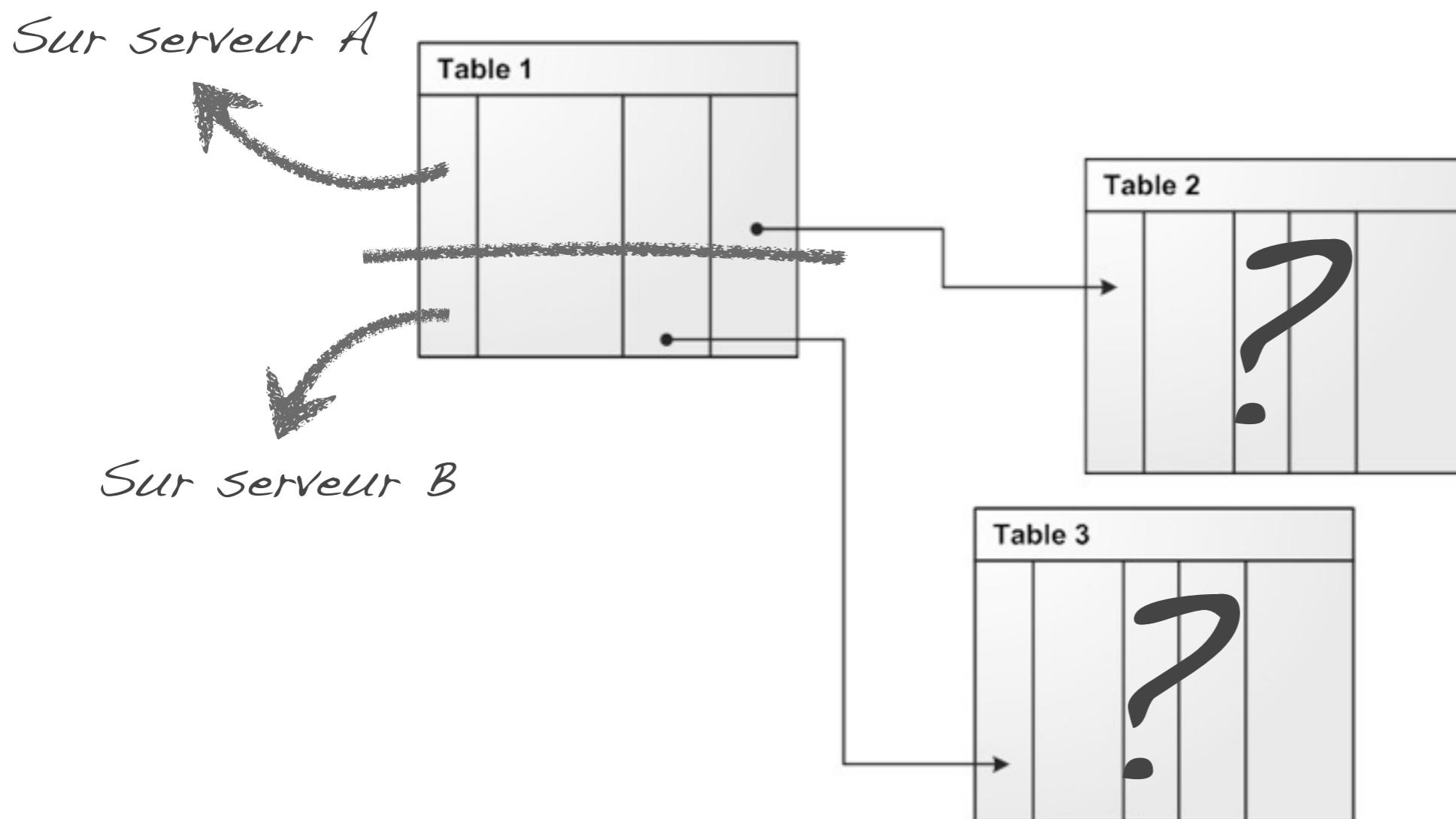
Sharding avec un SGBDR



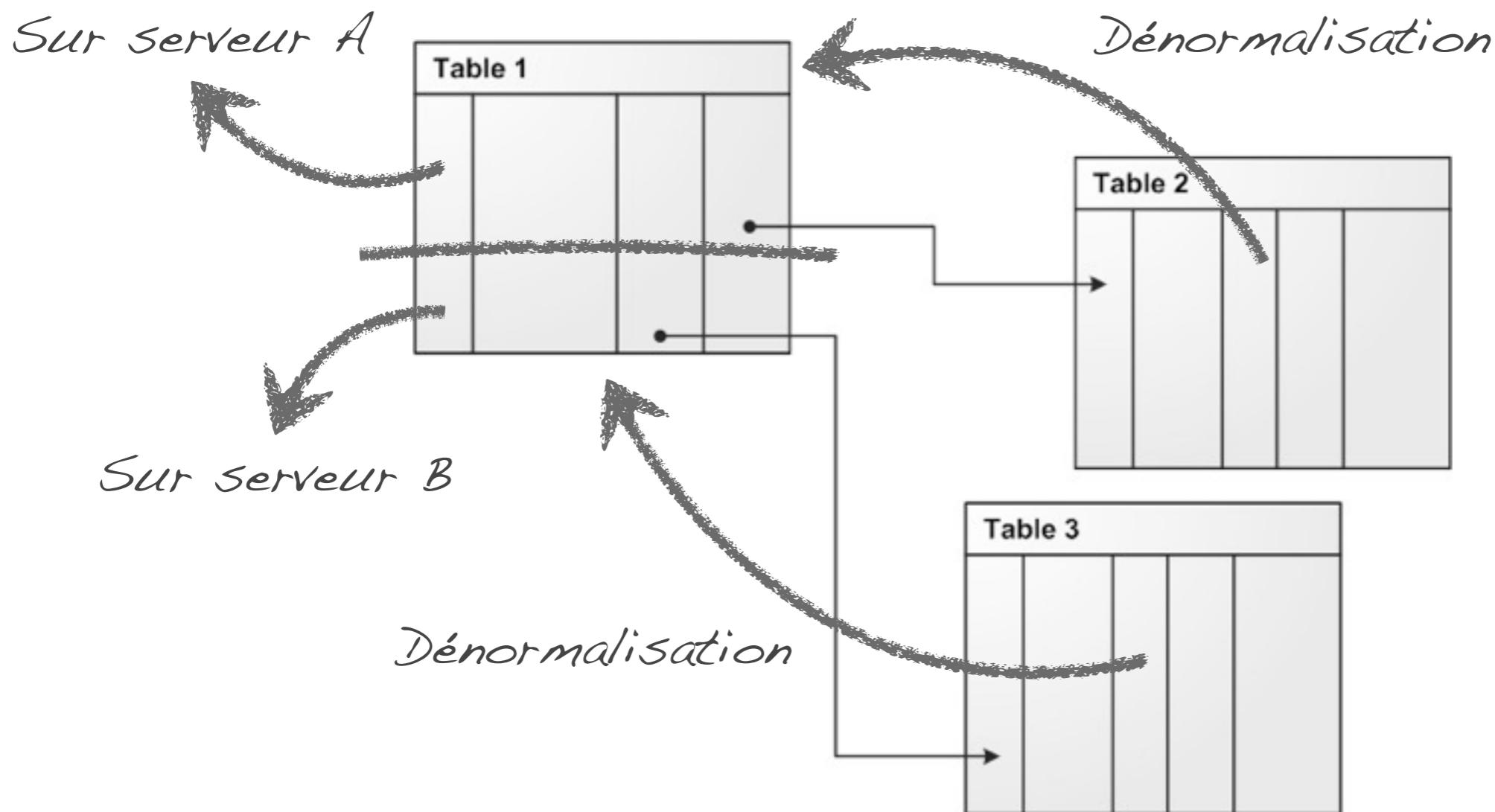
Sharding avec un SGBDR



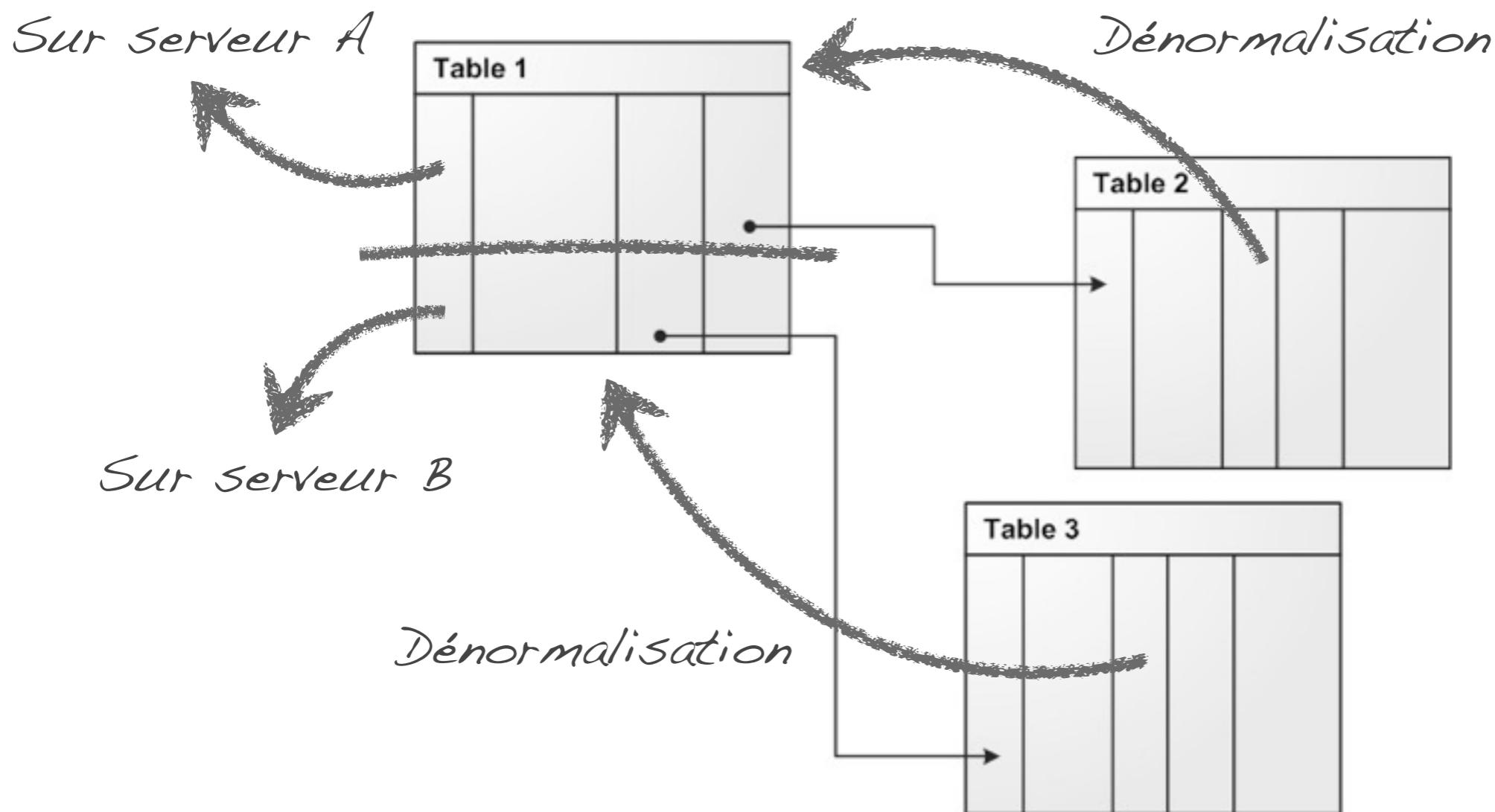
Sharding avec un SGBDR



Sharding avec un SGBDR



Sharding avec un SGBDR



On perd alors beaucoup de l'intérêt du relationnel !

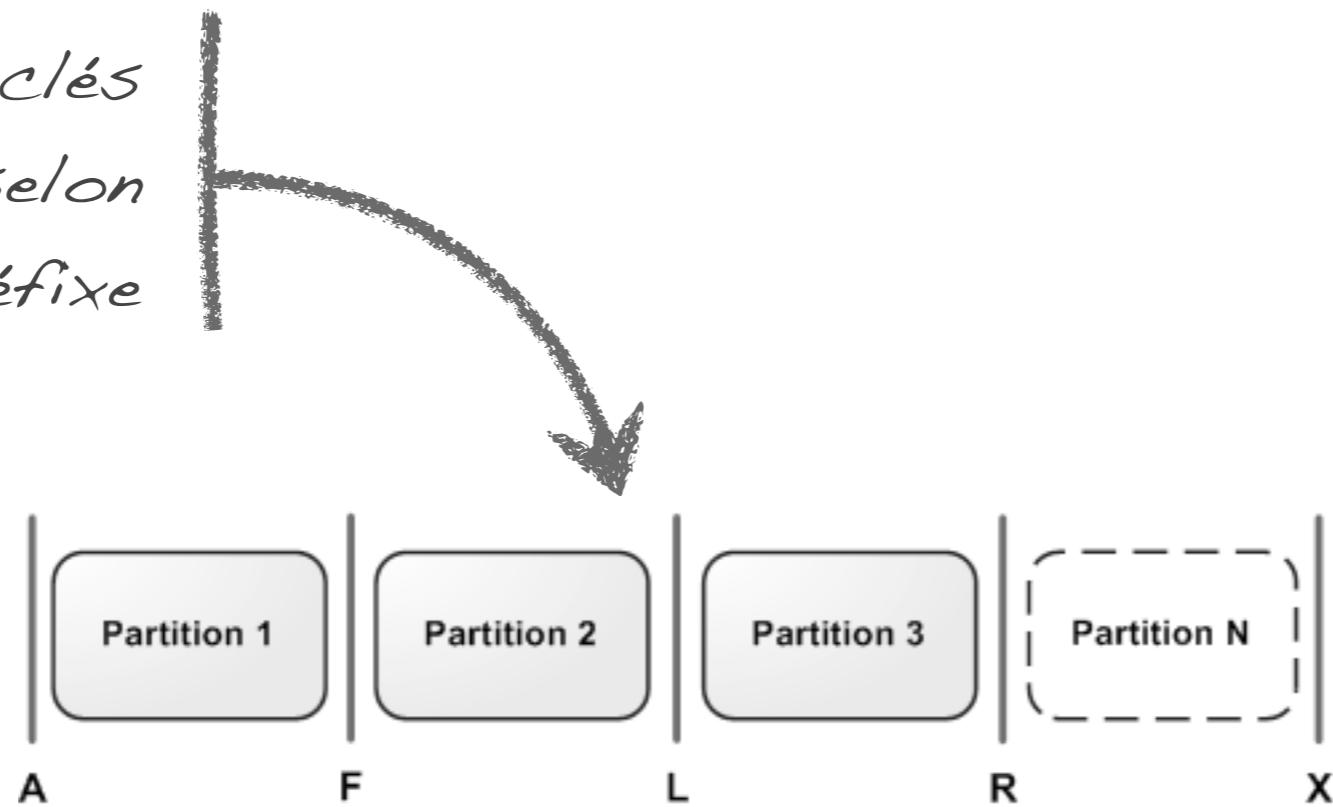
Sharding avec un SGBDR : les problèmes

- Pour garder de bonnes performances, les relations many-to-many et many-to-one nécessitent d'être dé-normalisées
- Gestion du resharding
- Code applicatif complexifié

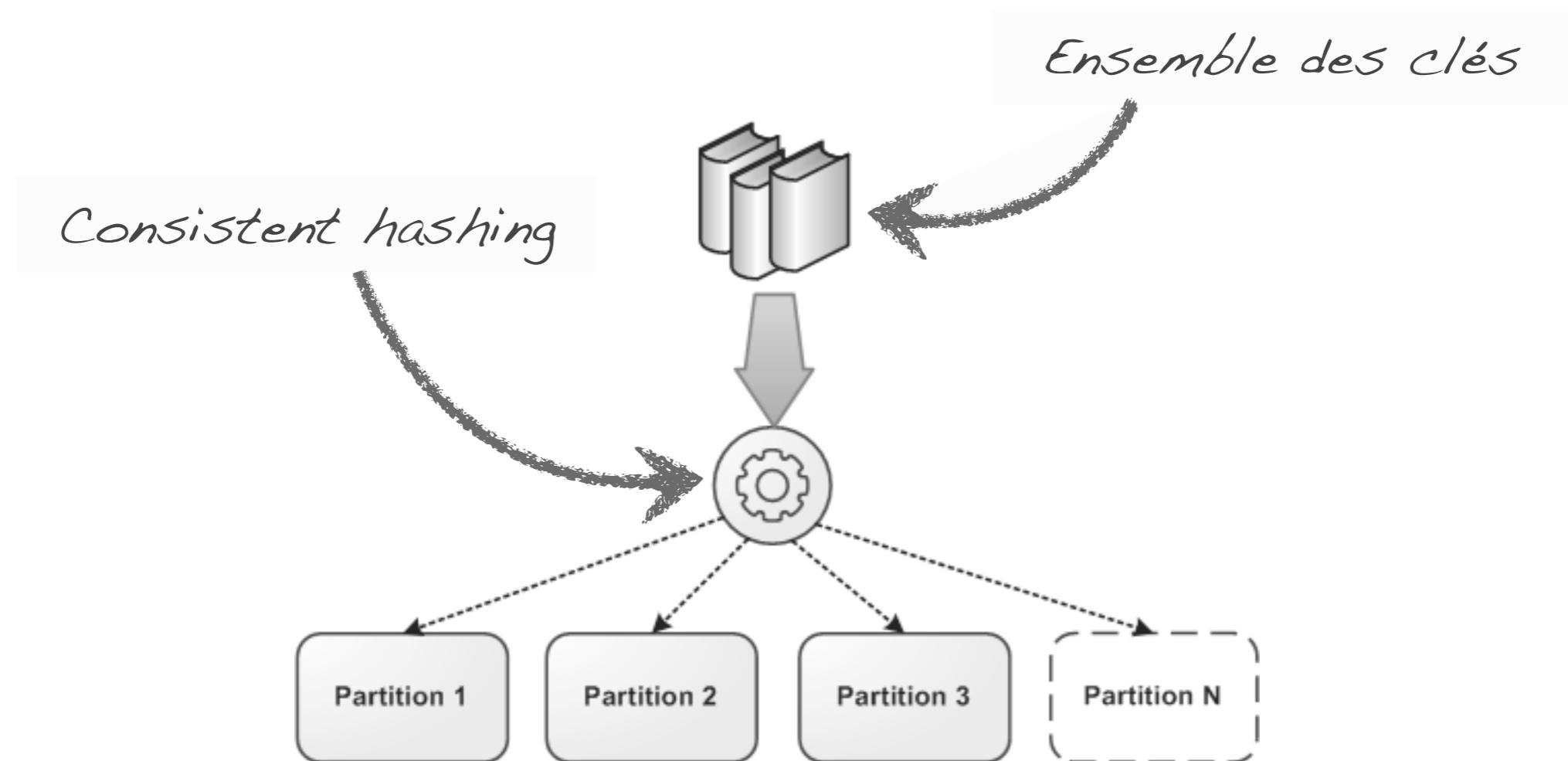
Une alternative

D'une table de hachage à une BDD clé-valeur

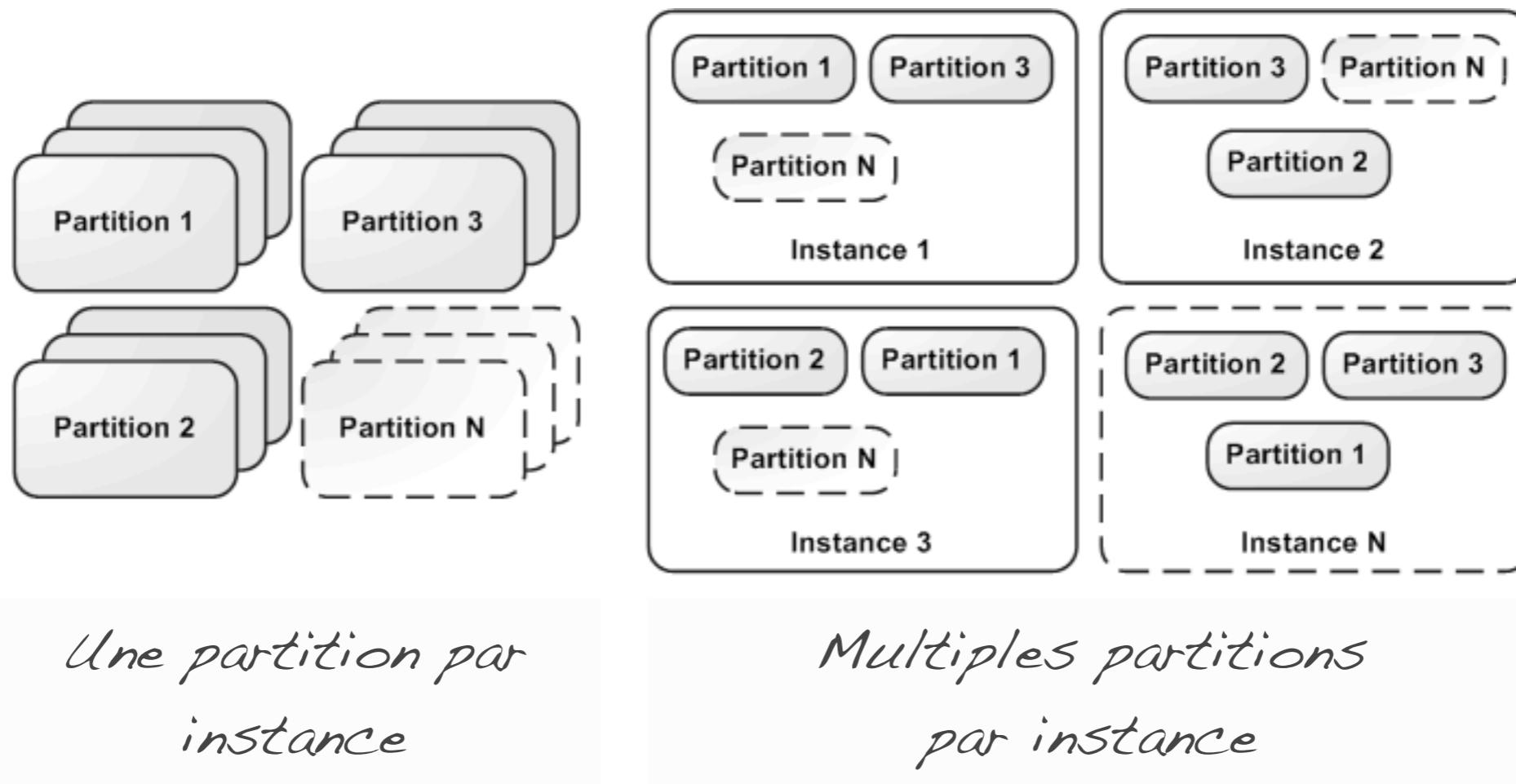
*Ensemble des clés
partitionnées selon
leur préfixe*



D'une table de hachage à une BDD clé-valeur

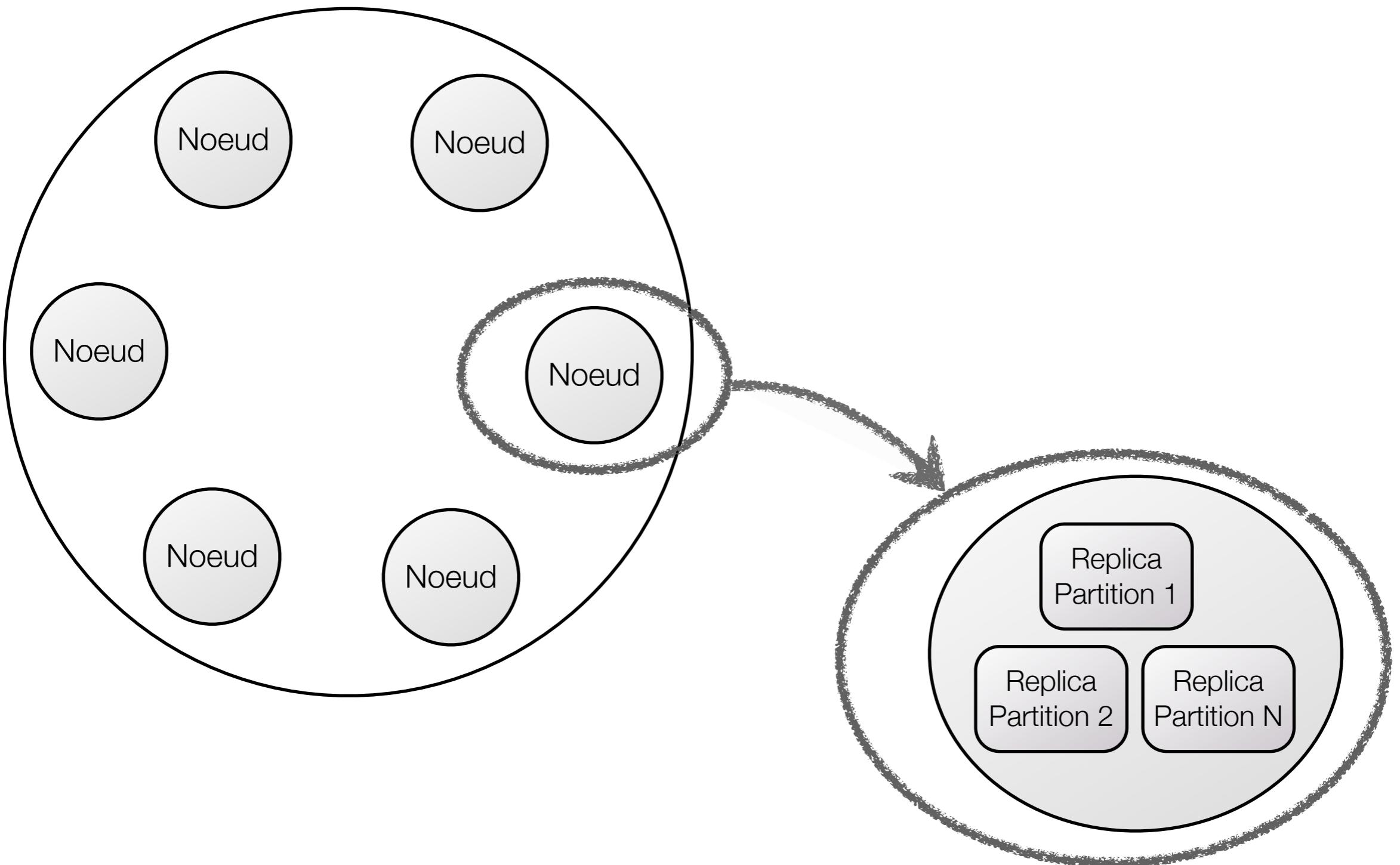


D'une table de hachage à une BDD clé-valeur

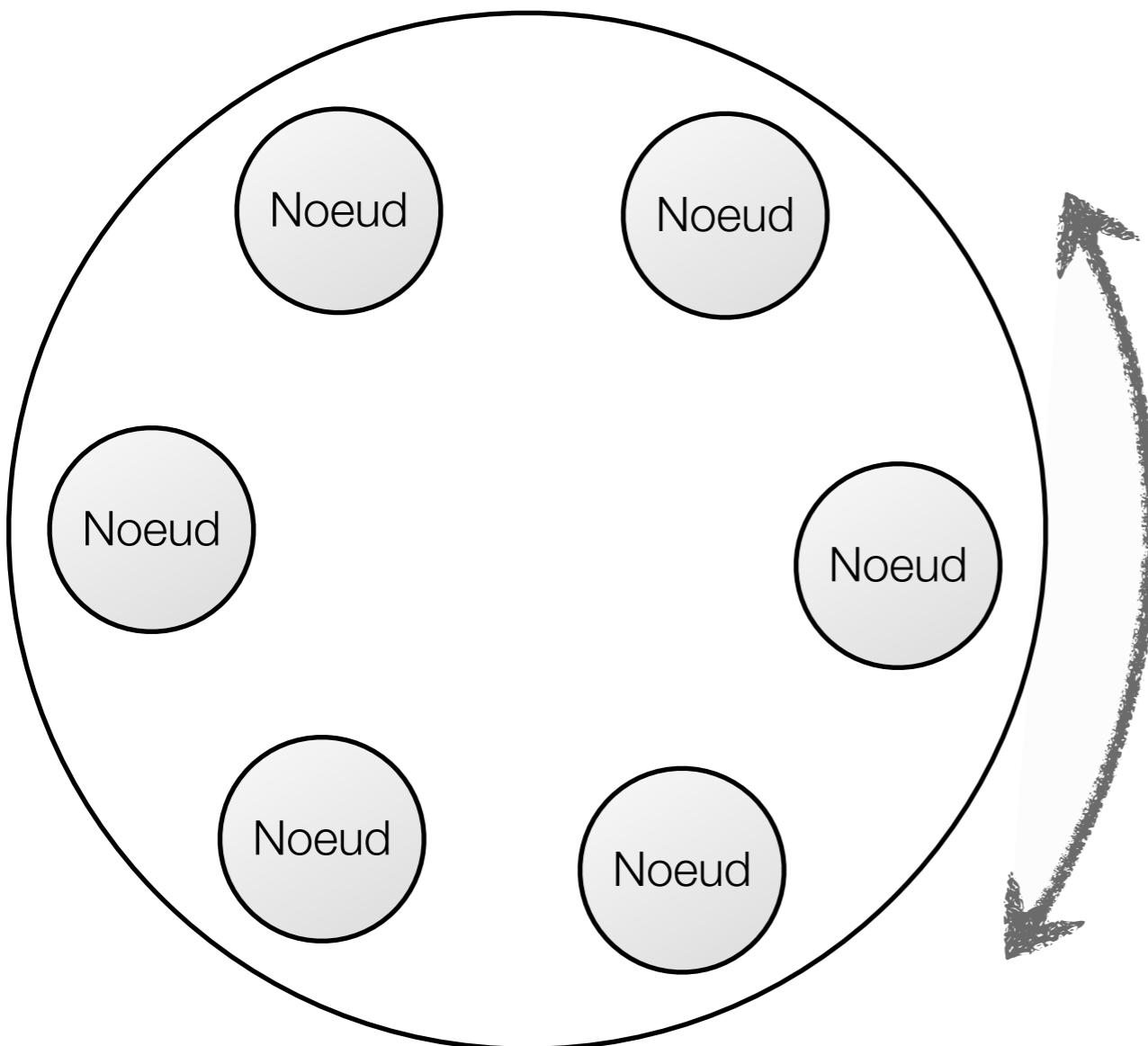


L'architecture de Cassandra

Organisation des noeuds en anneau

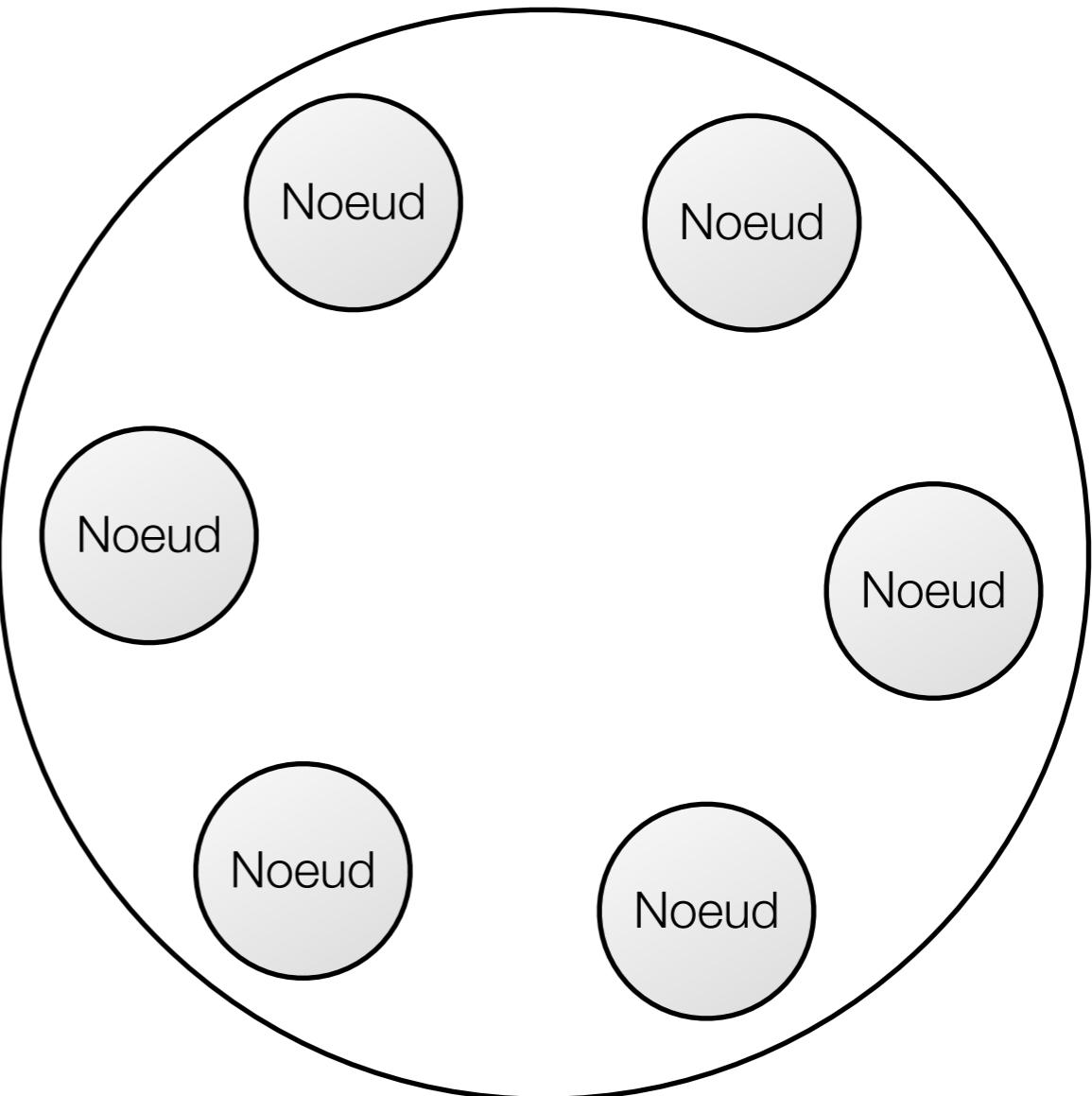
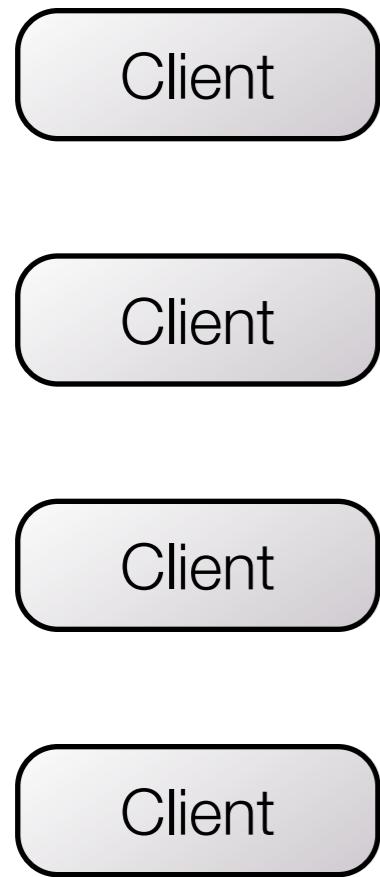


Organisation des noeuds en anneau

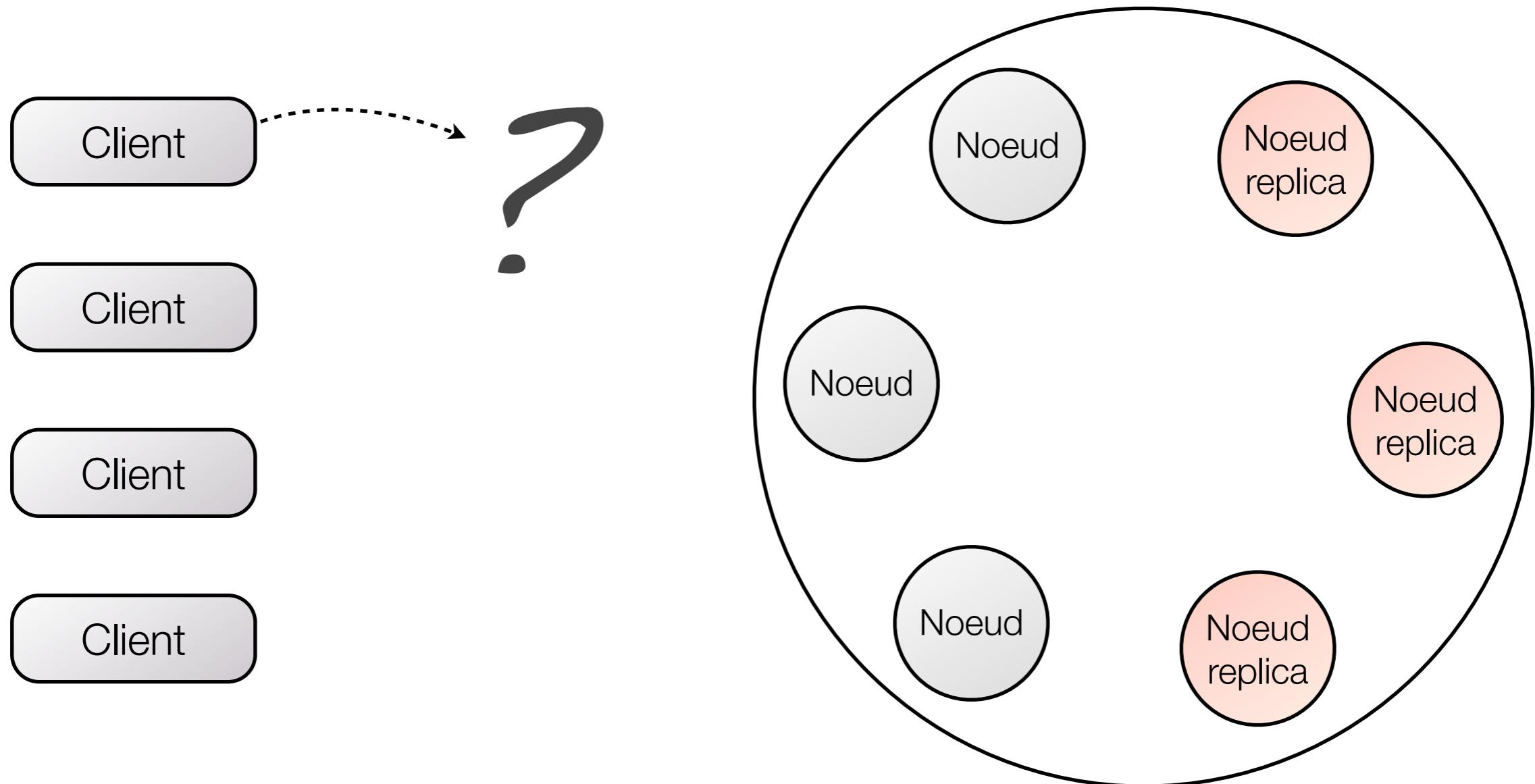


L'organisation en anneau permet de d'affecter un intervalle à chaque partition, facilitant ainsi les ajouts et suppressions d'instances

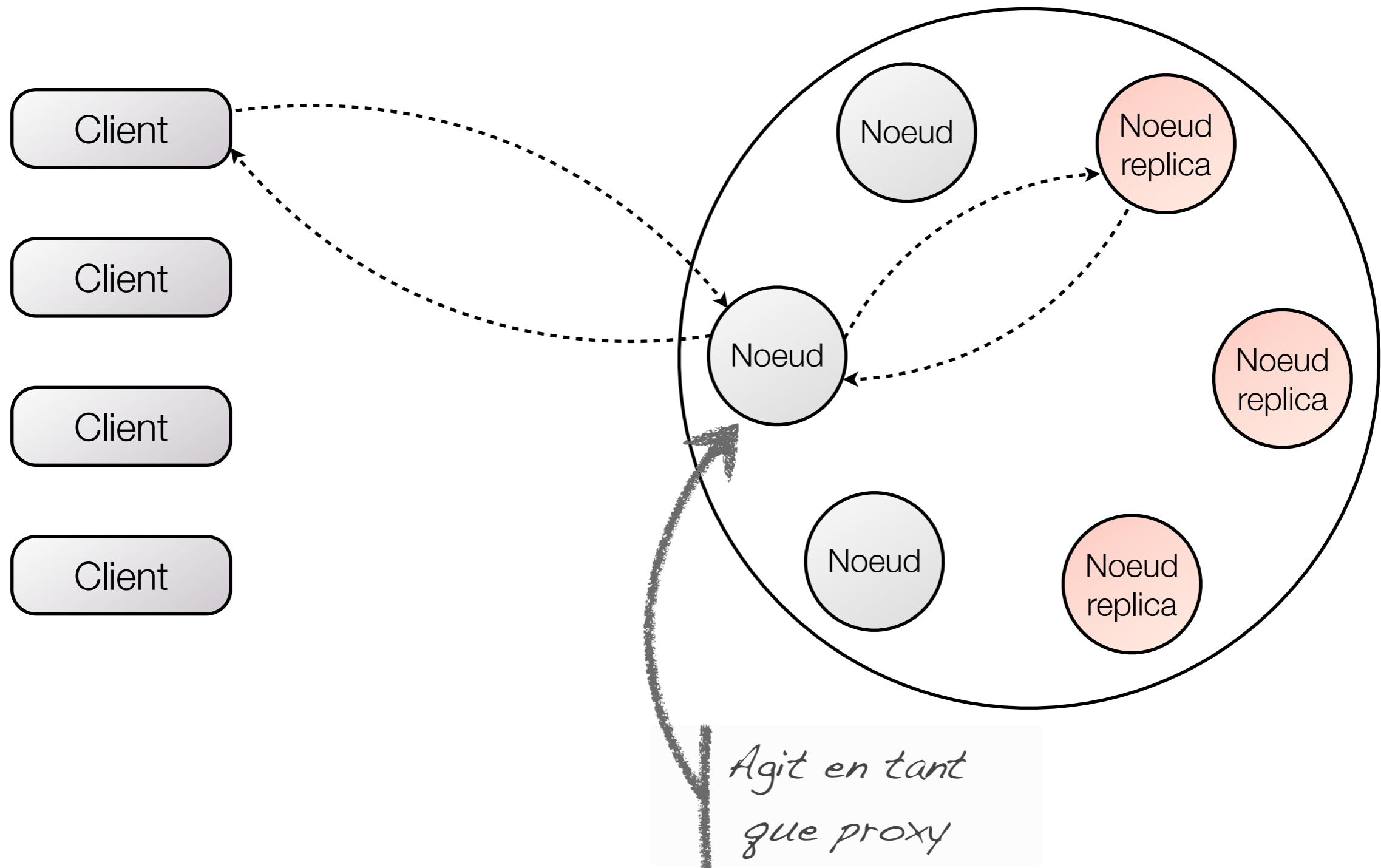
Interactions Client / Serveur



Interactions Client / Serveur



Organisation des noeuds en anneau



Gestion des défaillances

- Mécanisme d'anti-entropie, assurant des réplicas identiques

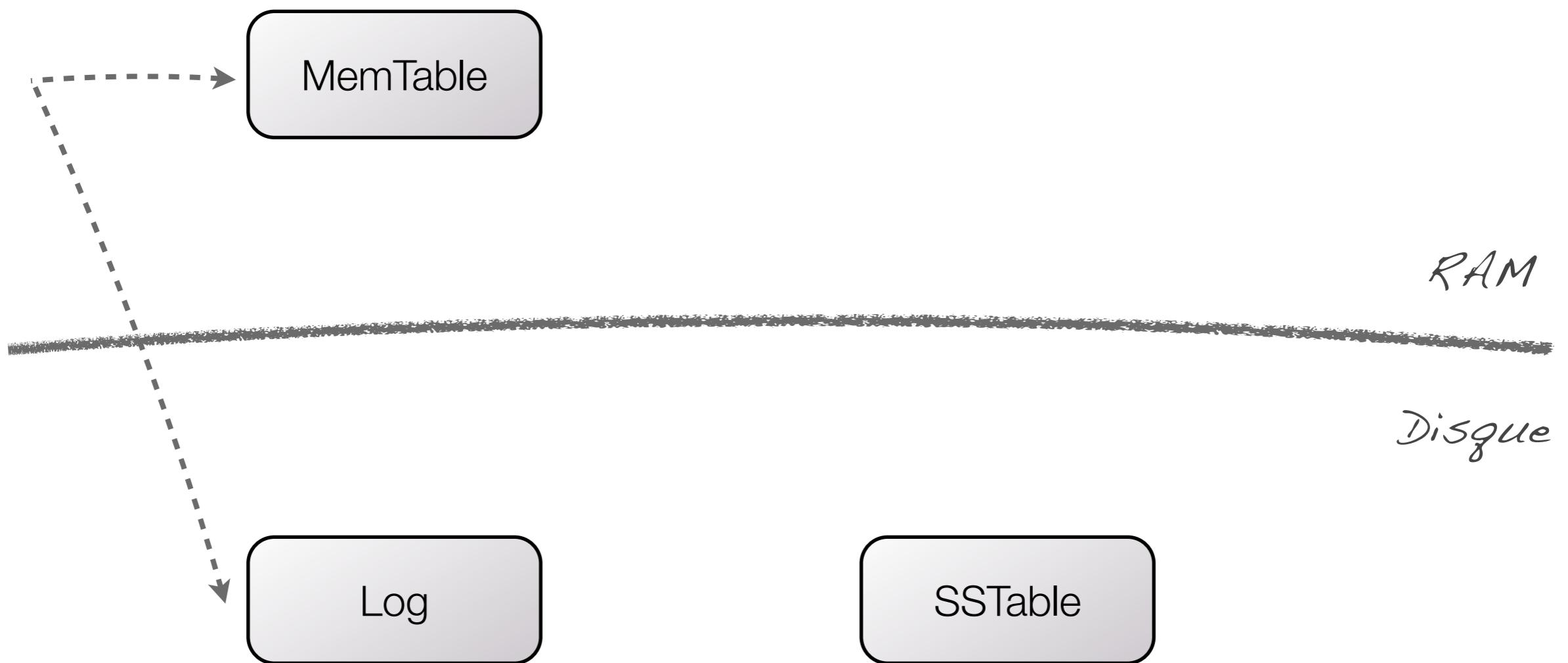
↳ *Echange des Hash des données entre réplicas*

- Hinted-Handoff stocke les écritures pendant l'absence d'un noeud

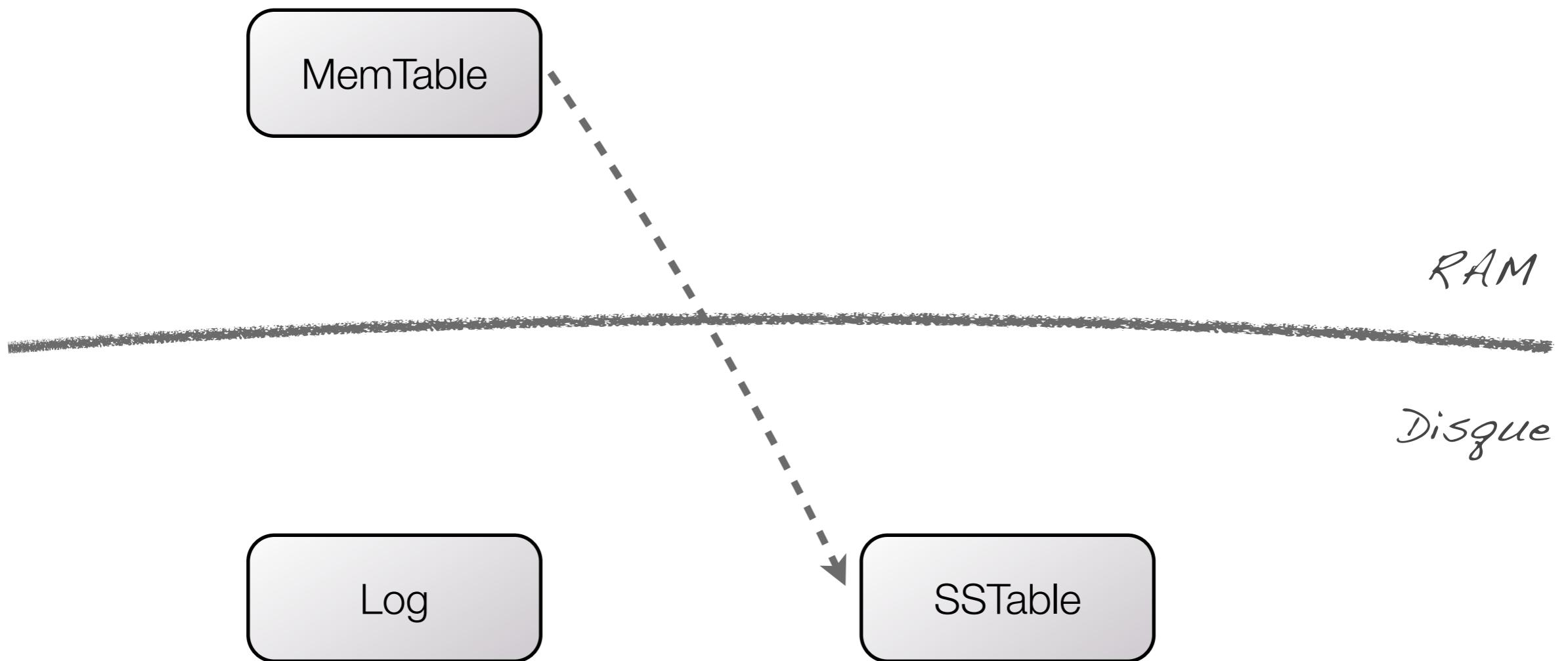
↳ *Semblable à une prise de messages*

Stockage sur disque

Architecture append-only de Cassandra



Architecture append-only de Cassandra

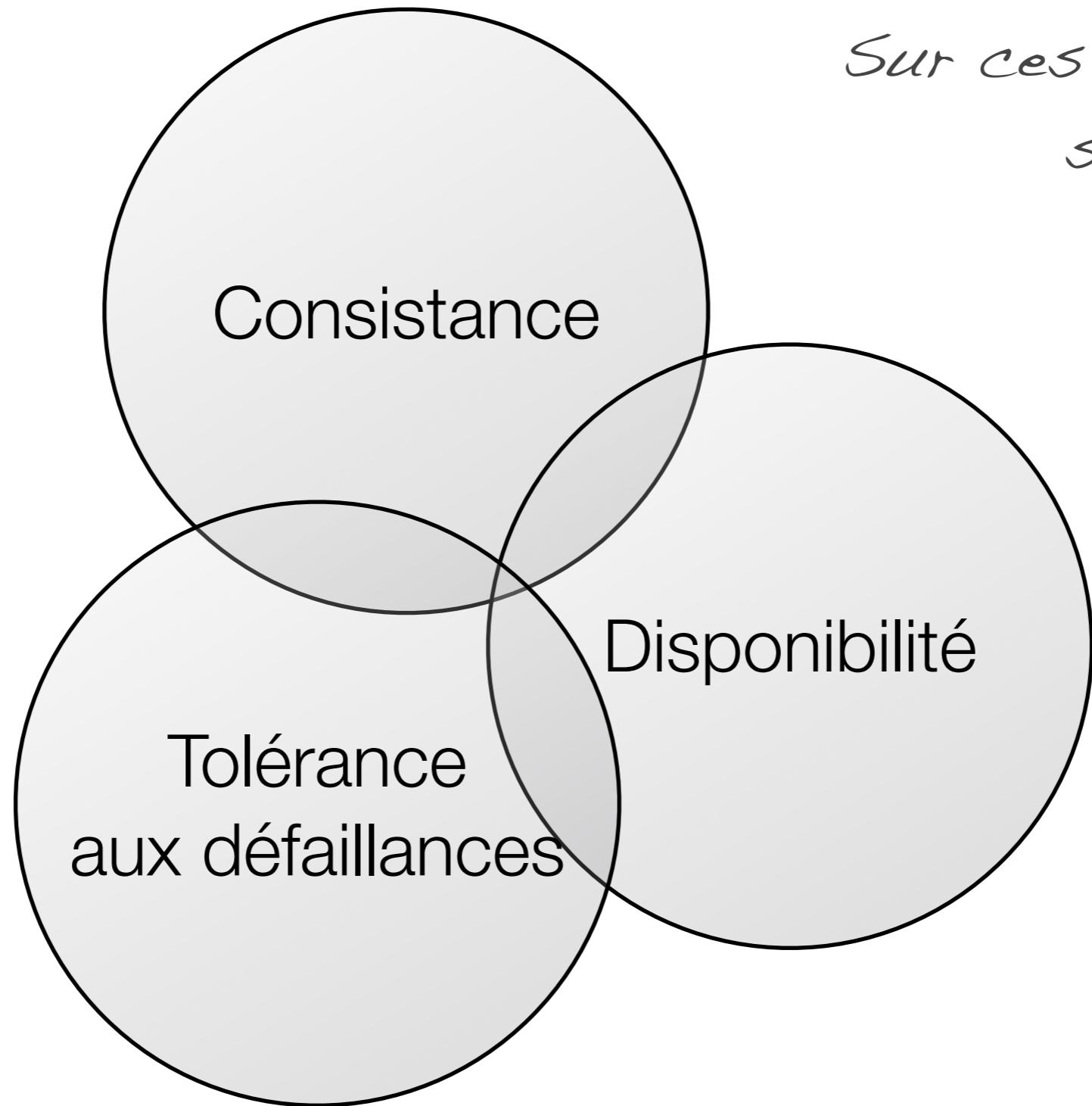


Que devient ACID ?

Que devient ACID ?

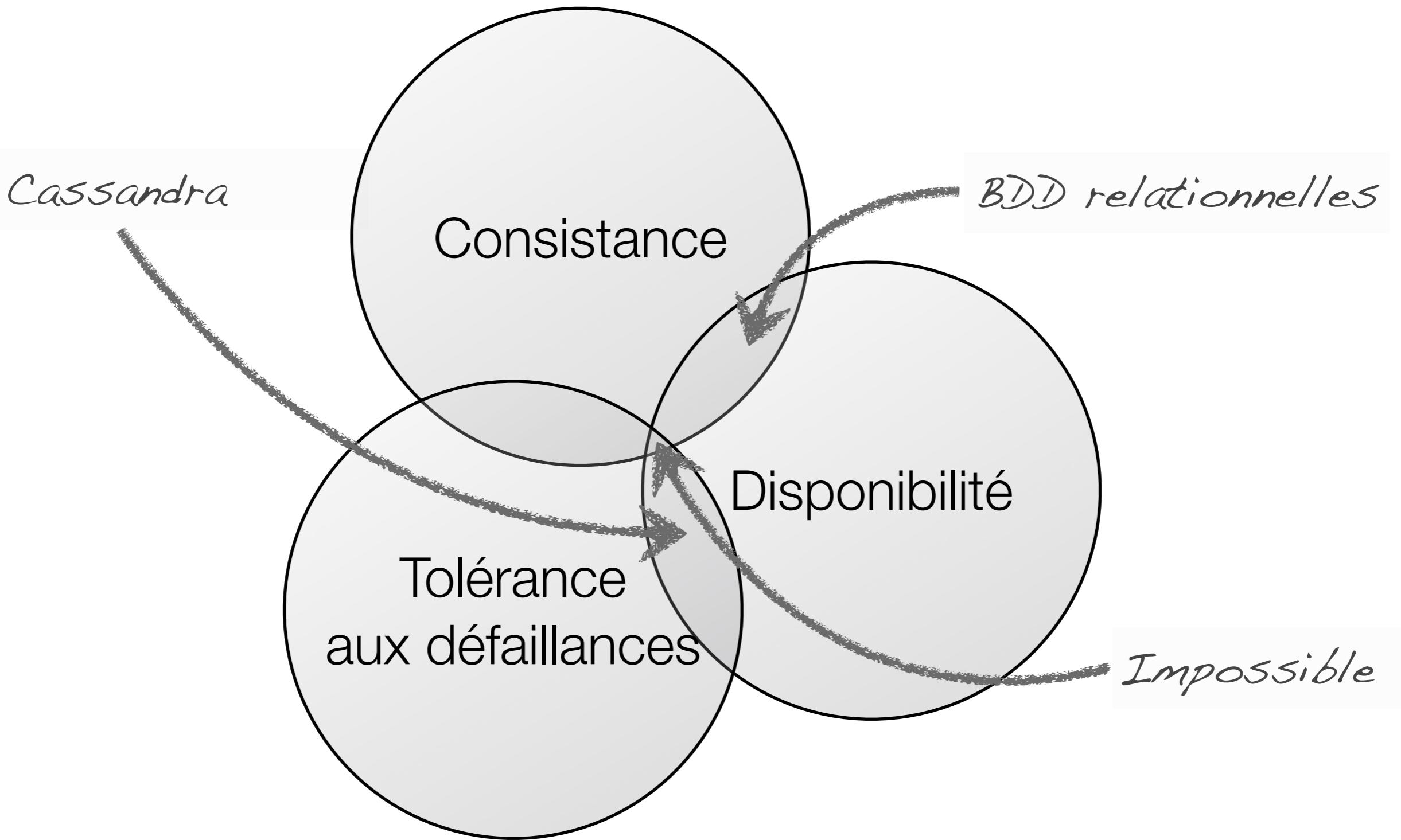
- Tout accès réseau est faillible
- Des concessions doivent être faites sur le modèle de données
- Des concessions doivent être faites sur la consistance

Le théorème CAP

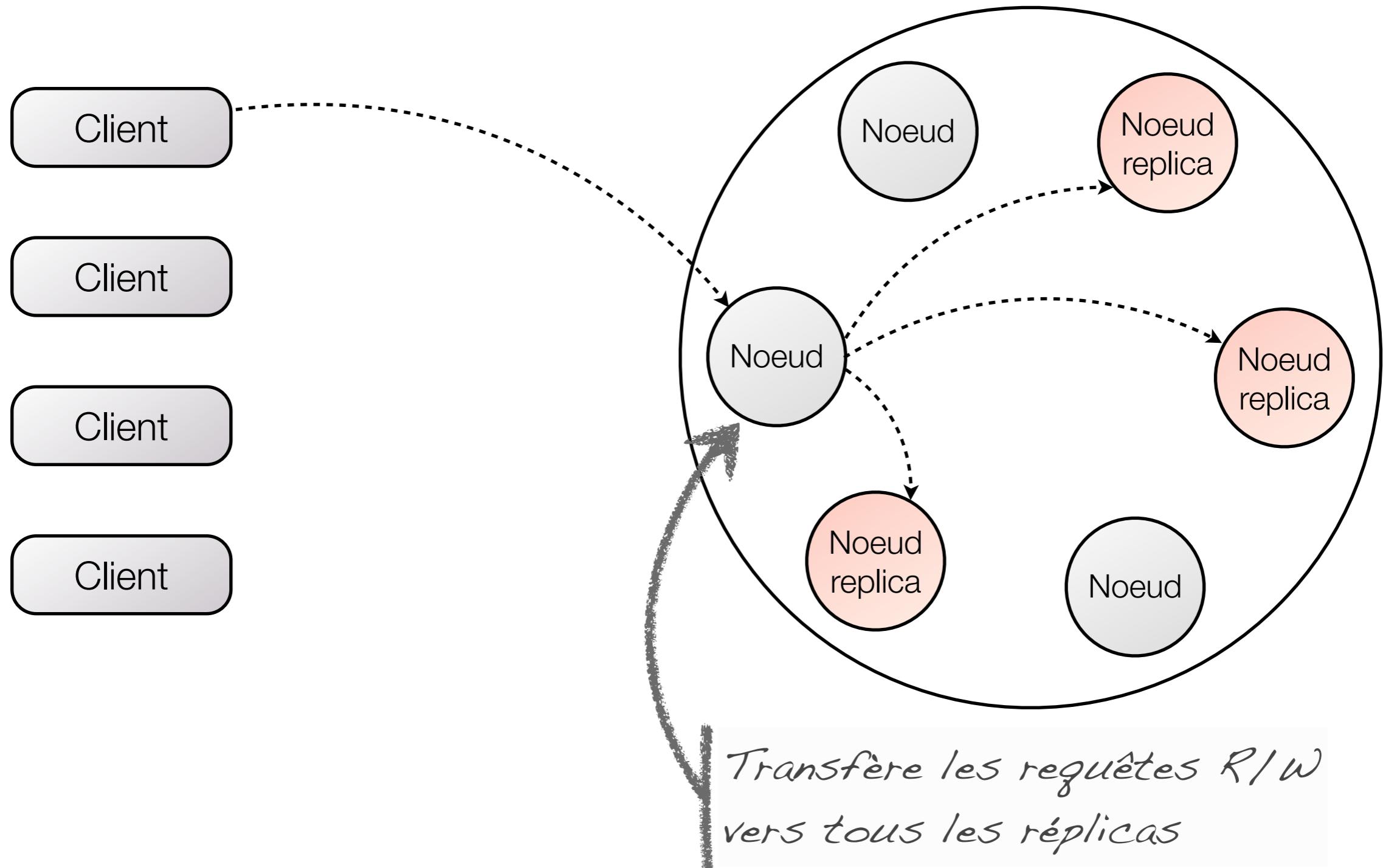


Sur ces 3 propriétés,
seules 2 sont
réalisables
à la fois

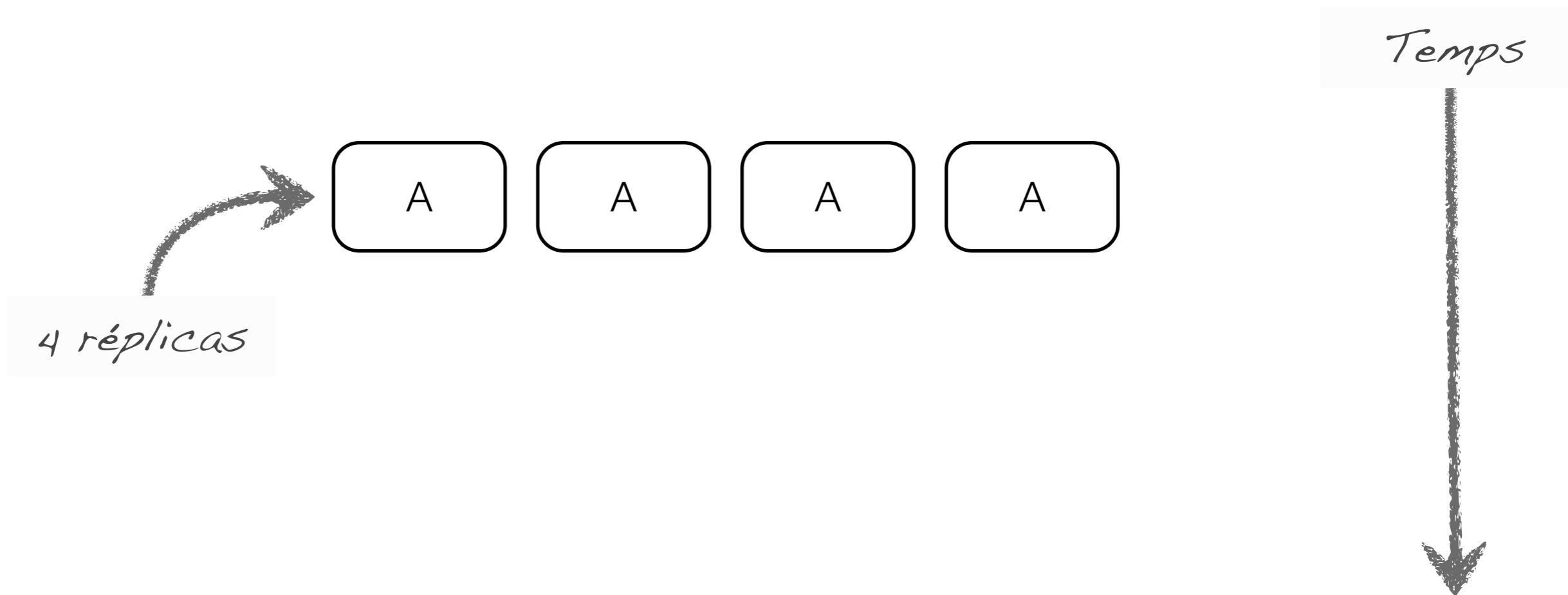
Le théorème CAP



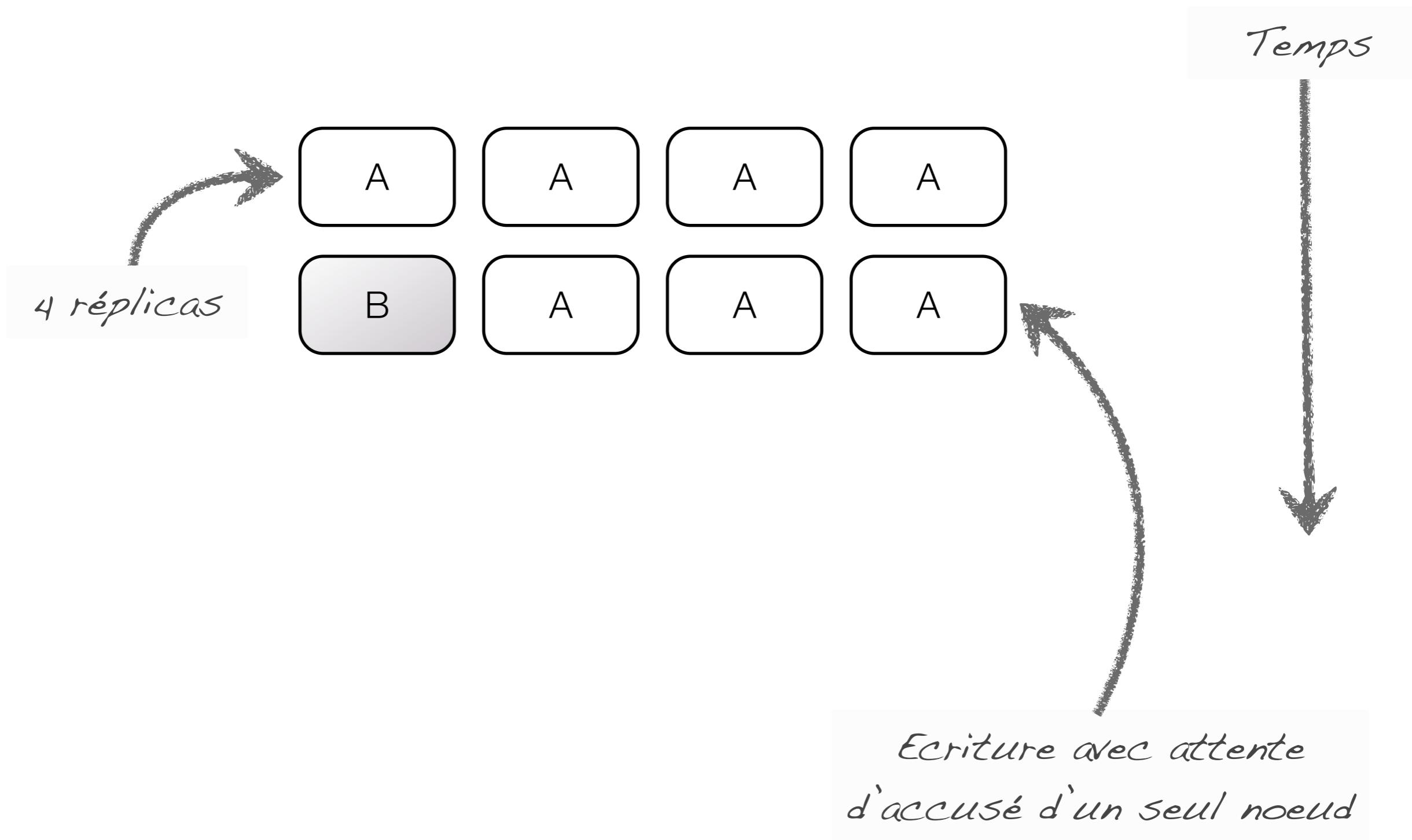
Consistance éventuelle



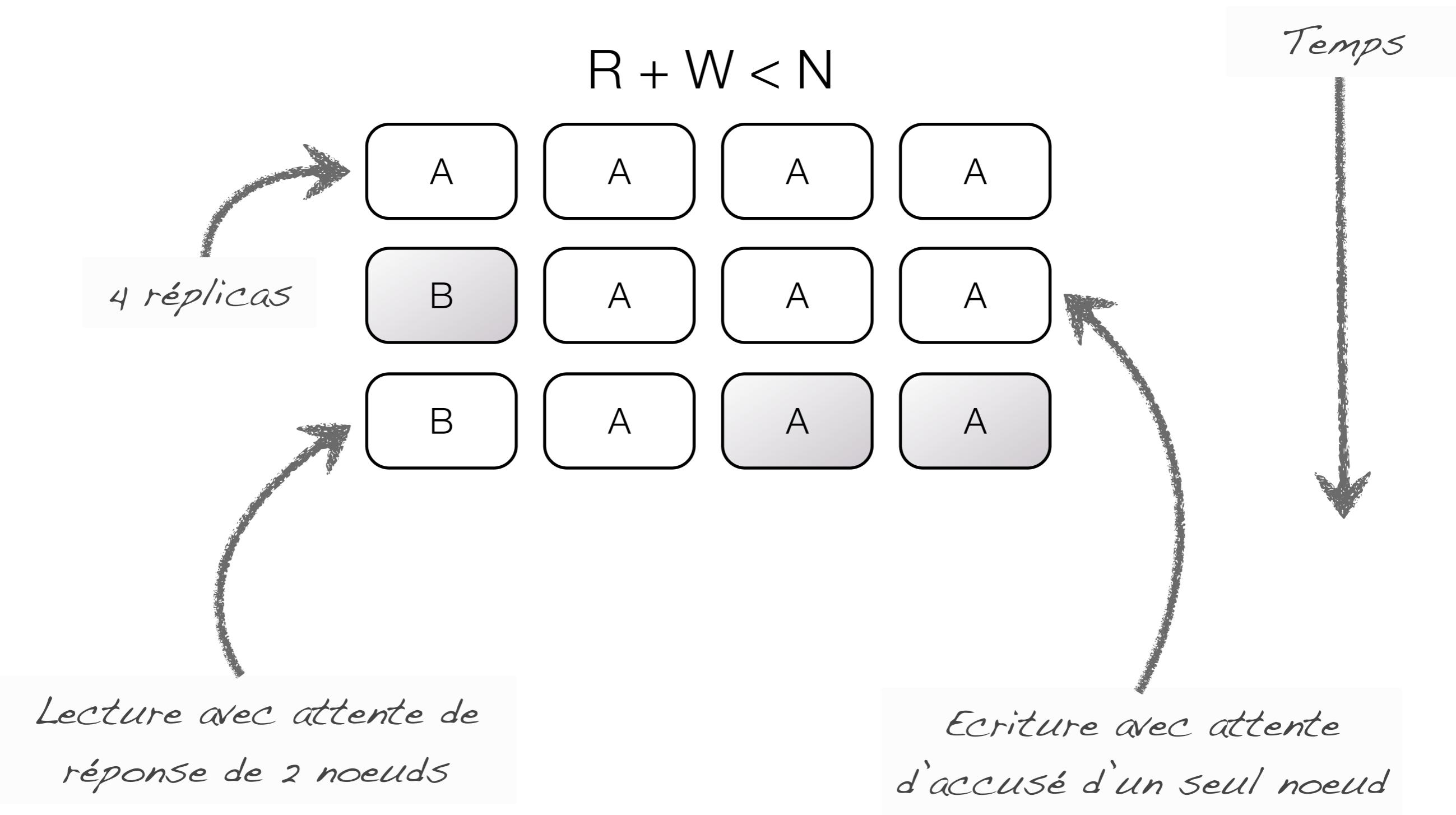
Consistance selon nombre de réponses attendues



Consistance selon nombre de réponses attendues

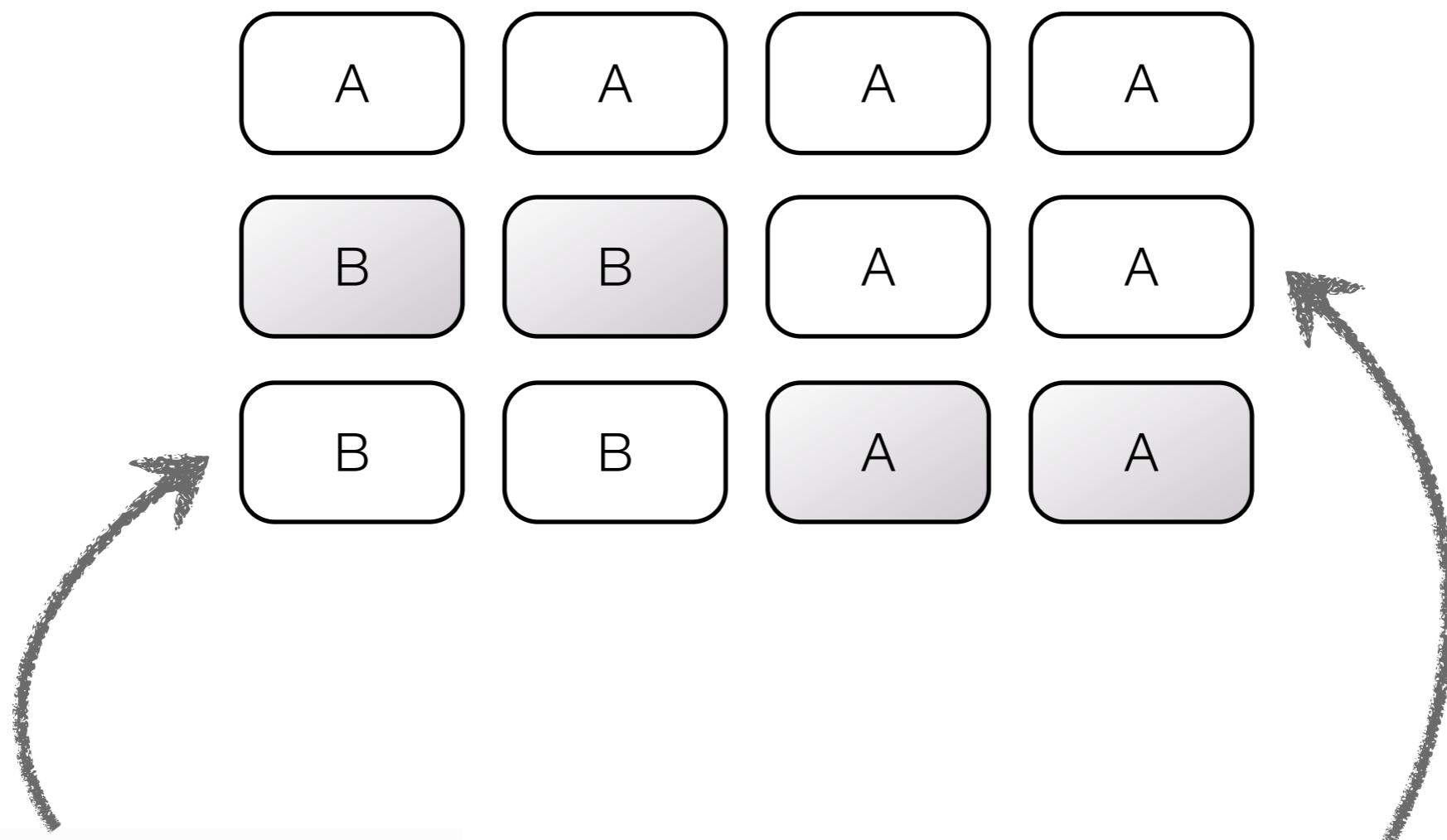


Consistance selon nombre de réponses attendues



Consistance selon nombre de réponses attendues

$$R + W = N$$

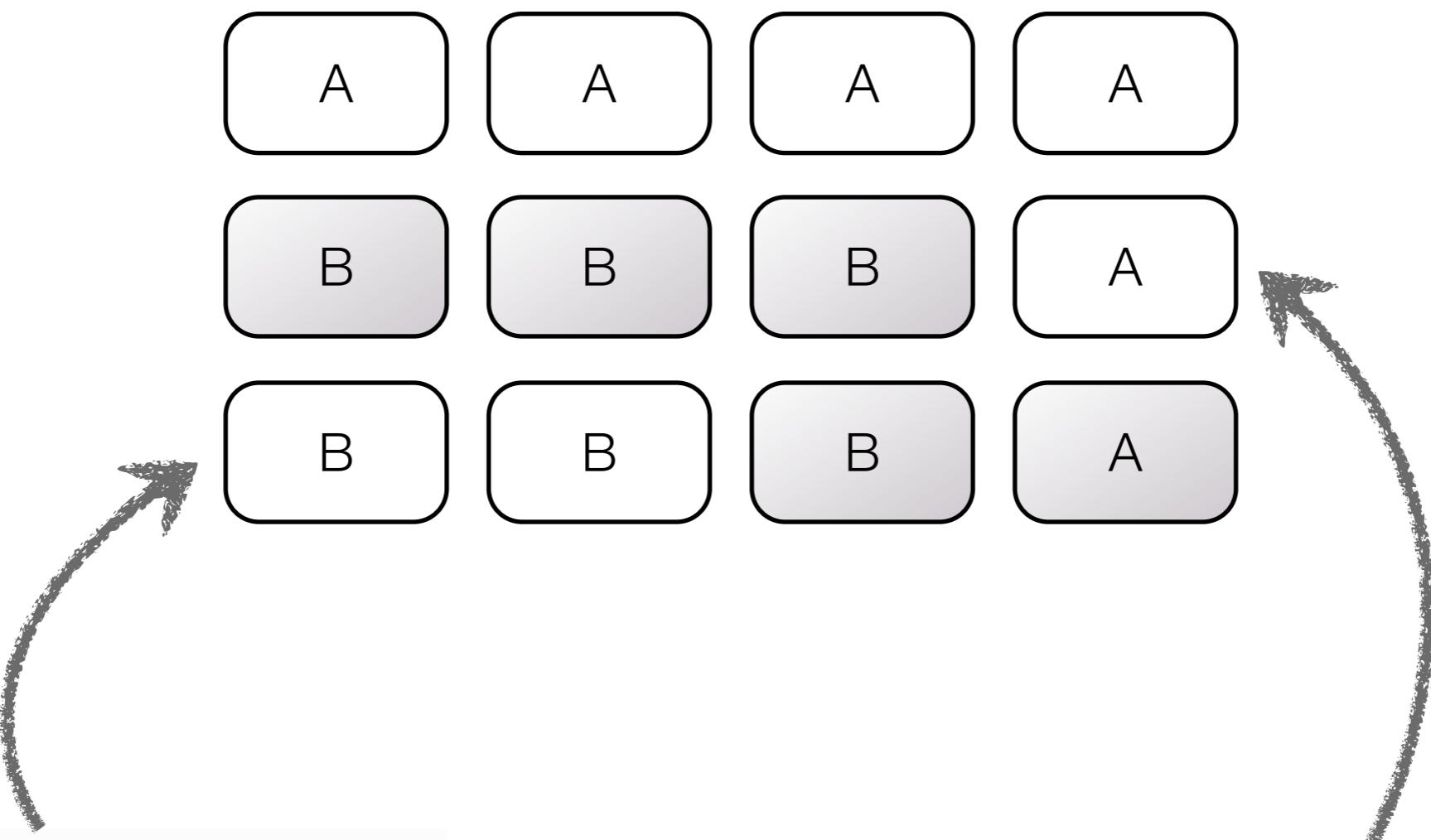


Lecture avec attente de
réponse de 2 noeuds

Écriture avec attente
d'accusé de 2 noeuds

Consistance selon nombre de réponses attendues

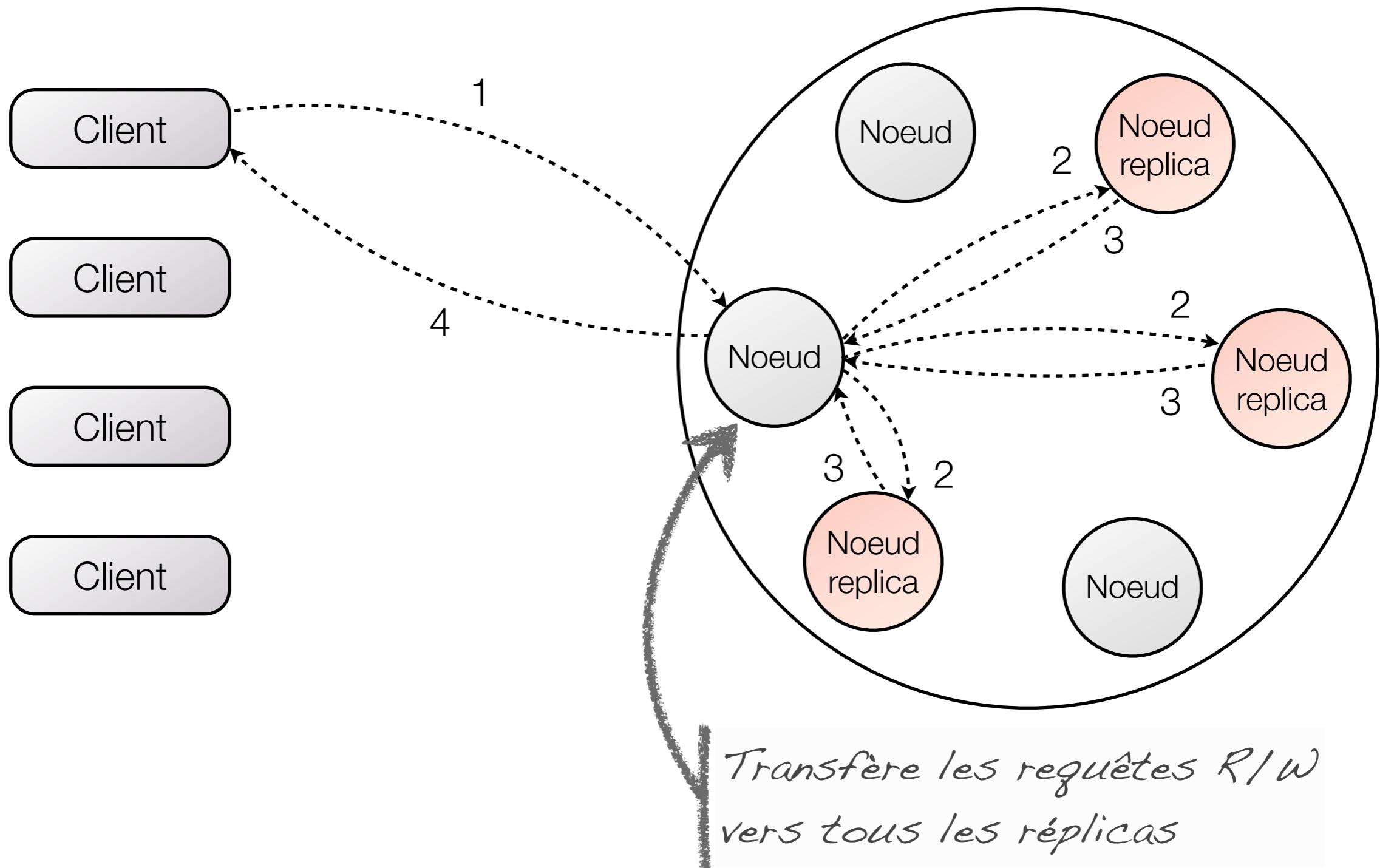
$$R + W > N$$



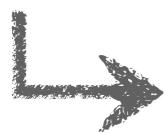
Lecture avec attente de
réponse de 3 noeuds

Écriture avec attente
d'accusé de 2 noeuds

Consistance apparente pour le client



Atomicité et Isolation

- Les données ne sont plus co-localisées
 *Localisation non prédictible dans le temps*
- Les transactions distribuées nuiraient à la disponibilité et aux performances
- Atomicité et Isolation par opération sur une clé

Durabilité

- Ecriture sur un ou plusieurs disques

↳ *La réPLICATION permet de renforcer la durabilité*

- Ecriture multiples en mémoire

↳ *La réPLICATION apporte la durabilité*

- En mémoire avec écriture asynchrone sur disque

↳ *Pas de durabilité*

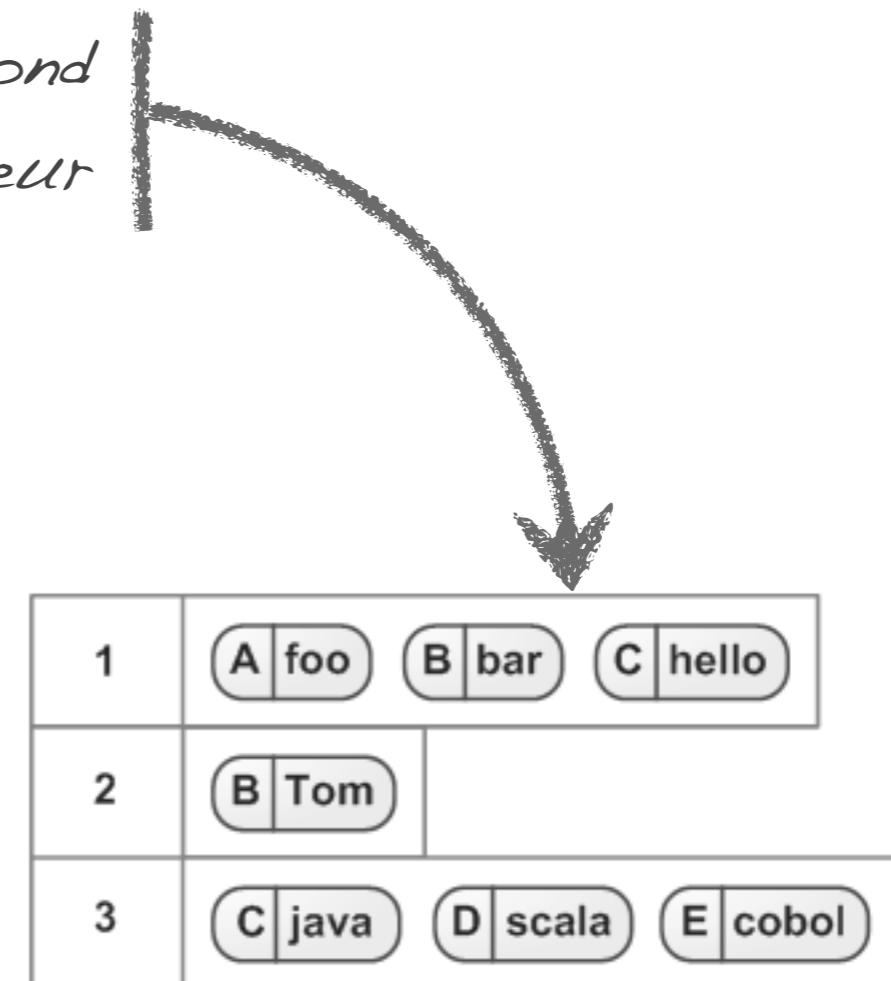
Modèle de données

Le modèle en famille de colonnes

A chaque ID de ligne correspond
une liste de couples clé-valeur

	A	B	C	D	E
1	foo	bar	hello		
2		Tom			
3			java	scala	cobol

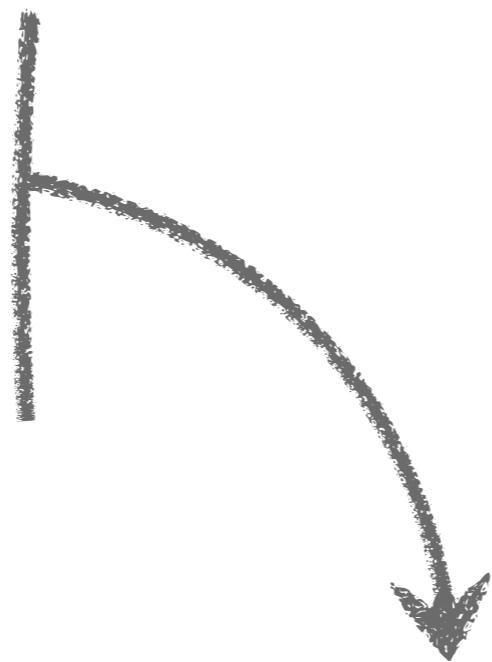
BDD relationnelle



BDD orientée colonnes

Super-colonnes

Les valeurs d'une super-colonne sont des collections de colonnes



1	Choses	A foo	B bar	C hello
2	Choses	C texte12	D texte	Personnes B Tom
3	Langages	C java	D scala	E cobol

Un modèle de données peu intuitif...

Re: Is SuperColumn necessary? Schubert Zhang to user

I don't know... I am not sure if it's necessary or not.

Correct data model for Cassandra Oleg Ivanov to user

Hello.
our company has a huge table in a relational database. It looks like the following:
SELECT COUNT(*) FROM account1 WHERE id = 'account2'
It's a huge table in a relational database, which served the transactional operations.

WTF is a SuperColumn? An Intro to the Cassandra Data Model Digg's data model

I am a newbie to bigtable like **model** and have to find a list users who dug a URL and also want to take privacy concerns into account, such that I am evaluating Cassandra and playing with some transactions.

Model Question Hi, I can't figure out how to use **model** sorted (by their name). Lets say I have a list of users in [cassandra-user@incubator.apache.org](#) on Mar 22 by [Julio Carlos Barrera Juez](#) — replies: 11

Modeling question Hi everyone, in my team, we are considering a (pseudo)relational solution, but I am not sure if it's the right way to go. In [cassandra-user@incubator.apache.org](#) on Mar 22 by [Julio Carlos Barrera Juez](#) — replies: 11

data model question Hi. I need to store a bunch of 'models' for **model** is right (seems to be the norm at first sight). In [cassandra-user@incubator.apache.org](#) on Nov 3 by [Julio Carlos Barrera Juez](#) — replies: 11

Digg's data model I am a newbie to bigtable like **model** and have to find a list users who dug a URL and also want to take privacy concerns into account, such that I am evaluating Cassandra and playing with some transactions.

How to model hierarchical structure? Hi all, I need an example of cassandra storage to understand better the datamodel. I mean, I need an example of a little application with his storage-conf.xml file like a little commerce with a list of categories, taxonomy, or folder/file, there will be multiple level hierarchical relationships.

Yet Another Data Model Description Hi all, just a quick note to make you aware of the (n+1)th Data Model post on my blog (<http://schabby.de/cassandra-getting-started/>). Comments/corrections are very welcome.

Model to store biggest score Hi, I'm trying to find a **model** where I can keep the list of biggest scores for users. It's stuck here. For example user1 score = 10 user2 score = 20...

Modelling assets and user permissions ...model the relationship between user and asset. I think I sort of see a second level of inheritance here.

Re: Inserting rows in columns inside a supercolumn OK, I have solved my problems with Cassandra data model. Now I am using Column Families of type Super and SuperColumns with many columns inside.

Thanks!
2010/4/16 Julio Carlos Barrera Juez <juliocarlos@gmail.com>

Hi again,
obviously, I have omitted the timestamps to make it easier the representation, not in the dev mailing list proposing a completely new naming scheme to alleviate some of the confusion. In this discussion I kept thinking: "maybe if there were some decent examples out there people would be more confused by the naming." So, this is my stab at explaining Cassandra's data model; It's intended to be a quick introduction to the basics of the data model, not a detailed explanation of every single detail but, hopefully, it helps clarify a few things.

BTW: this is long. If you'd rather have a PDF version of this presentation, you can download it from [here](#).

Exemple avec un panier d'achat

johndoe	17:21	Iphone	17:32	DVD Player	17:44	MacBook
willsmith	6:10	Camera	8:29	Ipad		
pitdavis	14:45	PlayStation	15:01	Asus EEE	15:03	Iphone

APIs disponibles

- Cassandra est accédé par Thrift, un RPC développé par Facebook
↳ *Thrift est disponible pour les principaux langages*
- Hector est un client de haut niveau pour Java
↳ *Offre un mapping de type JPA*
- PhpCassa est un client pour PHP, PyCassa est un client pour python

Exemple d'écriture avec Cassandra

```
Cluster cluster =
    HFactory.getOrCreateCluster("cluster", new CassandraHostConfigurator("server1:9160"));

Keyspace keyspace =
    HFactory.createKeyspace("EcommerceKeyspace", cluster,
                           new QuorumAllConsistencyLevelPolicy());

Mutator<String> mutator = HFactory.createMutator(keyspace, stringSerializer);

mutator.insert("johndoe", "ShoppingCartColumnFamily",
               HFactory.createStringColumn("14:21", "Iphone"));
```

*Insère une colonne dans une
ShoppingCartColumnFamily*

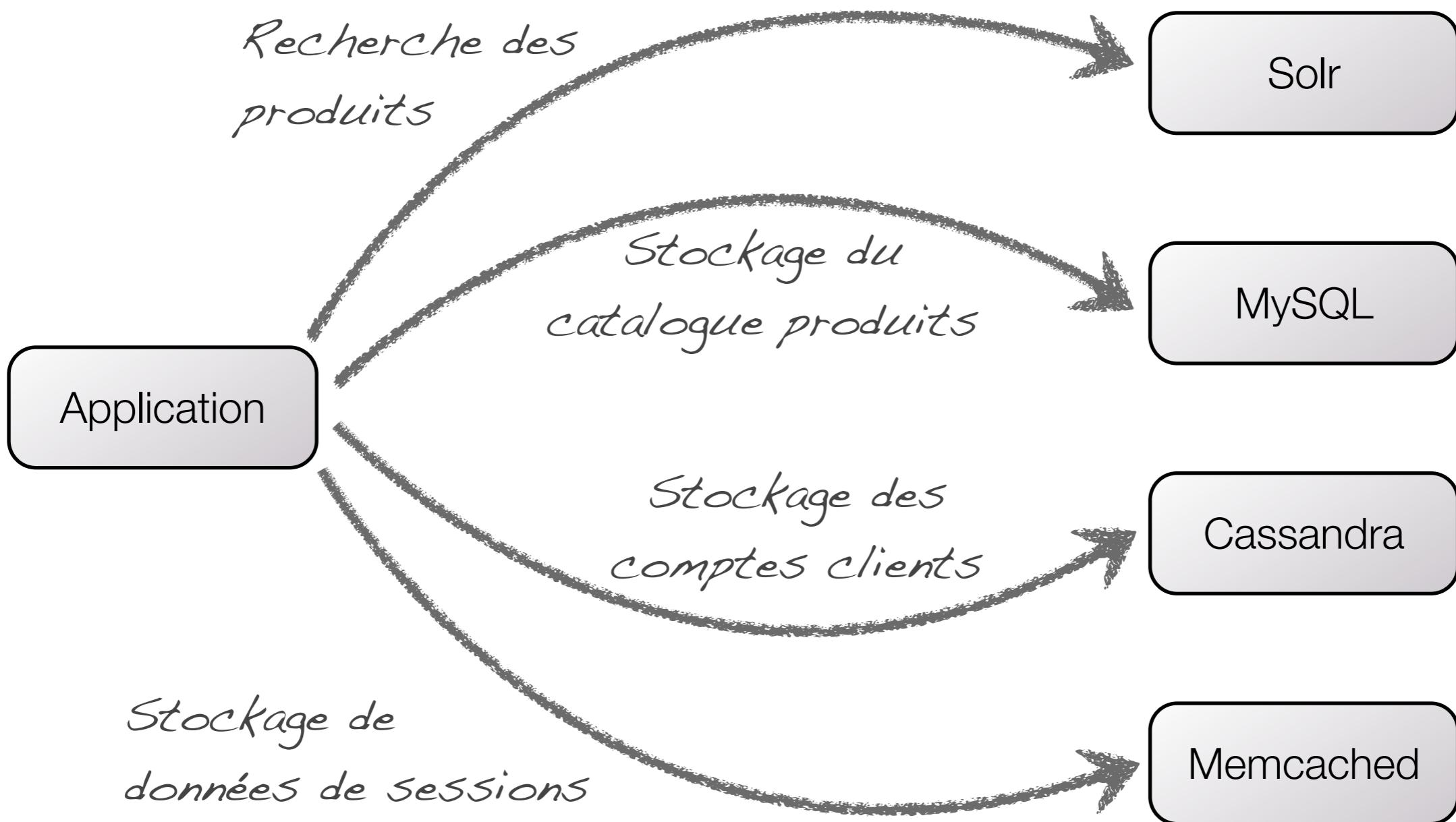
Exemple de lecture avec Cassandra

```
SliceQuery<String, String, String> query =  
    HFactory.createSliceQuery(keyspace,  
                               stringSerializer, stringSerializer, stringSerializer);  
  
query.setColumnFamily("ShoppingCartColumnFamily")  
    .setKey("johndoe")  
    .setRange("", "", false, 10);  
  
QueryResult<ColumnSlice<String, String>> result = query.execute();
```

*Lit un intervalle de 10 colonnes dans
une ShoppingCartColumnFamily*

Un cas d'usage

Un site de e-commerce

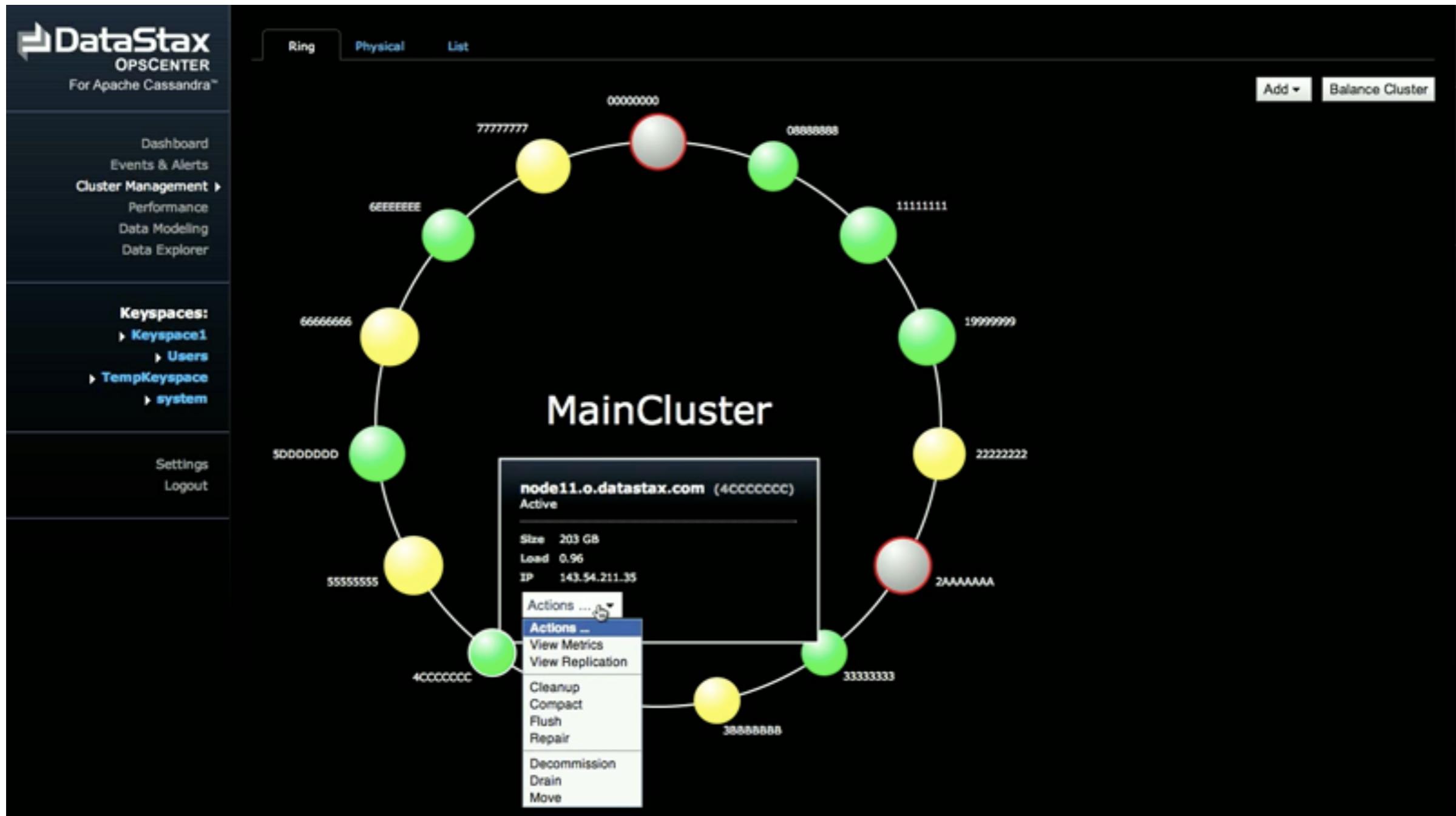


Cassandra en production

Cassandra en production

- En production chez de nombreux « Grands du Web »
- Outilage encore réduit
- Monitoring par JMX
- Les backups peuvent être problématiques avec des volumes importants
- La gestion du cluster requiert une équipe d'exploitation expérimentée

DataStax OpsCenter



L'intérêt pour l'entreprise

- Stockage polyglotte : une meilleure adéquation entre la BDD et les données
 - Scalabilité linéaire : être à même de répondre aux besoins les plus gourmands
 - Haute disponibilité : du multi-serveurs au multi-datacenters
 - Elasticité : une intégration naturelle à la logique du Cloud Computing
 - Curseur pour s'adapter : + de consistance ou + de fiabilité (Quorums)
- Et finalement... la possibilité crée le besoin !*

Questions / Réponses

?



@mfiguiere



blog.xebia.fr