Power Week

# Université IBM i 2019

## 22 mai

IBM Client Center Paris

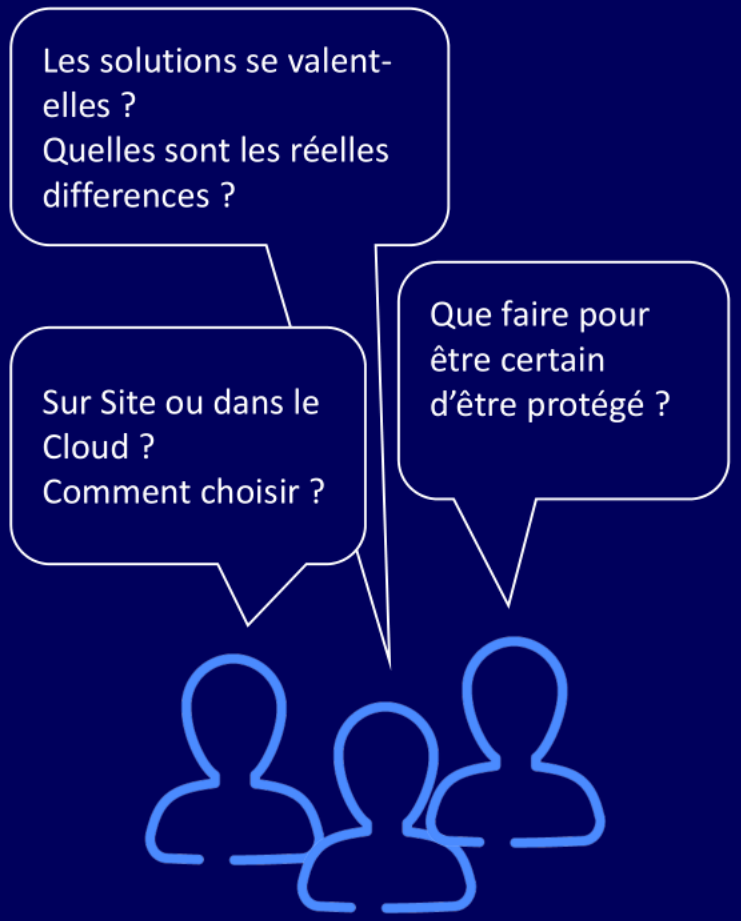
## S07 -Technologies de réplication sur IBM i

~~Benoît MASSIET du BIEST~~
~~ACMI~~
~~bmassiet@acmi.fr~~

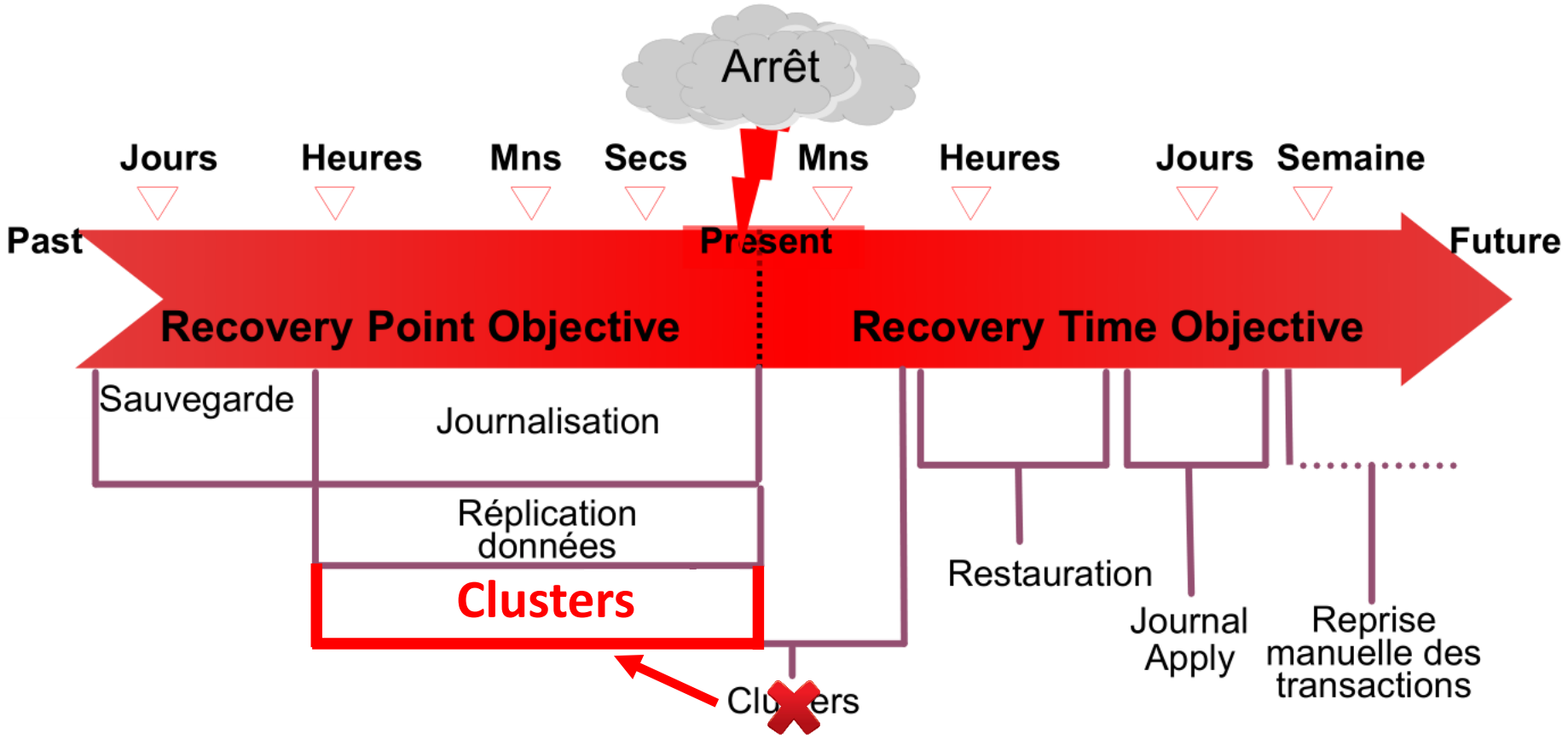
**Révision par Power IBM i**

# Agenda

- Le concept de résilience
- Les différences techniques entre réplication matérielle et réplication fondée sur la journalisation
- Les différences pratiques
- Considération concernant les topologies possibles (Cloud ?)
- Questions à se poser ?



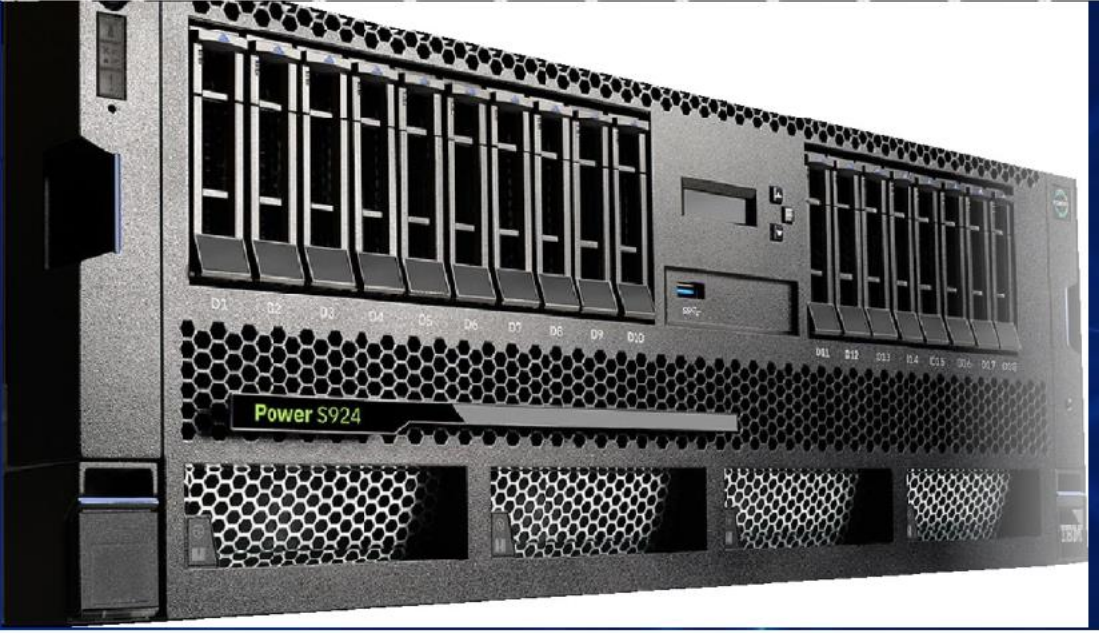
# Objectifs de reprise (RTO and RPO)



ACMi

# Objectifs de la Haute Disponibilité ?

- Couvrir les risques identifiables :
  - Panne matérielle
  - Panne logicielle
  - Erreurs Humaines, malveillances, grèves
  - Alimentation électrique, Télécom, Réseau,
  - Risques naturels, ou risques sur le bâtiment / Datacenter
- Garantir un temps de reprise (RTO)
- Garantir un point de reprise (RPO) et l'intégrité des données
- Couvrir les arrêts planifiés et les opérations de maintenance
- Permettre d'accroitre le temps de service IT (disponibilité)

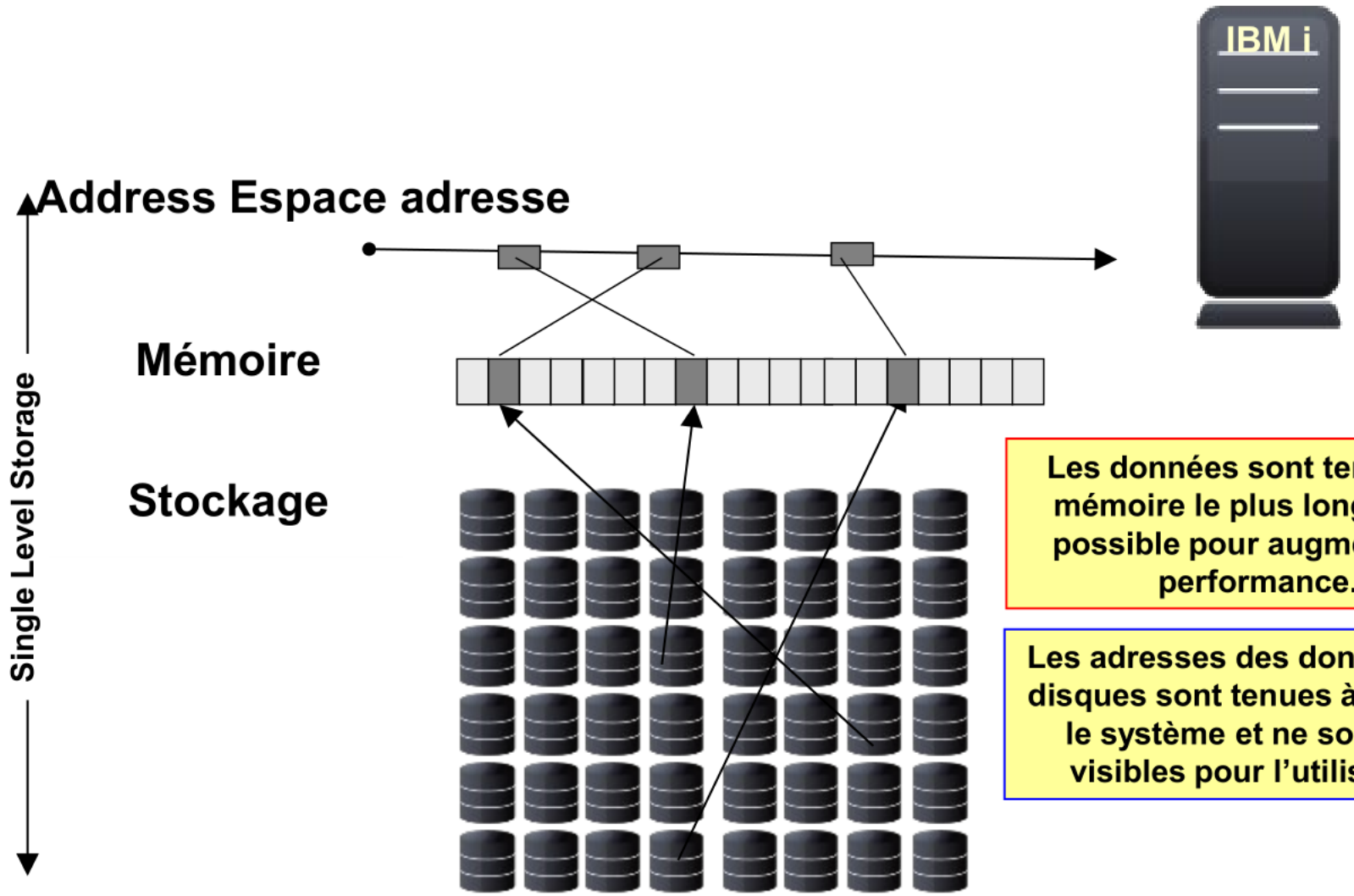




# Réplication matérielle ou réplication logicielle ?

## Le Power i : un serveur pas comme les autres.

# Single Level Store : l'espace adressable unique



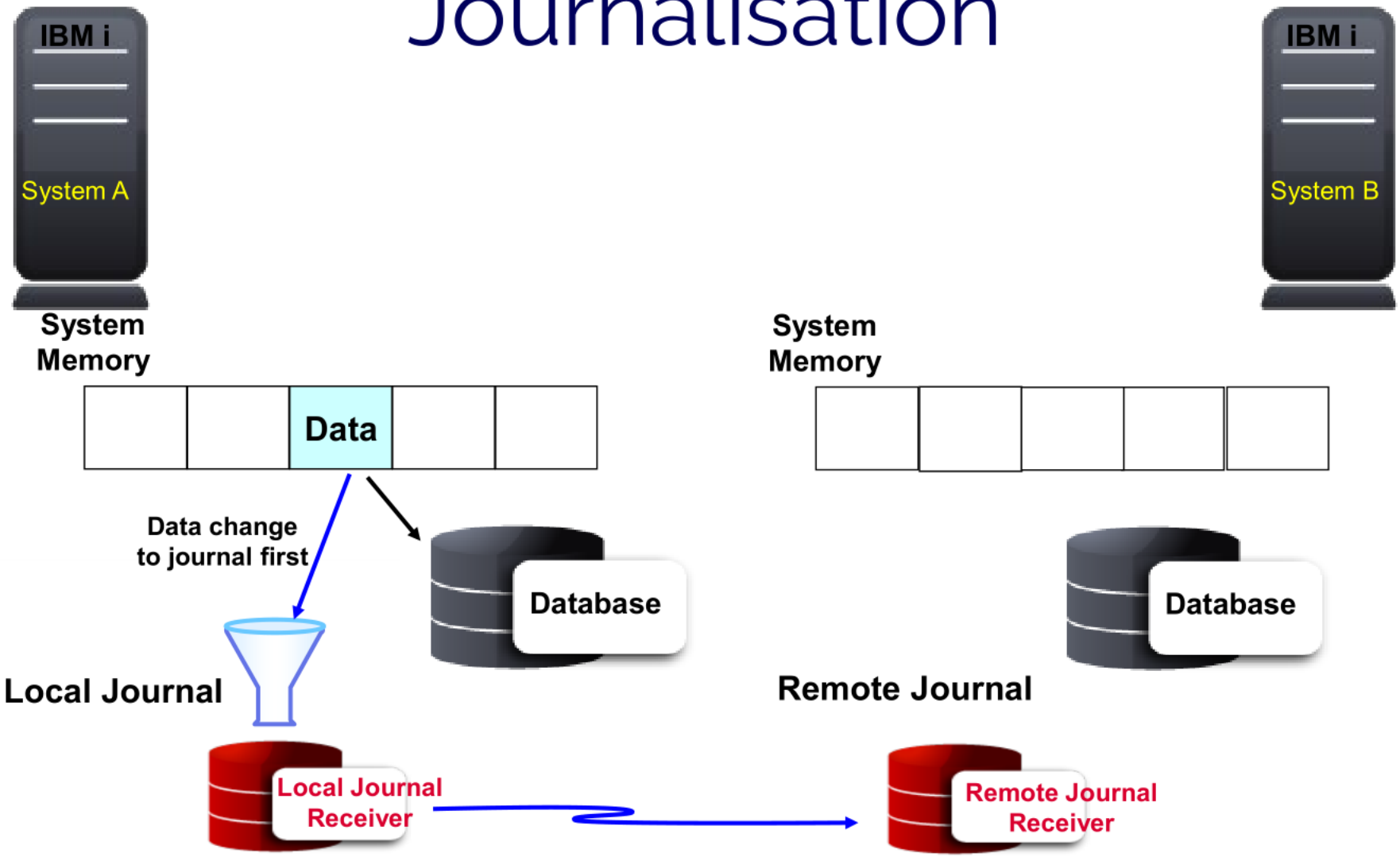

Les données sont tenues en mémoire le plus longtemps possible pour augmenter la performance.

Les adresses des données sur disques sont tenues à jour par le système et ne sont pas visibles pour l'utilisateur

# System i : Performance et écritures disques


**Les données sont tenues en mémoire le plus longtemps possible pour augmenter la performance**

# Journalisation



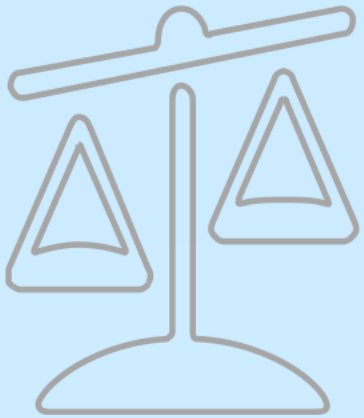
# Fonctions de protection et résilience sur IBM i

ACMi

- Les protections disques : Mirroring,  Raid 5, 6, 10
- La journalisation *Both, le commitment control
- La journalisation distante (V4R2, 1999)
- smapp : System managed access-path protection
- Le journal minimum data, la compression OS
- Le journal caching (Option 42)
- La journalisation implicite
- Les iASP et les “clustering administratives domains”
  - Et son application avec l’option 41 : XSM Power HA
- La fonction Quiesce (V6R1)
- Les API cluster
- Et ... en 2019 : Db2 Mirror for i (5770-DBM)

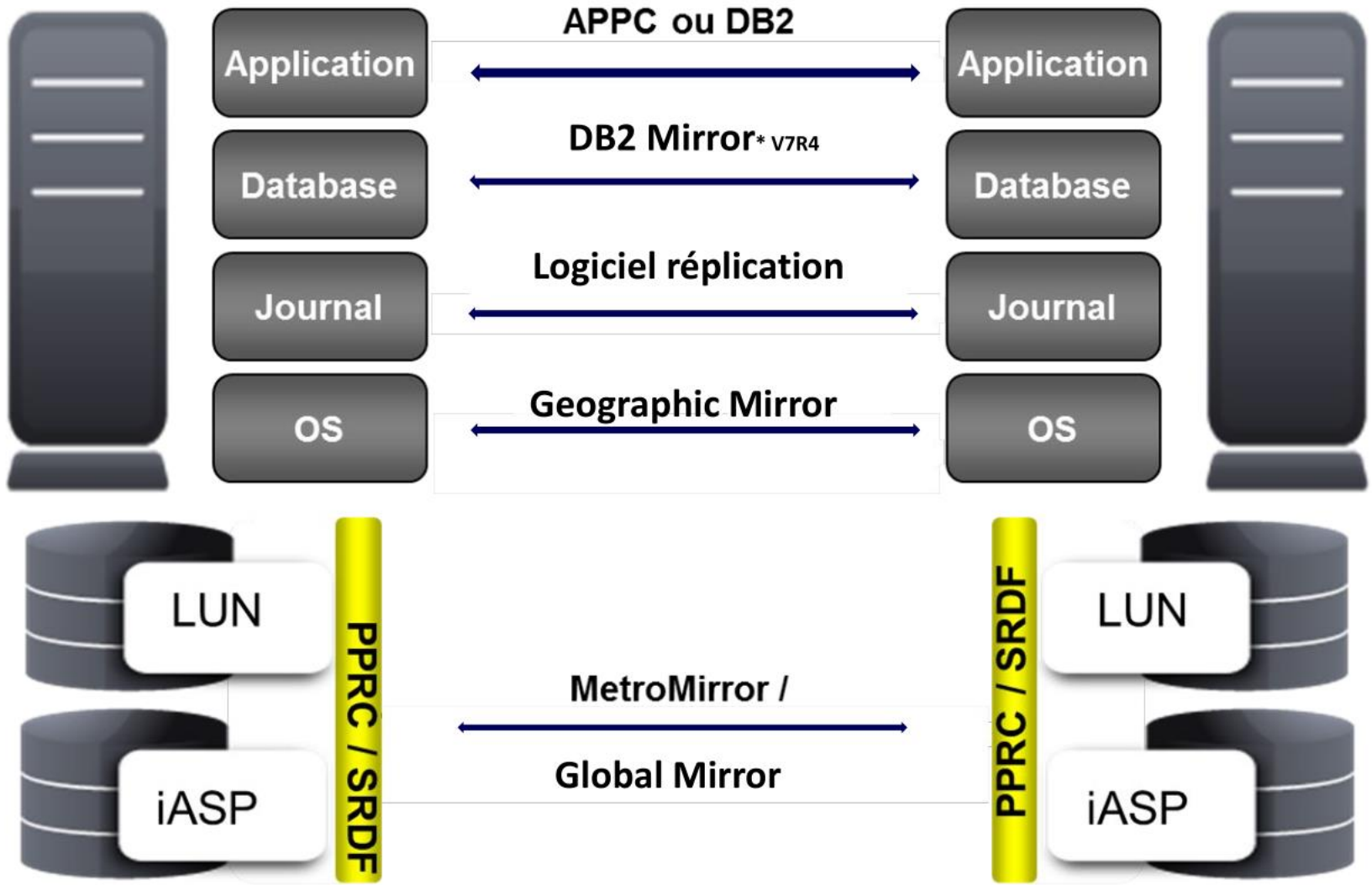
# Considérations Technologiques

## Comment peser les options possibles



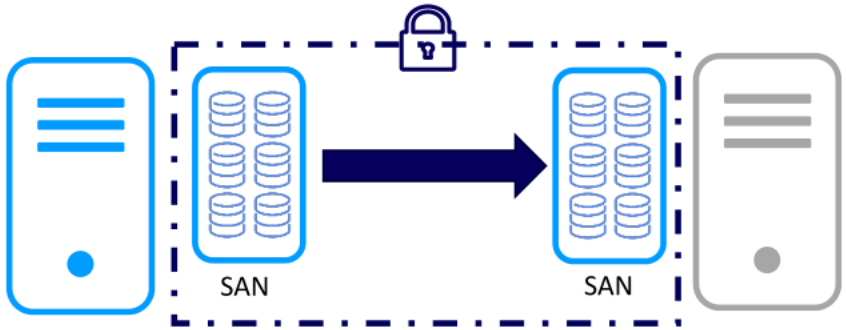
1. Les architectures
2. Réplication temps réel et point de reprise
3. Consommation de bande passante
4. Secours actif ou Off-line
5. Exigences des applications et des données
6. Intégrité des objets / objets corrompus ou endommagés
7. Flexibilité des topologie possible

# Niveaux de réplication



# Architecture sous jacente

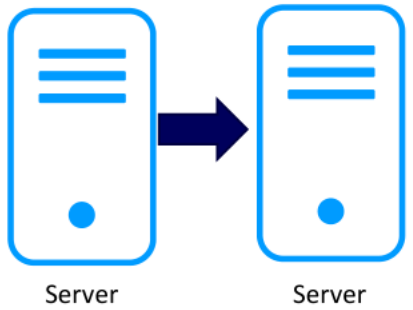
## Réplication du stockage




- Mécanismes de mirroring integrés de SAN à SAN.
- Réplication par secteurs de LUN ou blocs
- iASP
- Les données cible ne sont pas accessibles
- Basculement = Vary on + Récupération des données par ce serveur de secours ~~et IPL selon les cas~~

**Pas d'IPL requis**

## Réplication logicielle

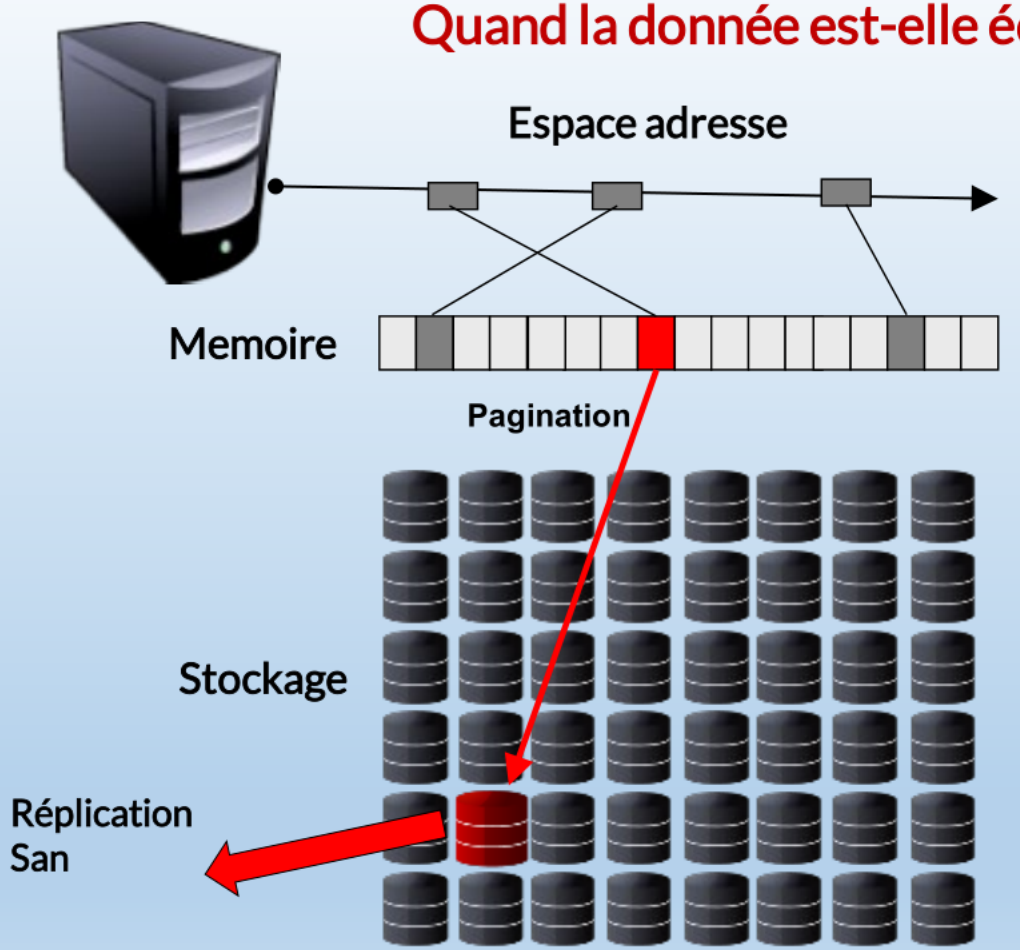



- Réplication par envoi en temps réel des postes de journaux (fichiers) et des entrées dans l'audit journal (objets).
- Les deux serveurs sont en miroir
- Indépendants actifs, les données sont accessibles.
- Supporte toutes les architectures matérielles
- Le basculement consiste à activer le réseau et les utilisateurs sur le serveur cible, pas d'IPL

# Dilemne : Intégrité des transactions

ACMi

## Quand la donnée est-elle écrite sur disque?




1. Le storage manager libère la mémoire selon un algorithme interne
2. La donnée est maintenue en mémoire pour augmenter les performances
3. La journalisation copie les données modifies sur disque
4. Fermer un fichier force son écriture sur disque
5. Le parameter de force write ration peut forcer l'écriture sur disque

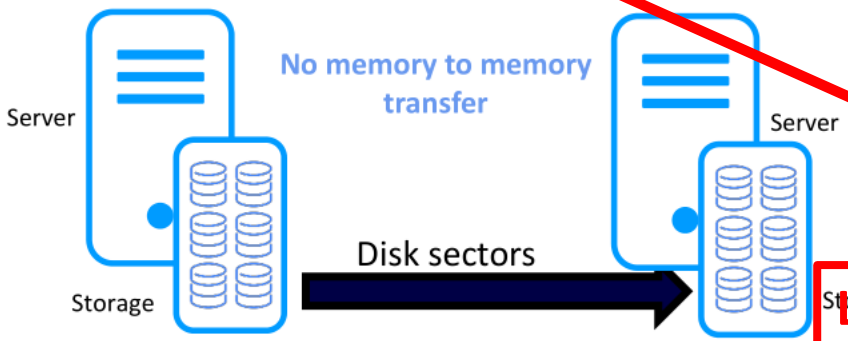
**La protection commence lorsque la page est écrite sur disque**

La journalisation est également active avec PowerHA, il s'agit d'une recommandation d'IBM

# Réplication et point de reprise

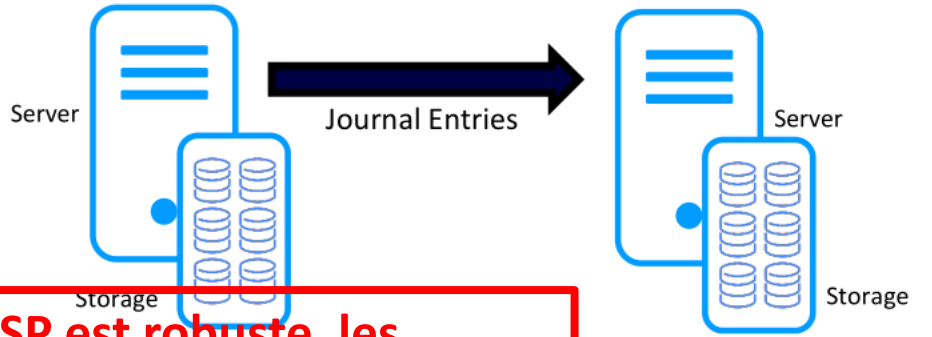
## Réplication du stockage

- L'OS utilise la mémoire comme une partie de l'espace de données
- Les données sont soit en mémoire soit sur disque
- Seules les données/objets sur disque sont protégées,
- Les données restant en mémoire sont perdues
- ~~Le RPO est imprévisible~~ **Le RPO est prévisible**
- La journalisation est requise pour récupérer les données perdues.
- Les mécanismes de récupération ~~rendent le RTO imprévisible~~



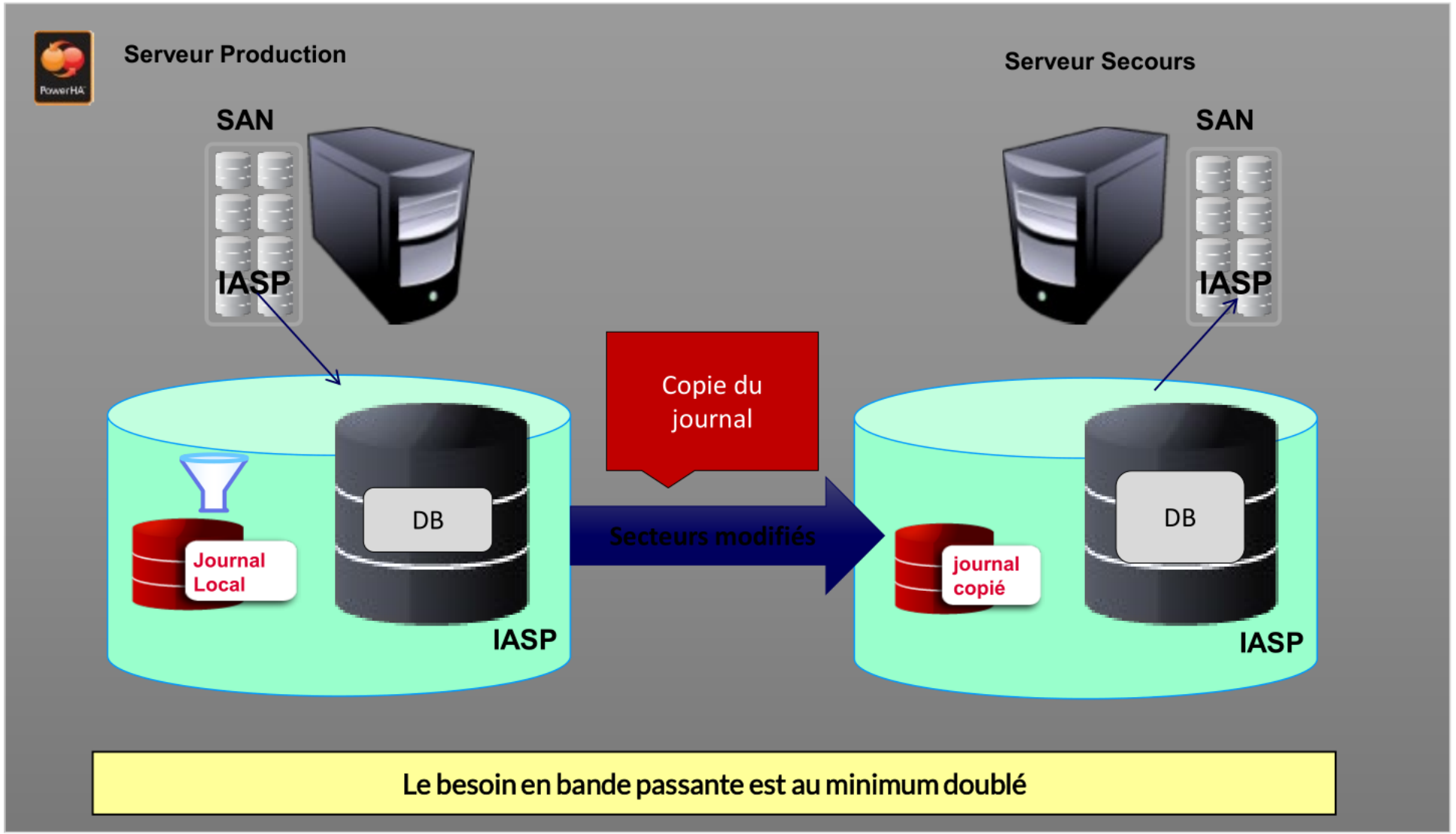

## Réplication logicielle

- La journalisation élimine les problèmes liés aux écritures disques ou mémoire
- Le poste de journal est répliqué en temps réel avant l'écriture dans la base de données source
- Le réplication synchrone ou asynchrone garantissent un RPO proche de zéro
- Pas de mécanisme de récupération, le RTO est rapide.



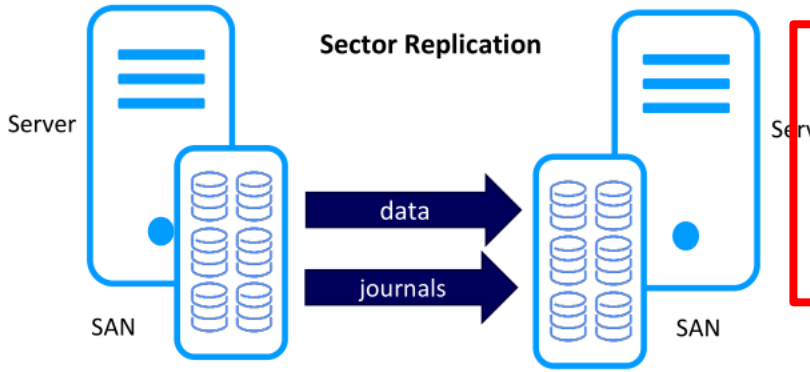

**Le démarrage des iASP est robuste, les différentes étapes sont relativement courtes**

# Besoin en bande passante



# Besoin en bande passante

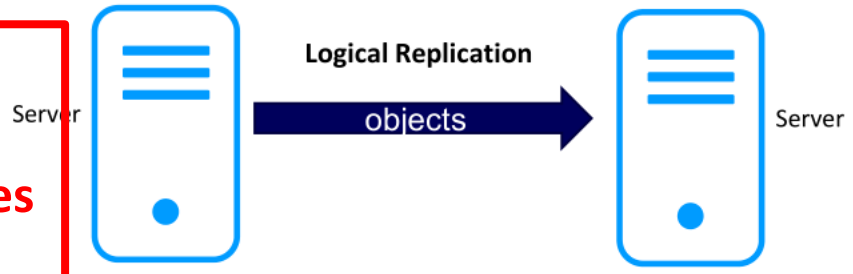
## Réplication du stockage




**Aucune copie de journal puisqu'il y a une copie binaire des secteurs disques**

- La bande passante nécessaire minimum x2
- Réplication par blocs de Luns + ~~copie des entrées des journaux eventuels~~
- La synchronisation initiale ~~et les processus de resynchronisation~~ réclament également beaucoup de bande passante

## Réplication logicielle



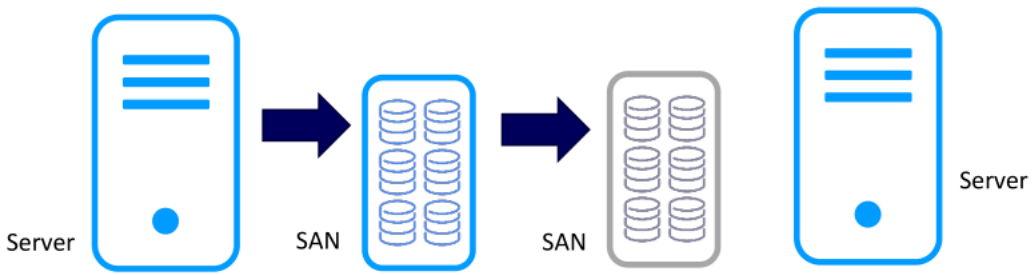

- La journalisation réplique uniquement les changement sur les données ou les objets
- La bande passante peut être encore réduite par filtrage des données non nécessaires, ou par le JRNMINDTA (Option 42 – Performance Journal HA) ou une compression OS (LZ 9, 10 ou 12)

**Les processus de resynchronisation ne resynchronisent que les secteurs modifiés.**

# Serveur cible Online ou Offline

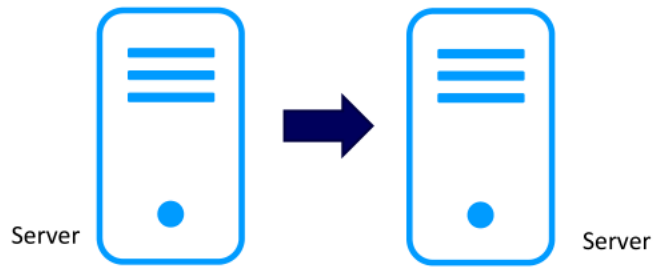
## Réplication du stockage

- Pas d'accès aux données de l'IASP distant
- Basculement = vary on de l'IASP distant
- Flashcopy possible mais :
- Seules les données écrites sur disques sauf si Quiesce
- Vary on nécessaire sur l'espace flashcopié




## Réplication logicielle

- Le serveur cible est toujours à jour et actif
- Le serveur cible utile pour :
  - ✓ Reporting,
  - ✓ DW, BI,
  - ✓ Sauvegardes en temps masqué
  - ✓ Base de données de test/ pre-prod.



# Contraintes sur les applications et les données

**Aucune conversion des données n'est requise.**

ACMi

## Réplication du stockage

- Les applications doivent s'exécuter dans un IASP pour être protégées
- ~~La conversion des données~~ et des applications peut s'avérer longue et fastidieuse
- **Toutes** ~~Seules~~ les données stockées dans l'IASP sont répliquées
- ~~Le Sysbas doit être protégé par une autre méthode~~
- ~~Nota: le domaine administré ne protège pas entièrement le SYSBAS~~

## Réplication logicielle

- Aucun changement applicatif n'est requis
- Tous les objets et toutes les données sont protégées
- Compatible avec une organisation SYSBAS + IASP
- Compatible avec tous types de stockage
- Toutes les applications connues sont supportées (par exemple banques, commerce de détail, logistique, santé, etc..)

**Il n'y a pas de données dans le SYSBAS, et s'il en reste, elles sont soit sous contrôle de l'Admin Domain, soit gérées par une procédure simple à l'aide de save file.**

# Objets endommagés

Les données sont journalisées, il n'y a donc pas de corruption

## Réplication du stockage

- Objets endommagés propagés
- ~~L'absence de protection de l'espace adresse unique, peut entrainer un corruption des données~~
- ~~Les logiques ne sont pas mise à jour en temps réel~~
- ~~En cas de basculement non planifié, les outils de récupération impactent fortement le RPO et le RTO~~

Le processus de clustering et de démarrage de l'iASP garantissent un accès rapide aux données.

## Réplication logicielle

- La réplication fondée sur le journal ne peut pas propager des objets endommagés de la source vers la cible
- Les éventuels objets endommagés sont détectés et réparés et font l'objet d'alertes
- les logiques sont mis à jour en permanence sur la cible

La réplication MetroMirror est une réplication binaire qui réplique TOUTES les mises à jour, y compris les logiques.

# Reconstruction des chemins d'accès (logiques)

ACMi

**En aucune façon, les chemins d'accès sont présents sur l'iASP, il n'y a qu'une fusion des métadonnées entre l'iASP et le système**

## Réplication du stockage

- ~~Un basculement non planifié entraine une reconstruction des chemins d'accès~~
- ~~Le temps de reconstruction dépend du stockage, du serveur, et de la taille des chemins d'accès~~
- ~~Ce temps n'est pas prévisible~~

**La durée est relativement courte et prévisible.**

## Réplication logicielle

- Les chemins d'accès (index) sont mis à jour en parallèle sur chaque serveur
- Pas de reconstruction nécessaire
- La réplication peut valider en temps réel leur intégrité et leur synchronisation
- Les chemins d'accès invalides ne peuvent être répliqués
- Le temps de basculement est réduit de 5 à 15 minutes

**Il n'y a pas de reconstruction des index, uniquement une consolidation des métadonnées, de plus cette étape est asynchrone.**

# Topologies possibles

NON, il peut s'agir de matériels totalement différents

## Réplication du stockage

- Avec une réplication SAN à SAN, les matériels source et cible doivent être ~~similaires,~~ compatibles
- ~~Limité à une réplication 1>1~~
- Les tailles de Lun identiques
- Les versions, releases microcodes, alignés
- La distance ~~peut être limitée~~ selon la nature de la réplication (synchrone ou asynchrone)
- ~~Pas compatible avec les solutions Cloud ou mutualisées~~

NON, pas avec la DS8000

Aucune contrainte à ce niveau

## Réplication logicielle

- La réplication supporte tout type de stockage, SAN ou disques internes, toute organisation
- Supporte toute distance (mode asynchrone)
- Permet des montées de versions OS décalées (2 releases d'écart supportées)
- Autorise les topologies 1>1, 1>N, N>1, A>B>C
- Parfaitement adaptée pour les solutions Cloud

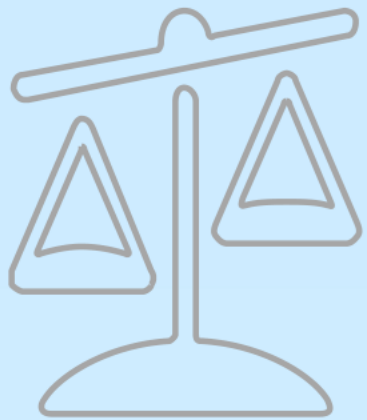
NON, le mode asynchrone Global Mirror étend la limite comme en HA logicielle

# Considérations pratiques

# Comportement de la solution

ACMi

## Bien choisir selon l'usage



1. Synchronisation initiale
2. Basculement non planifié (sinistre)
3. Basculement planifié (maintenance)
4. Contrôle d'intégrité
5. Reprise en cas de problème réseau
6. Maj du système d'exploitation
7. Mise à niveau stockage
8. Evénement sur stockage
9. Répartition de la charge de travail

# Synchronisation initiale

ACMi

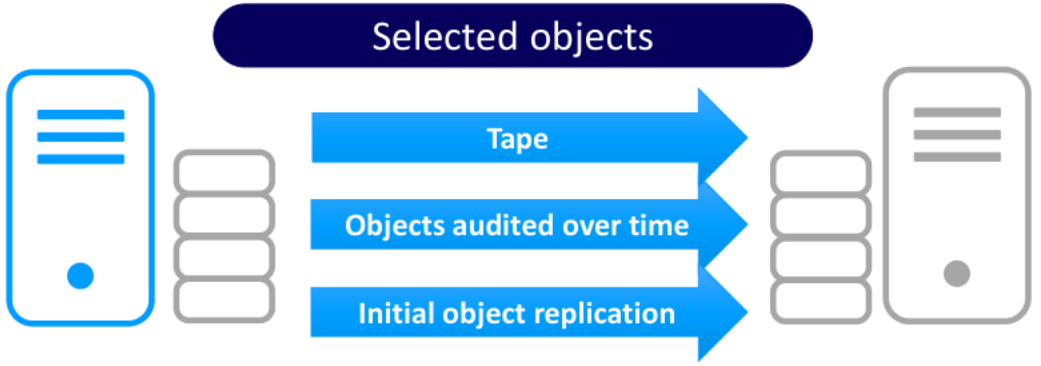
## Réplication du stockage

- Synchronisation initiale indispensable
- Tous les secteurs / blocs sont envoyés via le réseau **ou le SAN Fabric**
- ~~Risques sur les volumes~~ ( 10 To = 23 heures si 1Gb/s à 100%)

**La bande passante est taillée pour assurer la réplication**


## Réplication logicielle

- Options possibles :
- Save / restore
- Copie du stockage
- Puis resynchronisation via audits
- Réduit l'impact sur le réseau



# Basculement imprévu

ACMi

Cela n'a rien à voir avec un IPL, il s'agit simplement d'un device particulier

## Réplication du stockage

L'iASP

- ~~La baie~~ cible doit passer par un vary-on
- Puis un contrôle d'intégrité (~~semblable à un IPL anormal~~)
- 34 étapes de recovery dont la durée est imprévisible
- Certaines étapes peuvent conduire à :
  - ~~Objets endommagés~~
  - ~~Chemin d'accès invalide~~
  - ~~Nécessite une intervention manuelle~~

A considérer : ~~le SYSBAS est répliqué par une autre méthode (OS ou Logiciel)~~

Le SYSBAS n'a pas à être répliqué, on utilise l'Admin Domain

~~RPO incertain, RTO indéterminé~~

## Réplication logicielle

- Le système cible est actif et son espace adressable cohérent
- Pas besoin d'IPL
- Les données sont intègres, et leur synchronisation controlée
- Les chemins d'accès sont tenus à jour en temps réel
- Le basculement des différents flux est parallélisable
- L'environnement système est inclus
- Pas de contrôle nécessaire

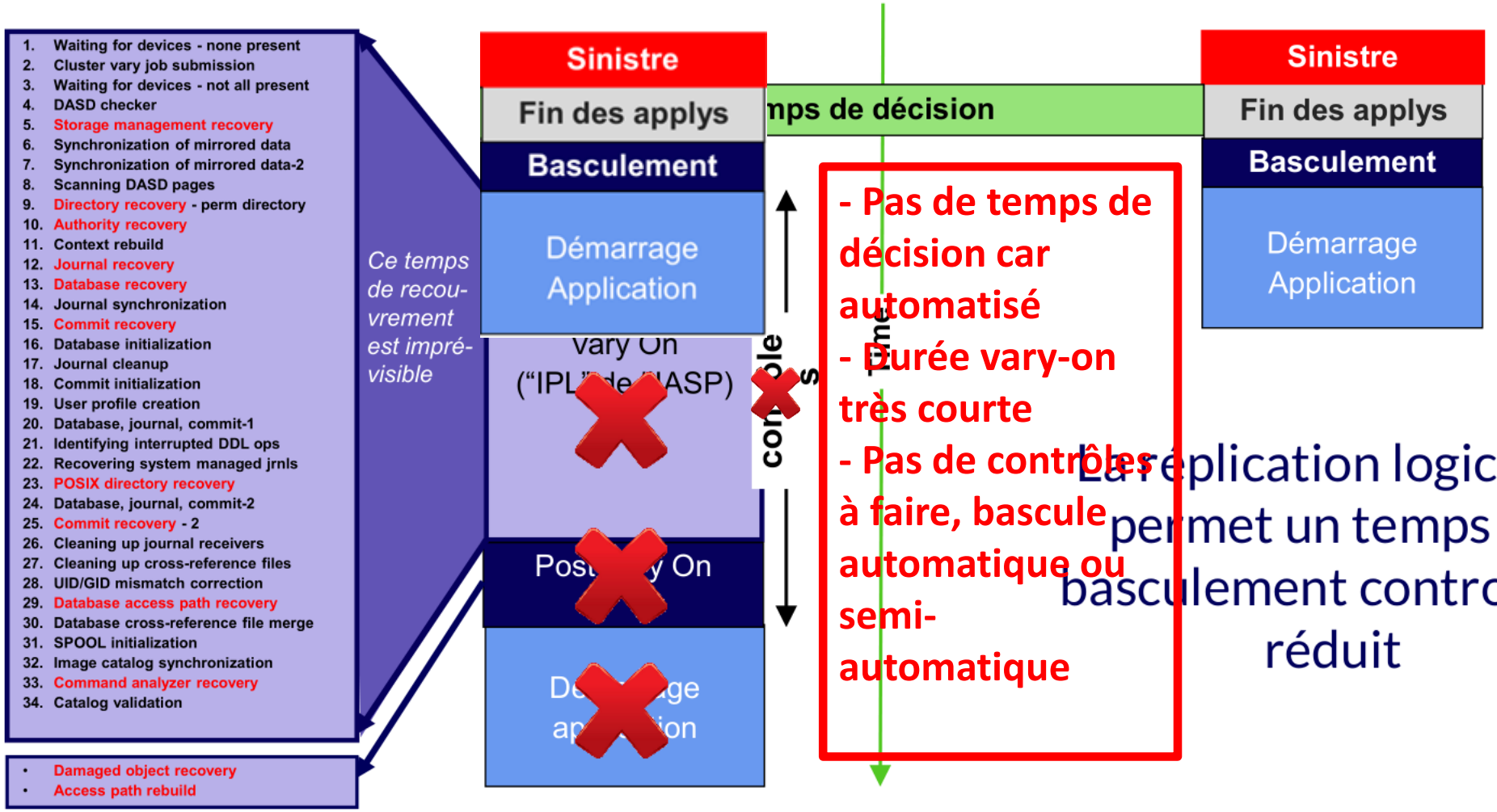
Cette opération est totalement automatique et sécurisée

RTO et RPO garantis

# Chronogramme d'un basculement imprévu

## Réplication du stockage

## Réplication logicielle




La réplication logicielle permet un temps de basculement controlé et réduit

# Basculement planifié

ACMi

## Réplication du stockage

- Quiesce et/ou Vary off sur ~~la baie de stockage~~ l'iASP coté production
- 34 étapes de recovery
- Et Vary on ~~de la baie~~ l'iASP coté secours

A considérer : ~~le SYSBAS est répliqué par une autre méthode (Os ou Logiciel)~~

**Le SYSBAS n'a pas à être répliqué, on utilise l'Admin Domain**

RPO proche de zéro, RTO imprévisible

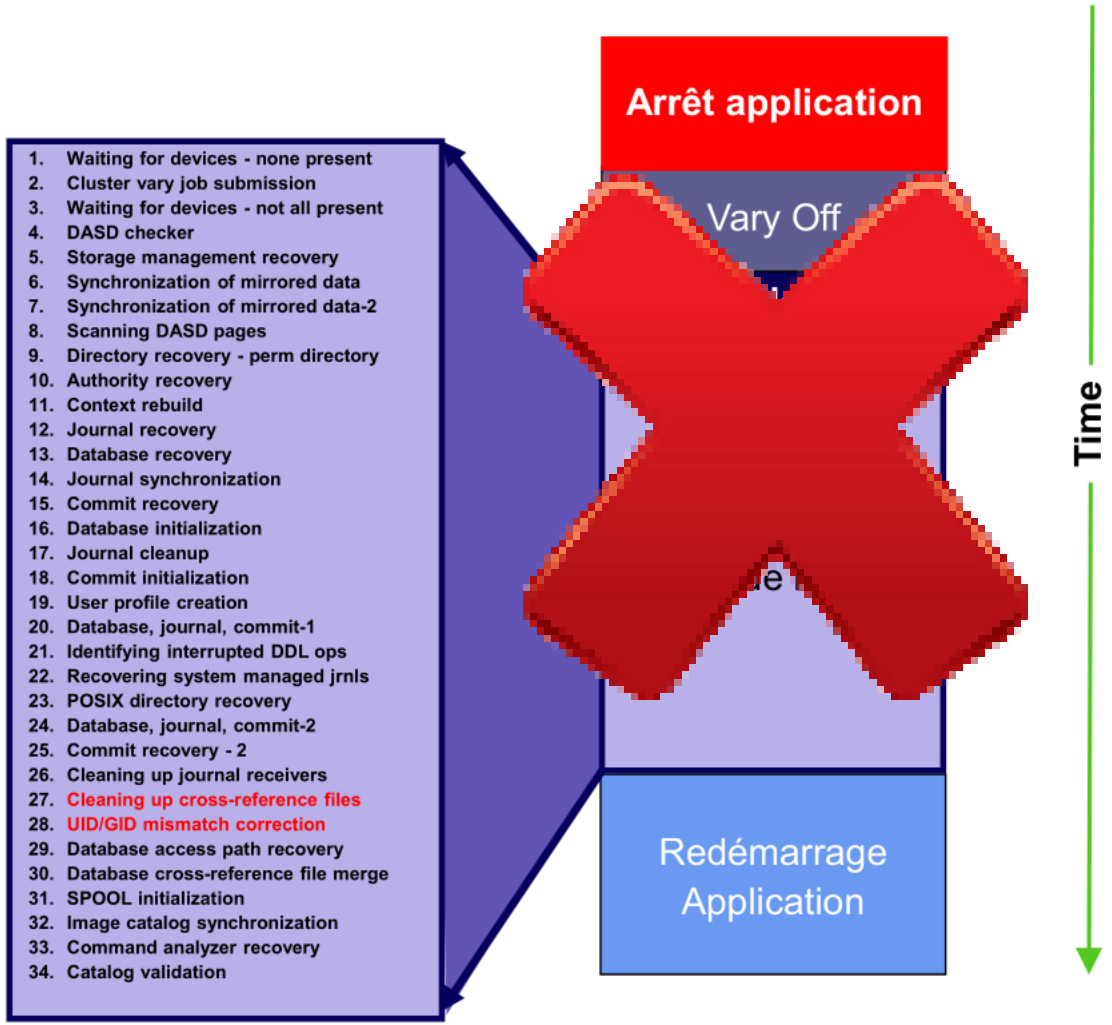
## Réplication logicielle

- Le serveur cible est actif
  - ✓ Pas de vary on
  - ✓ Pas de contrôle au moment du basculement
- Les scripts de basculement sont fournis
  - ✓ Les différents environnements peuvent basculer en parallèle
  - ✓ Le processus de basculement est controlable à chaque étape

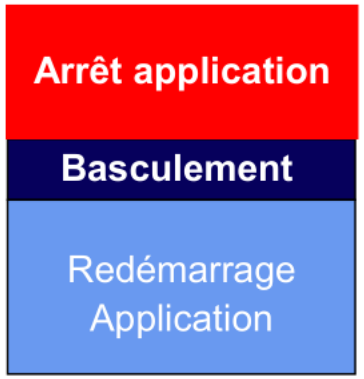
RPO proche de zéro, RTO prévisible

# Chronogramme : basculement planifié

## Réplication du stockage




## Réplication logicielle

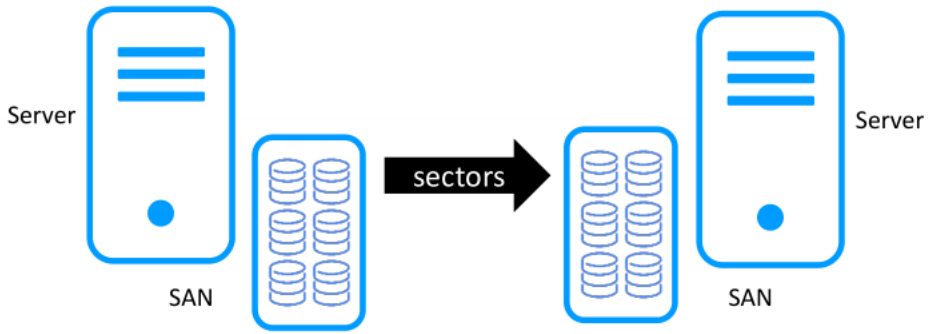



La réplication logicielle permet un temps de basculement controlé et réduit

# Reporting et requêtes

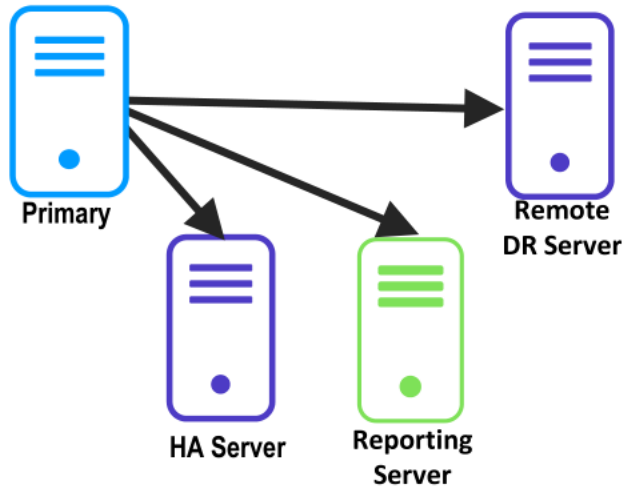
## Réplication du stockage

- Les données sur le SAN cible ne sont pas accessibles
- Il faut activer des “flashcopy” et remonter les données dans une autre partition pour y accéder




## Réplication logicielle

- Les rapports, queries , extraction et tous travaux read-only peuvent être effectués à tout moment
- Une réplication multiple (Y et plus) permet de répartir les données



# Validation de l'intégrité des données

**Intégrité garantie par le SAN**

## Réplication du stockage

- ~~Les données de la cible San ne peuvent pas être contrôlées~~
- En cas de problème une resynchronisation partielle ~~et totale~~ peut être nécessaire
- Les corruptions fichiers sont indétectables avant le basculement


## Réplication logicielle

- Différentes méthodes de contrôle d'intégrité peuvent être utilisées (jusqu'à 8)
- Des fonctions de basculements virtuels permettent de tester une reprise sans gêner les utilisateurs
- Des fonctions du "Remote Journaling" permettent de contrôler les problèmes réseau

# Reprise après une rupture réseau

**Comme toutes les autres solutions HA**

## Réplication du stockage

- Selon l'ampleur de la panne, une synchronisation partielle ou totale des secteurs stockées peut être nécessaire
- SAN n'est pas protégé qu'à l'issue de la resynchronisation

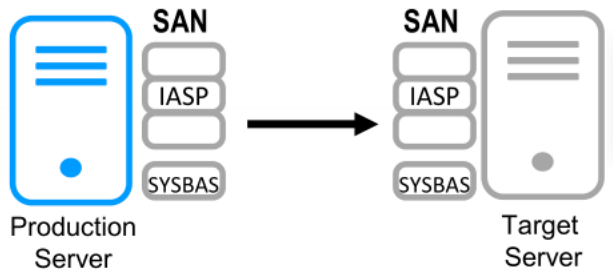

## Réplication logicielle

- En cas d''arrêt temporaire du réseau, la réplication rattrape le retard avec les écritures lues dans le journal
- Le remote journal package les écritures pour les transférer plus rapidement
- La protection est assurée dès la fin de ce transfert

# MAJ de l'operating System

## Réplication du stockage

- Une fois que vous basculez vers le serveur cible, vous ne pouvez revenir en arrière qu'après avoir mise à jour également le serveur de production
- Pas de retour arrière possible hormis un save / restore



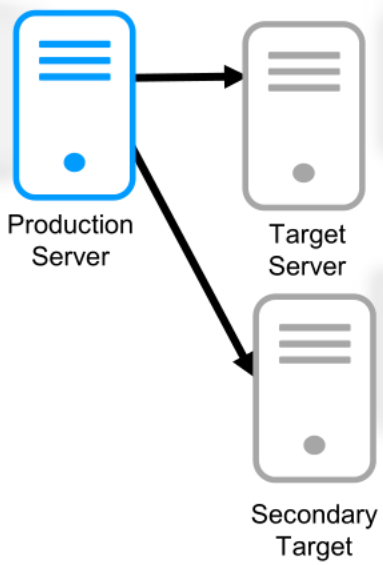

3 *Upgrade original production server to the same level as the target. Switch back.*

1 *Upgrade the target server to the new OS version and validate.*

2 *Switch to the target server. The IASP image will be upgraded as part of Vary On. This is non-reversible and must complete successfully.*

## Réplication logicielle

- La réplication logicielle supporte des écarts en release et version de l'OS
- Upgrade , Test et Bascule permettent de réduire le risque et le temps d'arrêt
- Un 3ème serveur peut assurer tout risque pendant la migration




2 *Upgrade the production server to the new OS version. Switch back to the production server.*

1 *Upgrade the target server to the new OS version and validate. Switch to the target server.*

3 *Replication to a secondary target is recommended to protect against outages during the upgrade.*

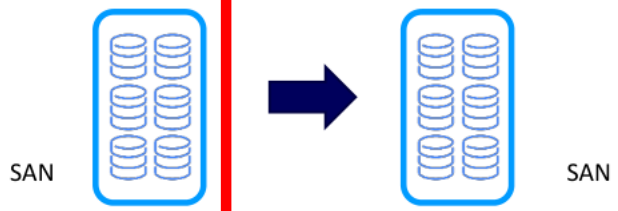
# Evénement sur le stockage

**C'est quand même le but pour synchroniser des données !**

## Réplication du stockage

- Les SAN source et cible sont étroitement couplés
- Un arrêt du SAN source ~~ou cible~~ peut nécessiter un arrêt de production
- ~~Par exemple un Reclaim Storage (RCLSTG) provoque un arrêt de la protection~~

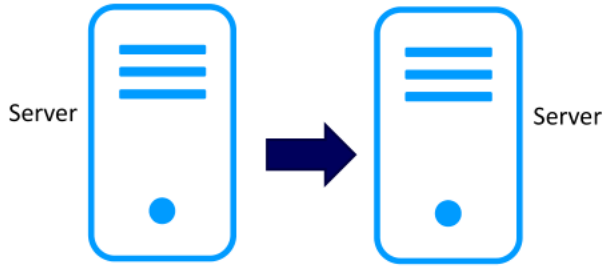
**En aucun cas !**




**La protection étant assurée par le SAN (hardware), le Reclaim Storage n'a aucune incidence sur la réplication**

## Réplication logicielle

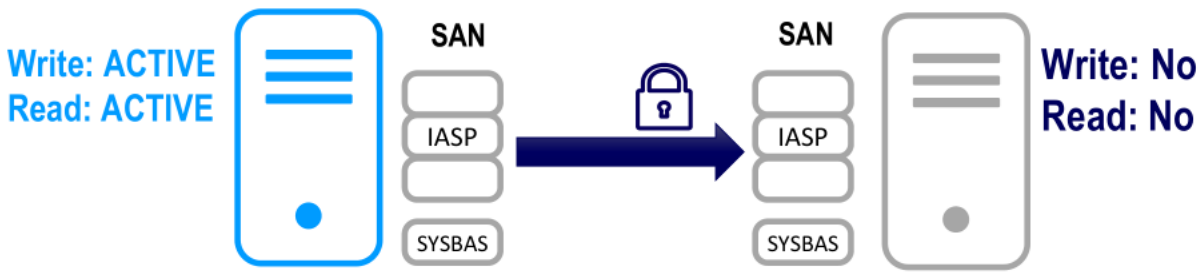
- Indépendante des types de stockages
- Les événements sur un disque ou une baie peuvent nécessiter un basculement avec un très faible impact sur la production
- Par ex : Reclaim Storage (RCLSTG ) nécessite un basculement et une suspension de la protection



# Répartition de charge

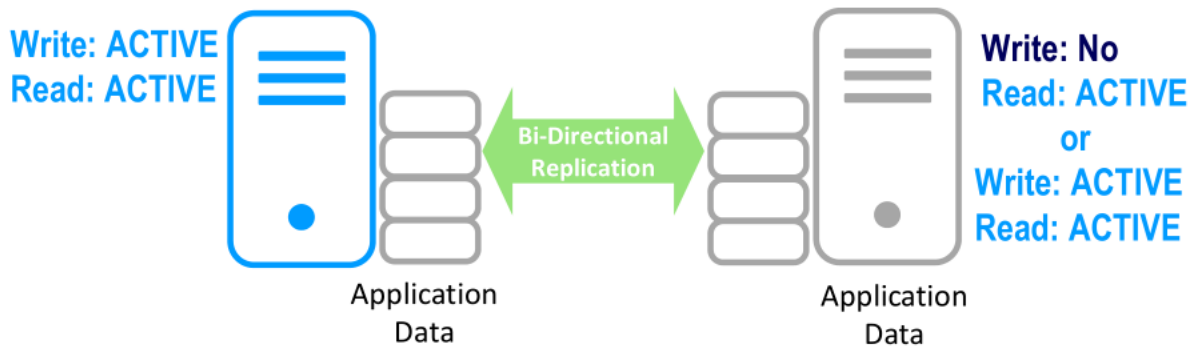
## Réplication du stockage

- La cible SAN est verrouillée pendant la réplication, il n'y a pas de répartition de charge possible
- Pour utiliser les données cible , il faut utiliser les technologies de flashcopy / snapshot




## Réplication logicielle

- Les topologies de réplication possibles sont 1-vers-1, 1 vers N, A->B->C, ou bidirectionnel)
- Les données sur le serveur cible sont mises à jour en temps réel, accessibles et protégées contre l'écriture
- Une réplication bi directionnelle par clé permet de bâtir des solutions active/ active



# Résumé : SLA et contraintes

Quelques minutes

Oui

Plus avec DS8000

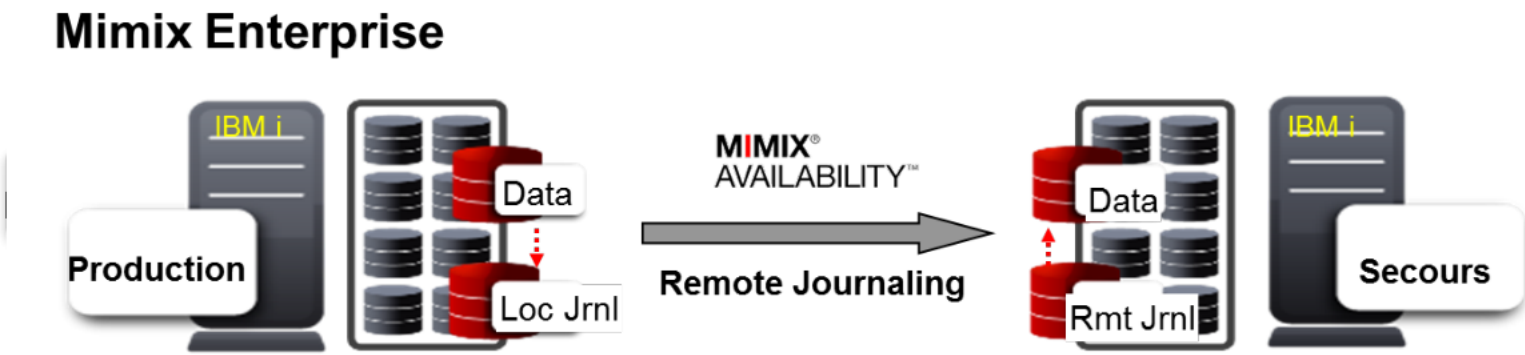
| Solutions | RTO | RPO | Distance | Besoin Débit réseau | Cloud ready | Architecture nécessaire | Nombre de nœuds |
|---|---|---|---|---|---|---|---|
| **Réplication HW** | 2 à 4h | ~ 0 , Incertain sur SVC | Limitée si synchrone | Fort 10 Mb/s à 100 Mb/s | Non | Espace disque et serveurs doublés à l'équivalent | 1 vers 1 Ou cascade |
| **Db2 Mirror** | ~0 | ~0 | <1km | Très fort 40 Gb/s à 100 Gb/s | Non | Espace disque et serveurs doublés à l'équivalent | 1 vers 1 |
| **Réplication SW** | ≤1h | ~ 0 | Illimitée si asynchrone | Faible à moyen de 1 Mb/s à 50 Mb/s | Oui | Espace disque et serveurs doublés peuvent être mutualisés et/ou dégradés | 1 vers N<br>N vers 1 |

# Résumé : Les usages

| Solutions | RTO | RPO | Accès aux data / cible | CDP | Sélection périmètre | Basculement | Surveillance /Charge Gestion | Support /hotline |
|---|---|---|---|---|---|---|---|---|
| **Réplication SAN** | ~~Imprévisible ordre 2 à 4 h~~ | ~~incertain sur IBM i~~ | non | Non | Non | ~~Basculement SAN Basculement Sysbas~~ **Sous contrôle opérateur rapide** | Oui/Faible | ~~Selon solution~~ **Point Service IBM** |
| **DB2 Mirror** | ~0 | ~0 | Oui idem source | Non | Non | Pas de basculement | faible | IBM |
| **Réplication Logicielle** | **De 15mn à 30 mn** | ± 0 | Ok | Oui | Oui | Sous contrôle opérateur rapide | Oui/Normale | ACMI |

# Cas concret

Client IBM i : deux serveurs à 200 km avec 48 partitions, dont 13 de production 48 processeurs actifs, 2 baies DS 8800 , env 200 To de capacité disque
Installation Power HA Global Mirror avec groupe de cohérence.
La réplication du Sysbas par le logiciel Quick-EDH

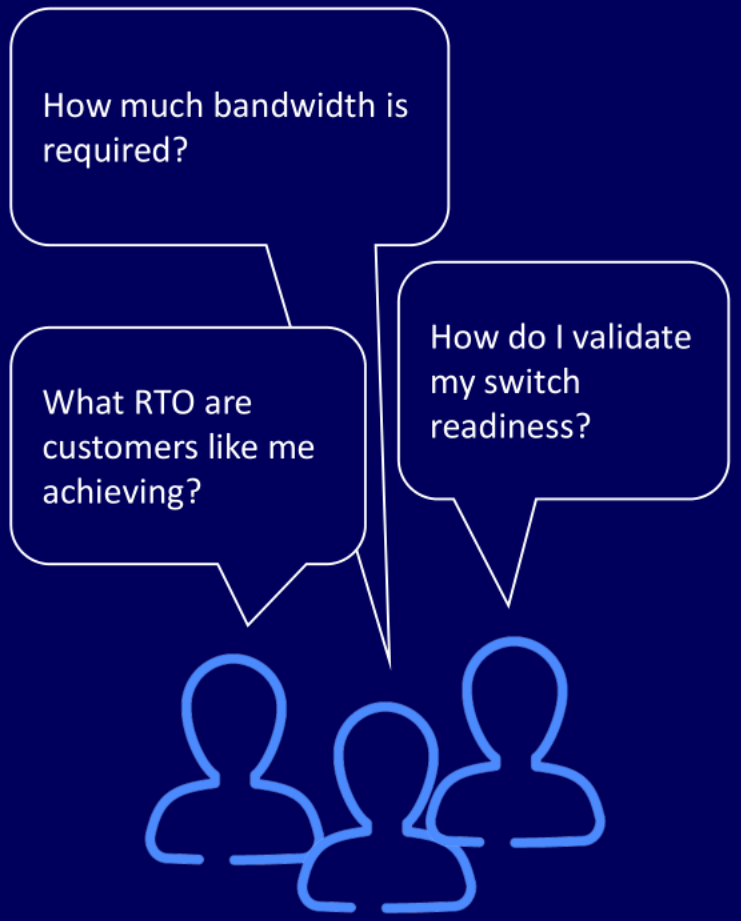



- Installé POWER HA GM en 2013
  - Un seul test réalisé sur une partition : basculement 40 mn
  - mais 2h à 7h de recovery dont la reconstruction des chemins d'accès
  - Pas d'accès aux données coté cible : d'où un ensemble de partition pour Flashcopy

- POC Mimix en Juillet 2018
  - après installation & basculement : 3 jours
  - Résultat : RPO <1 mn, RTO <30 mn pour lancement des tous les applications
  - Ni phase de recovery, ~~ni phase de reconstruction des logiques.~~

# Questions à se poser

- Quels sont mes objectifs de RTO / RPO ?
- Quelle distance et quel réseau entre les serveurs ?
- Quels avantages en terme de service (sauvegarde/ répartition de charge, maintenance facilitée) ?
- Comment puis je garantir l'intégrité des données ?
- Quelle flexibilité sur les architectures (local, distant , plusieurs serveurs)
- Puis je rencontrer des clients installés pour leur expérience de basculements ?
- Comment s'effectue la synchronisation initiale et quelle bande passante est nécessaire ?
- Quelle bande passante est nécessaire pour la réplication de l'état d'équilibre ?
- Quel RTO/RPO serai-je en mesure de réaliser ?
- Puis-je tester sans impacter la production ?
- Ce que je peux avoir des nœuds pour HA, DR et rapports ?
- De quels usages quotidiens pourrais je bénéficier ?
- Quel est le coût total de tous les composants matériels et logiciels requis ?



# QUELQUES RÉFÉRENCES ACMI avec MiMiX

- Traitement des flux bourse Euronext
- **Hautement Critique :**24/7 RTO 15 min , constaté **RTO ~3 mn**
- 7x2 partitions IBM i en Cluster sur 2x IBM Power P7+ 20 cpu
- Flash System V9000 100 To

- Industrie et distribution
- 2 partitions IBM Power 7+, 25 cpu, FS900 att.Direct 500 To
- **RTO 15 min**

- Distribution Matériel Electrique
- 2 serveurs IBM Power 880 36 Cpu – 6 Partitions
- 2x DS 8870 100 To
- **RTO < 1 h**

# QUELQUES RÉFÉRENCES SYNCSORT

**En Europe**

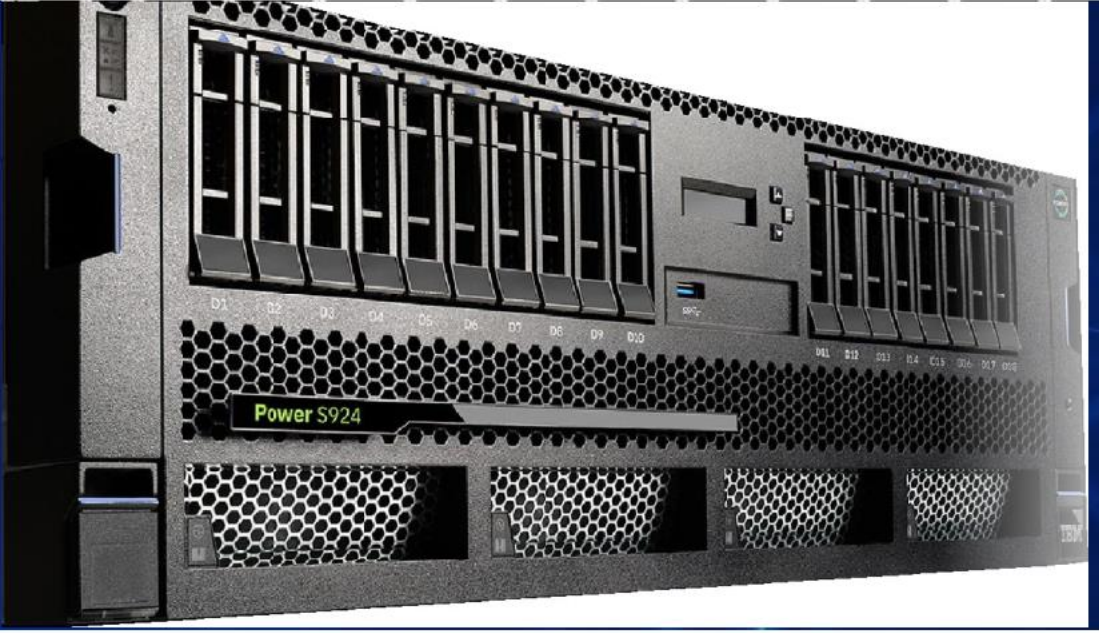
- **KNAUF Allemagne** 2 x880 50 CPU with SAP, disk IBM FS900 with 250TB, **RTO <8 mn** On SAP BW with 2.8TB of JRNRCV per hour 100 Millions tr/hour
- GLS Europe
- Hermes Germany GmbH
- Nike

**En Russie**

- Alfa Bank **2x 880 MHE 192 cpu, 2 Tb/day**

**En Asie**

- (Indonesie) :Bank Rakyat, Bank Mandiri, (Philippines) : Banco de Oro, (Chine): China Merchants Bank, CITIC Bank, Bank of China Under Mimix and Mimix Global (3 sites )





# Merci pour votre attention !

ACMi

Benoît MASSIET du BIEST
bmassiet@acmi.fr