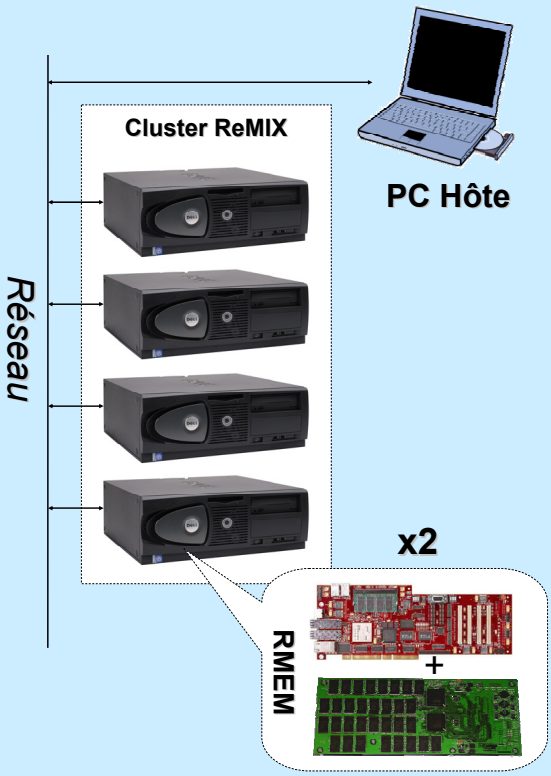
IRISA

VALORIA

ReMIX

Mémoire Reconfigurable pour l'Indexation de Masses de Données

Plate-forme matérielle

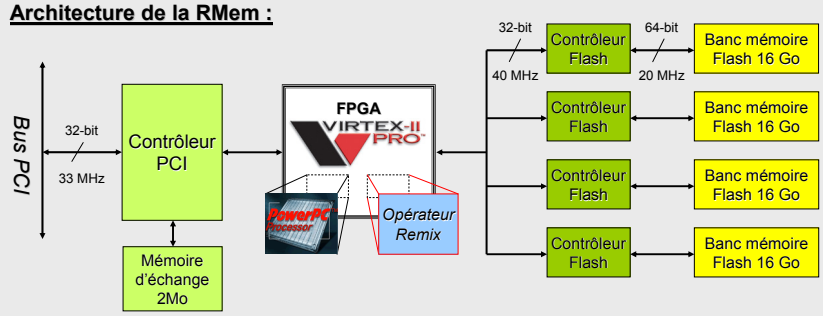


L'ACI ReMIX propose l'élaboration d'une mémoire spécialisée de très grande taille, dans le but d'accélérer la recherche d'informations dans des bases de données indexées. Une architecture matérielle dédiée a été développée. Trois champs disciplinaires représentatifs serviront de support pour valider cette proposition : la génomique, la recherche d'images par le contenu et la recherche documentaire basée sur les textes et leurs structures.

Caractéristiques :

✓ L'architecture ReMIX est un cluster de 4 PCs. Chaque nœud est doté de 2 cartes mémoires reconfigurables (RMem) de 64Go, portant la mémoire d'index globale à 512Go.
✓ Les index sont distribués sur 8 supports physique indépendants.
✓ La technologie mémoire « Flash Nand » utilisée offre une latence d'accès en lecture 200 à 500 fois inférieure à un disque dur.
✓ La bande passante agrégée en sortie de mémoire est d'environ 5Go/s.
✓ Des opérateurs matériels reconfigurables, implantés dans un FPGA, permettent un filtrage efficace des données en sortie de la mémoire.
✓ Un système de fichiers dédié permet la gestion des bancs mémoires Flash.
✓ Un environnement de programmation générique de haut niveau parallélise les traitements au niveau du cluster.

Architecture de la RMem :



Recherche par le contenu dans les bases de séquences génomiques

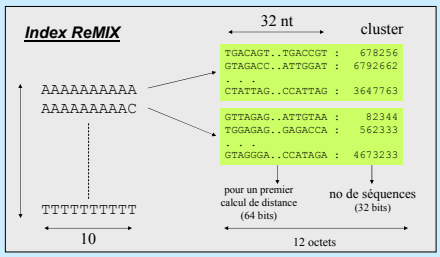
✓ **Objectif** : Indexation du programme BLAST sur ReMIX.

✓ **BLAST** (*Basic Local Alignment Search Tool*) : Programme le plus utilisé en biologie moléculaire pour la recherche d'alignement entre des séquences génomiques.
✓ **Temps de calcul** proportionnel à la taille des bases de données qui double tous les 14 mois.

✓ **Heuristique** : Dans un alignement les 2 séquences partagent au moins W caractères.

```
Banque
     tggcggatggcacgaccgagactggcgataacgattga
                |     |||||||||||||     |
Requête         gttgagacgagactggccgga        W=4
```

✓ **Principe de BLAST** : Indexation de la séquence requête, et parcourt de toute la banque génomique.
=> Temps de calcul limité par le débit des données.
✓ **Indexation sur ReMIX (iBLAST)** : Indexation de la banque et parcourt de la requête.



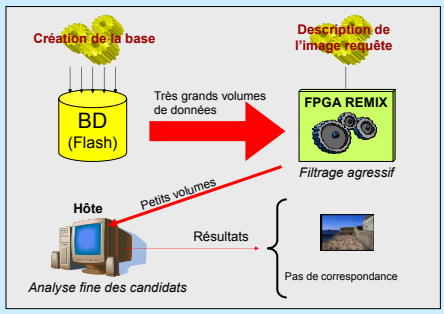
✓ **Opération à implémenter sur FPGA :**
Calcul d'alignement sur le voisinage de la clé d'index.

Recherche par le contenu dans des banques d'images

✓ **Objectif :** Retrouver une image éventuellement modifiée dans une banque d'images protégées par le copyright.
✓ 94% du chiffre d'affaire généré par le marché de la distribution de photos à usage professionnel (plus de 1600M Euros) est lié au copyright.
=> Importance de la vérification du respect du copyright.

✓ **Recherche par le contenu** : basée sur la présence de similitudes visuelles entre l'image piratée et l'image originale.
✓ **Nécessité** d'être robuste aux altérations éventuellement sévères.
✓ Une image est définie par des descripteurs locaux de 24 dimensions : 50 à 1000 descripteurs par image.
✓ Description d'une image (base ou requête) par détection automatique des points d'intérêts (Harris), puis traitement du signal autour de chaque point (convolution, dérivation, mélange spécifique).

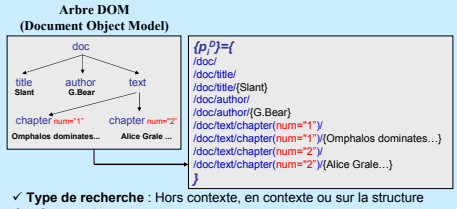
✓ **Processus de recherche :**
=> Description de l'image requête,
=> Interrogation de la base pour chaque descripteur,
=> Comptage du taux d'apparition de chaque image de la base,
=> Sélection des meilleures images.



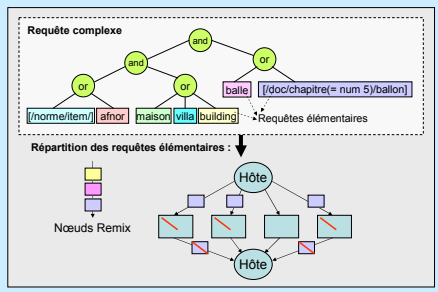
✓ **Opération à implémenter sur FPGA :** Calcul de distance euclidienne entre descripteurs multidimensionnels.

Recherche approchée d'information dans des bases de documents semi-structurés

✓ **Objectif :** Indexation et interrogation de bases de documents semi-structurés de taille de l'ordre de 50 à 100 Go.
✓ **Documents** représentés sous la forme d'une arbre DOM, assimilable à un ensemble de chemins $\{P_i^D\}$.



✓ **Type de recherche** : Hors contexte, en contexte ou sur la structure des documents.
✓ **Requêtes complexes** fragmentées en requêtes élémentaires réparties sur les nœuds ReMIX.
✓ **Opérations assemblistes** et création du résultat réalisés sur l'hôte.



✓ **Opération à implémenter sur FPGA :** Calcul de distance entre deux chemins (Distance d'édition de Levenstein).

Contacts :

Dominique LAVENIER
IRISA
Campus de Beaulieu
35042 Rennes Cedex
Tel : (33) 2 99 84 72 17
Email : lavenier@irisa.fr

Gilles GEORGES
IRISA
Campus de Beaulieu
35042 Rennes Cedex
Tel : (33) 2 99 84 74 54
Email : georges@irisa.fr

CNRS CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE
ministère délégué recherche et nouvelles technologies
DGA
INRIA