Flash : Optimisation de graines et indexation des banques génomiques sur mémoire flash reconfigurable

Journées nationales des ARC 2007

Pierre Peterlongo

IRISA, Symbiose, INRIA / CNRS / Université Rennes 1

http://www.irisa.fr/remix/arc.html

1-2 Octobre 2007

Overview

Overview

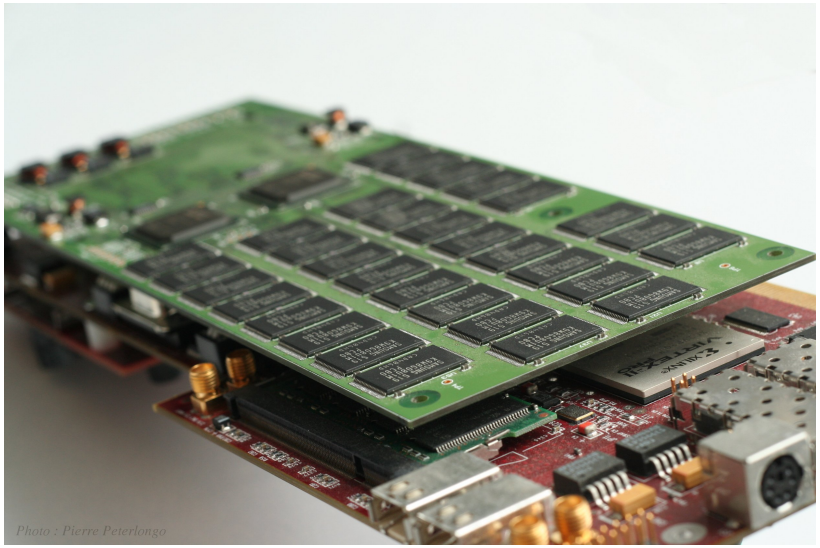
Motivations biologiques

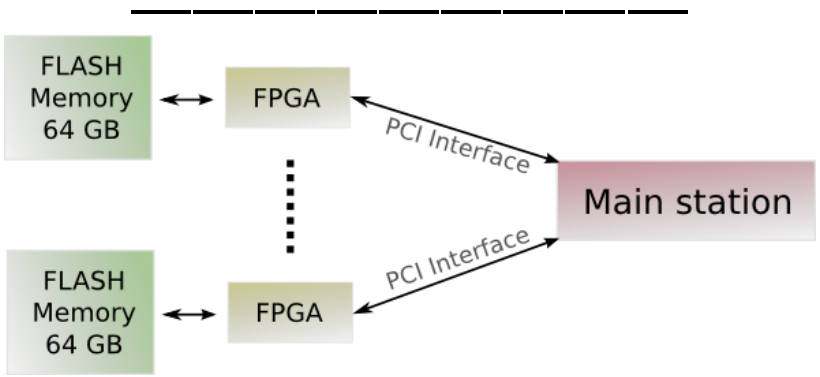
Détection de similarités dans les séquences biologiques

- Étape indispensable d'études biologiques
- Besoins grandissants :
  - Augmentation exponentielle de la taille des banques de données
  - Intérêts dans des similarités moins "fortes"

ReMIX

Point de départ...





A Reconfigurable Memory for IndeXation

- Stockage d'index de très grosse taille
- Traitement d'information rapide

Qualités principales

- Large stockage (512 GB)
- Parallélisme à gros grain
- Parallélisme à grain fin

Concrètement

Trois axes de recherche principaux

1. Synthèse d'architecture

Équipe "IP Design" LESTER Lorient

E. Casseau

Équipe-projet "Symbiose", INRIA Rennes

H. Daroles
G. Georges
D. Lavenier

Concrètement

Trois axes de recherche principaux

1. Synthèse d'architecture

   Équipe "IP Design" LESTER Lorient — E. Casseau

   Équipe-projet "Symbiose", INRIA Rennes — H. Daroles, G. Georges, D. Lavenier

2. Comparaison massive de génomes eucaryotes contre procaryotes

   INSERM U694 Angers — M. Ferre, Y. Tourmen

   Équipe-projet "Symbiose", INRIA Rennes — G. Georges, D. Lavenier

Concrètement

Trois axes de recherche principaux

1. Synthèse d'architecture

   Équipe "IP Design" LESTER Lorient — E. Casseau

   Équipe-projet "Symbiose", INRIA Rennes — H. Daroles, G. Georges, D. Lavenier

2. Comparaison massive de génomes eucaryotes contre procaryotes

   INSERM U694 Angers — M. Ferre, Y. Tourmen

   Équipe-projet "Symbiose", INRIA Rennes — G. Georges, D. Lavenier

3. Optimisation Algorithmique / matérielle

   Équipe-projet "Sequoia", LIFL, INRIA Futurs Lille — M. Giraud, G. Kucherov, L. Noé

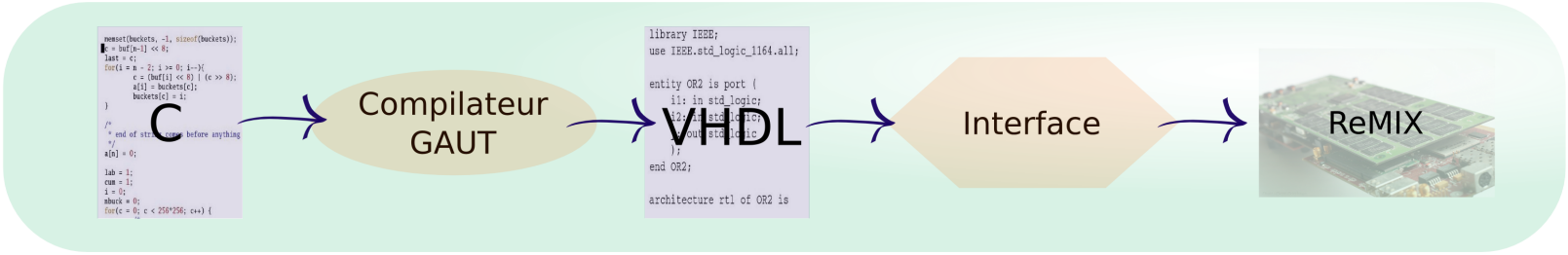
   Équipe-projet "Symbiose", INRIA Rennes — G. Georges, J. Jacques, D. Lavenier, P. Peterlongo

Overview

Synthèse d'architecture (LESTER, Symbiose)

[H. Darolles. Synthèse automatique sur plateforme FPGA, juin 2006]

[H. Darolles, plateforme ReMIX manuel d'utilisation, mai 2006]



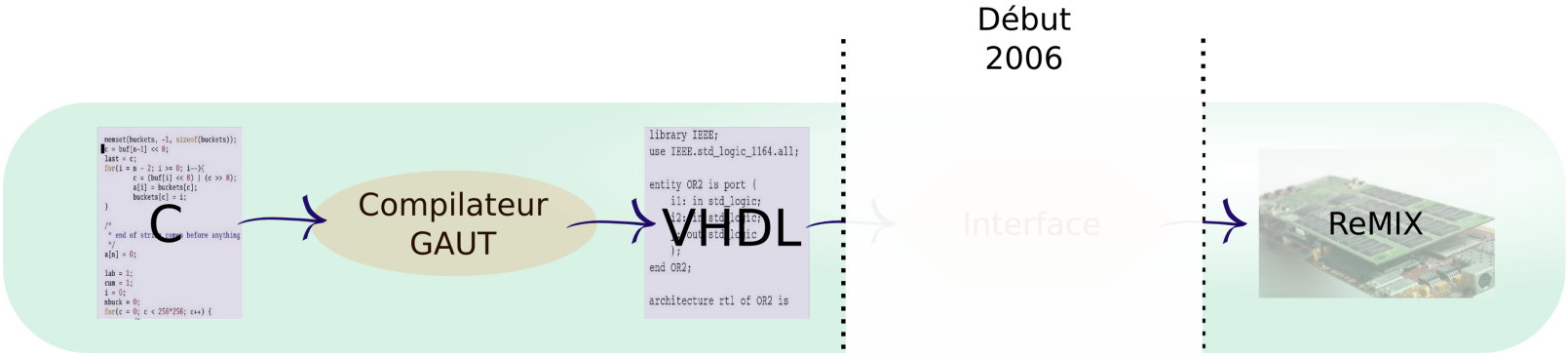
Contexte

- Programmation des FPGA
- **Problématique** : Adapter GAUT à ReMIX

Synthèse d'architecture (LESTER, Symbiose)

[H. Darolles. Synthèse automatique sur plateforme FPGA, juin 2006]

[H. Darolles, plateforme ReMIX manuel d'utilisation, mai 2006]



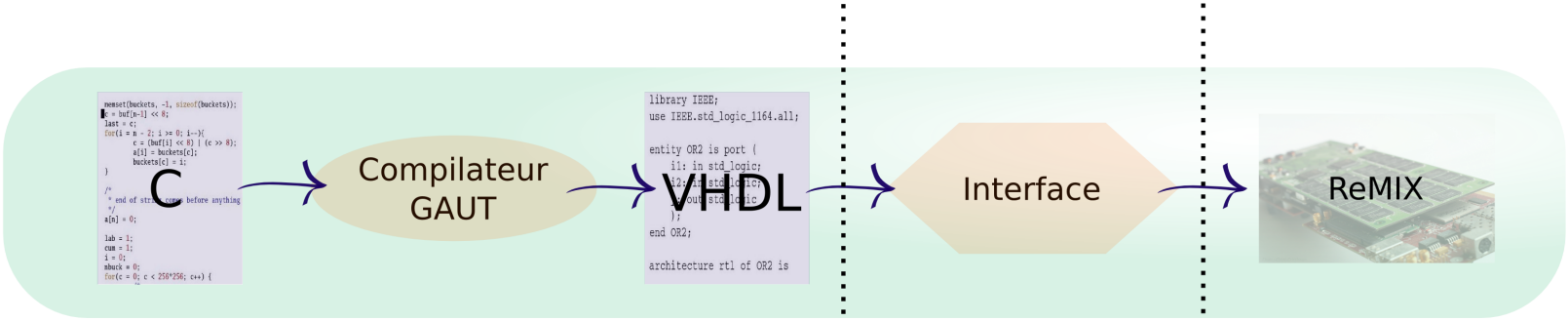
Contexte

- Programmation des FPGA
- **Problématique** : Adapter GAUT à ReMIX

Synthèse d'architecture (LESTER, Symbiose)

[H. Darolles. Synthèse automatique sur plateforme FPGA, juin 2006]

[H. Darolles, plateforme ReMIX manuel d'utilisation, mai 2006]



Contexte

- Programmation des FPGA
- **Problématique** : Adapter GAUT à ReMIX

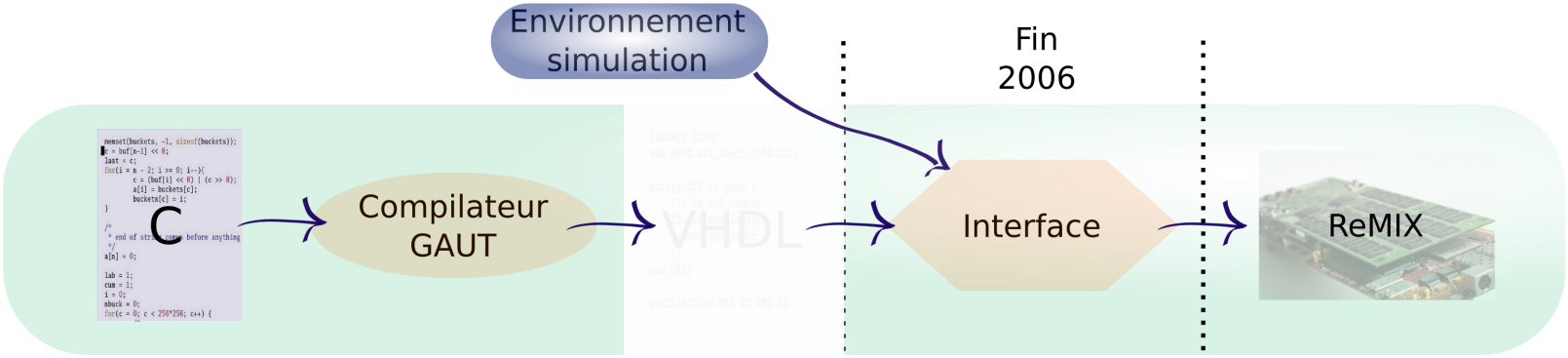
Résultats obtenus

- Stage M2 H. Darolles
  - Interface GAUT/ReMIX

Synthèse d'architecture (LESTER, Symbiose)

[H. Darolles. Synthèse automatique sur plateforme FPGA, juin 2006]

[H. Darolles, plateforme ReMIX manuel d'utilisation, mai 2006]



Contexte

- Programmation des FPGA
- **Problématique** : Adapter GAUT à ReMIX

Résultats obtenus

- Stage M2 H. Darolles
  - Interface GAUT/ReMIX
  - Environnement de simulation

Difficultés rencontrées

- GAUT initialement prévu pour le traitement du signal
- Départ de notre contact au LESTER.

Overview

Comparaison massive de génomes (INSERM U694 Angers, Symbiose)

[D. Lavenier, X. Xinchun, G. Georges, IEEE FPT 2006]

Contexte

- **But :** Détection de protéines d'origine mitochondriales
- **Moyen :** Comparaison (blast) de 270 génomes d'archeae et d'eubactéries contre des génomes eucaryotes
- **Difficulté :** Temps de calculs (estimé à 27 mois pour le génome humain)

Comparaison massive de génomes (INSERM U694 Angers, Symbiose)

[D. Lavenier, X. Xinchun, G. Georges, IEEE FPT 2006]

Contexte

- **But :** Détection de protéines d'origine mitochondriales
- **Moyen :** Comparaison (blast) de 270 génomes d'archeae et d'eubactéries contre des génomes eucaryotes
- **Difficulté :** Temps de calculs (estimé à 27 mois pour le génome humain)

Fait

- Mise en oeuvre sur ReMIX
- 400 000 protéines vs. génome humain
  - **11 jours contre 27 mois**
- Résultats en cours de traitement

Comparaison massive de génomes (INSERM U694 Angers, Symbiose)

[D. Lavenier, X. Xinchun, G. Georges, IEEE FPT 2006]

Contexte

- **But :** Détection de protéines d'origine mitochondriales
- **Moyen :** Comparaison (blast) de 270 génomes d'archeae et d'eubactéries contre des génomes eucaryotes
- **Difficulté :** Temps de calculs (estimé à 27 mois pour le génome humain)

Fait

- Mise en oeuvre sur ReMIX
- 400 000 protéines vs. génome humain
  - **11 jours contre 27 mois**
- Résultats en cours de traitement

Comment ?

- Utilisation de graine "simple" ###
- Stockage d'index sur FLASH
- Filtrage des données sur FPGA

Comparaison massive de génomes (INSERM U694 Angers, Symbiose)

[D. Lavenier, X. Xinchun, G. Georges, IEEE FPT 2006]

Contexte

- **But :** Détection de protéines d'origine mitochondriales
- **Moyen :** Comparaison (blast) de 270 génomes d'archeae et d'eubactéries contre des génomes eucaryotes
- **Difficulté :** Temps de calculs (estimé à 27 mois pour le génome humain)

Fait

- Mise en oeuvre sur ReMIX
- 400 000 protéines vs. génome humain
  - **11 jours contre 27 mois**
- Résultats en cours de traitement

Comment ?

- Utilisation de graine "simple" ###
- Stockage d'index sur FLASH
- Filtrage des données sur FPGA

$\Rightarrow$ **base pour les améliorations futures**

Comparaison massive de génomes (INSERM U694 Angers, Symbiose)

[D. Lavenier, X. Xinchun, G. Georges, IEEE FPT 2006]

Contexte

- **But :** Détection de protéines d'origine mitochondriales
- **Moyen :** Comparaison (blast) de 270 génomes d'archeae et d'eubactéries contre des génomes eucaryotes
- **D**

```
Query: 102 LQFDRPFLLAMANAGPGTNGSQFFITVVPTPHLNNHHTIFGEVTD 146
           L+   P +L+ ANAGP TN S FFI++  T  L+  H +FG + +
Sbjct: 131 LKHTGPGILSTANAGPNTNCSWFFISIAKTESLDGQHVVFGNMKE 175
```

Fait

- Mise en oeuvre sur ReMIX
- 400 000 protéines vs. génome humain
  - **11 jours contre 27 mois**
- Résultats en cours de traitement

Comment ?

- Utilisation de graine "simple" ###
- Stockage d'index sur FLASH
- Filtrage des données sur FPGA

⇒ base pour les améliorations futures

Overview

Avancées algorithmiques (Séquoia Lille, Symbiose)

[P. Peterlongo, L. Noé, D. Lavenier, G. Georges, J. Jacques, G. Kucherov, M. Giraud, PBC 2007]
[M. Giraud, G. Kucherov, D. Lavenier, L. Noé, P. Peterlongo, LAW 2007]

Contexte

- **But** : Amélioration algorithmique (temps et sensibilité)
- **Moyen** : Conception de graines avancées, appelées **graines subset**

```
Query: 102 LQFDRPFLLAMANAGPGTNGSQFFITVVPTPHLNNHHTIFGEVTD 146
           L+   P +L+ ANAGP TN S FFI++  T  L+  H +FG + +
Sbjct: 131 LKHTGPGILSTANAGPNTNCSWFFISIAKTESLDGQHVVFGNMKE 175
```

Avancées algorithmiques (Séquoia Lille, Symbiose)

[P. Peterlongo, L. Noé, D. Lavenier, G. Georges, J. Jacques, G. Kucherov, M. Giraud, PBC 2007]
[M. Giraud, G. Kucherov, D. Lavenier, L. Noé, P. Peterlongo, LAW 2007]

Contexte

- **But** : Amélioration algorithmique (temps et sensibilité)
- **Moyen** : Conception de graines avancées, appelées **graines subset**

```
Query: 102 LQFDRPFLLAMANAGPGTNGSQFFITVVPTPHLNNHHTIFGEVTD 146
           L+   P +L+ ANAGP TN S FFI++  T  L+  H +FG + +
Sbjct: 131 LKHTGPGILSTANAGPNTNCSWFFISIAKTESLDGQHVVFGNMKE 175
```

Avancées algorithmiques (Séquoia Lille, Symbiose)

[P. Peterlongo, L. Noé, D. Lavenier, G. Georges, J. Jacques, G. Kucherov, M. Giraud, PBC 2007]
[M. Giraud, G. Kucherov, D. Lavenier, L. Noé, P. Peterlongo, LAW 2007]

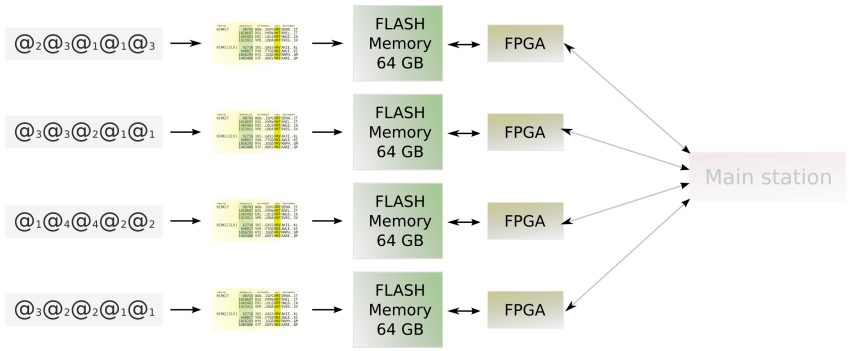
Contexte

- **But** : Amélioration algorithmique (temps et sensibilité)
- **Moyen** : Conception de graines avancées, appelées **graines subset**

```
Query: 102 LQFDRPFLLAMANAGPGTNGSQFFITVVPTPHLNNHHTIFGEVTD 146
           L+   P +L+ ANAGP TN S FFI++  T  L+  H +FG + +
Sbjct: 131 LKHTGPGILSTANAGPNTNCSWFFISIAKTESLDGQHVVFGNMKE 175
```

Fait

- Création d'ensembles de graines subset $@_1@_2@_3@_2$
- Répartition des graines sur les cartes ReMIX



Avancées algorithmiques (Séquoia Lille, Symbiose)

[P. Peterlongo, L. Noé, D. Lavenier, G. Georges, J. Jacques, G. Kucherov, M. Giraud, PBC 2007]
[M. Giraud, G. Kucherov, D. Lavenier, L. Noé, P. Peterlongo, LAW 2007]

Contexte

- **But :** Amélioration algorithmique (temps et sensibilité)
- **Moyen :** Conception de graines avancées, appelées **graines subset**

```
Query: 102 LQFDRPFLLAMANAGPGTNGSQFFITVVPTPHLNNHHTIFGEVTD 146
           L+   P +L+ ANAGP TN S FFI++  T  L+  H +FG + +
Sbjct: 131 LKHTGPGILSTANAGPNTNCSWFFISIAKTESLDGQHVVFGNMKE 175
```

Résultats

- Accélération algorithmique : 25%
- (Accélération matérielle : $\times 13$)

Avenir

- **But :** Réduction taille d'index
- **Pourquoi ?** Réduction espace et temps

Overview

Pour résumer

Buts initiaux

1. Synthèse d'architecture
2. Application à la comparaison massive
3. Avancées algorithmiques

Fait

1. GAUT/ReMIX
   - Environnement de simulation
   - interface ReMIX
2. Détection d'alignements
   - 11 jours contre 27 mois
3. Mise en place théorique et pratique de subset seeds
   - Amélioration algorithmique pure
   - Accélération matérielle

Pour résumer

Buts initiaux

1. Synthèse d'architecture
2. Application à la comparaison massive
3. Avancées algorithmiques

Fait

1. GAUT/ReMIX
   - Environnement de simulation
   - interface ReMIX
2. Détection d'alignements
   - 11 jours contre 27 mois
3. Mise en place théorique et pratique de subset seeds
   - Amélioration algorithmique pure
   - Accélération matérielle

Travaux en cours et à venir

- Limitation de taille d'index
- Préparation d'une version étendue PBC
- Exploration de nouvelles techniques
  - d'indexation
  - de comparaison

En pratique

- Collaboration fructueuse Séquoia/Symbiose
- $\longrightarrow$ Nombreuses visites
  - Passées
  - À venir