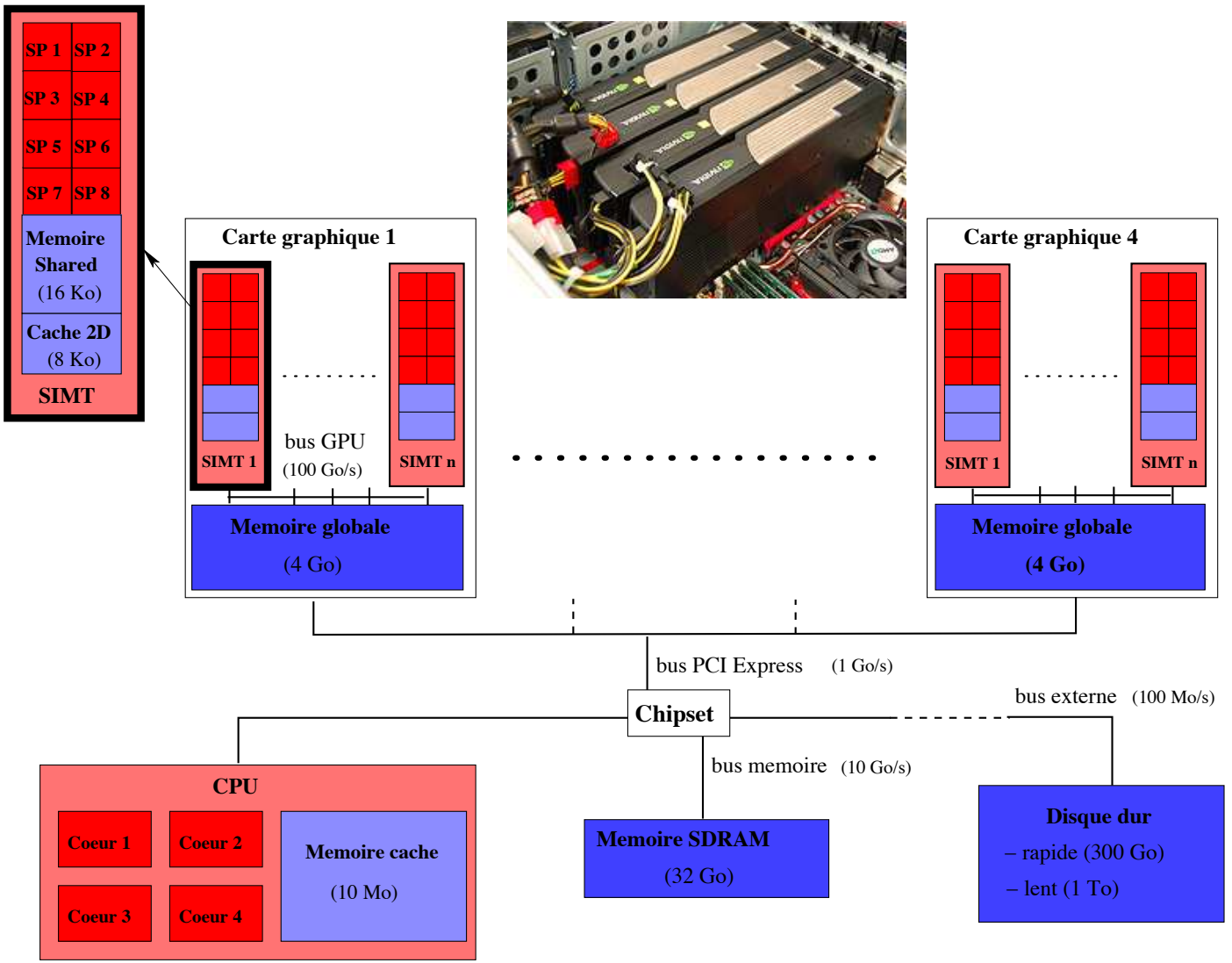
Parallélisation des calculs sur serveur multi-GPUs pour la résolution de problèmes inverses

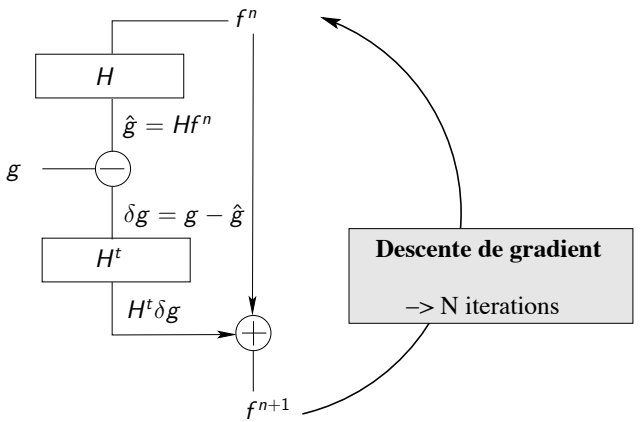
Nicolas GAC, MCF Université Paris Sud
Groupe Problèmes Inverses (GPI)
Pôle Signaux
L2S (CentraleSupélec/CNRS/Univ Paris Sud)

NVIDIA/CDS/UPSaclay meeting, LAL, 30 Mars 2015

Serveur de calcul multi GPU



Algorithme itératif



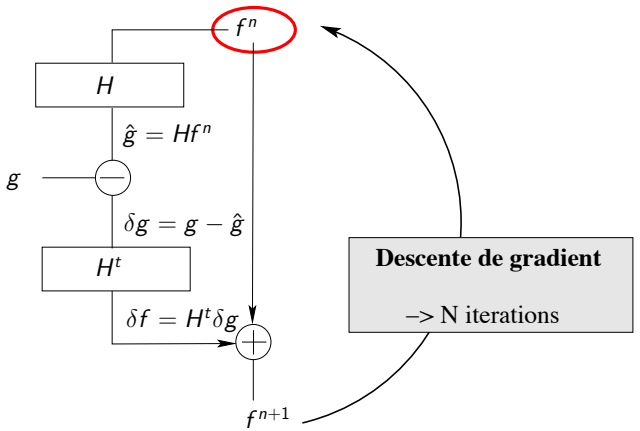
$\mathbf{g = Hf + \epsilon}$

f : objet observé
g : mesure de l'instrument
H : modèle d'acquisition
ϵ : bruit

Descente de gradient

$$J(f) = ||g - Hf||^2$$
$$f^{n+1} = f^n - \alpha \cdot \nabla J(f^n)$$
$$\nabla J(f) = -2 \cdot H^t(g - Hf)$$

Algorithme itératif



f^n : **Estimée du volume**

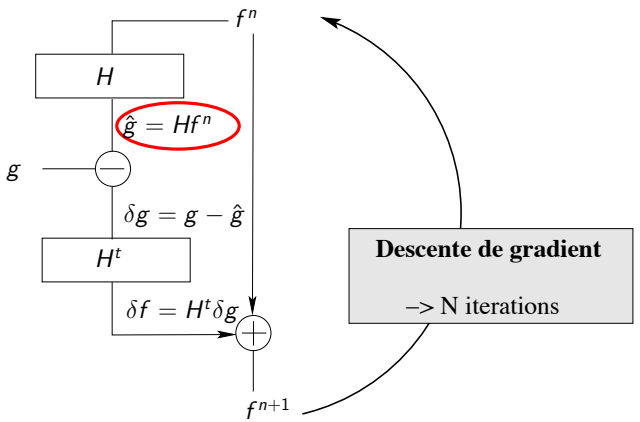
$\mathbf{g = Hf + \epsilon}$

f : objet observé
g : mesure de l'instrument
H : modèle d'acquisition
ϵ : bruit

Descente de gradient

$$J(f) = ||g - Hf||^2$$
$$f^{n+1} = f^n - \alpha \cdot \nabla J(f^n)$$
$$\nabla J(f) = -2 \cdot H^t(g - Hf)$$

Algorithme itératif



$\hat{g}$: **Estimée des données**

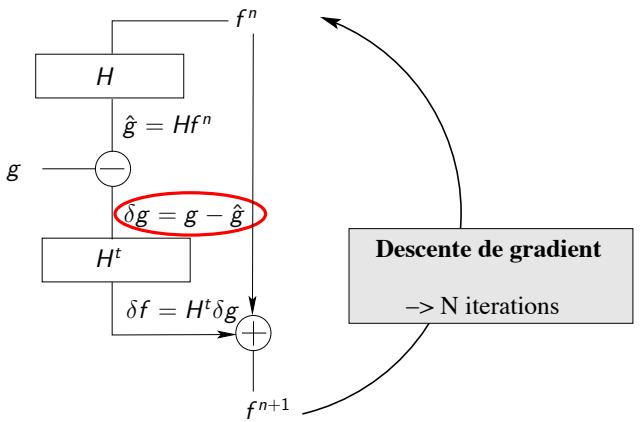
$g = Hf + \epsilon$

f : objet observé
g : mesure de l'instrument
H : modèle d'acquisition
ϵ : bruit

Descente de gradient

$$J(f) = ||g - Hf||^2$$
$$f^{n+1} = f^n - \alpha \cdot \nabla J(f^n)$$
$$\nabla J(f) = -2 \cdot H^t(g - Hf)$$

Algorithme itératif



δg : Correction des données

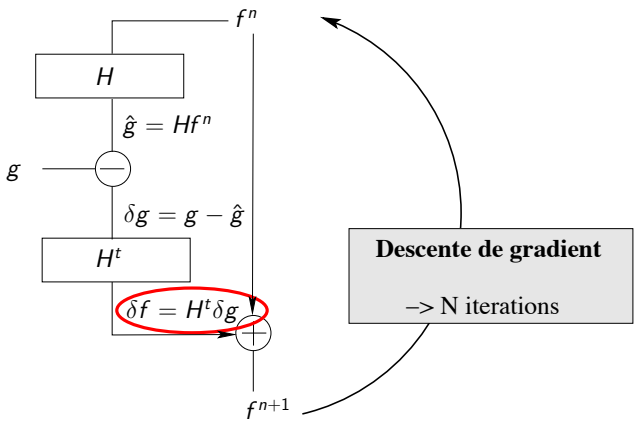
$\mathbf{g = Hf + \epsilon}$

f : objet observé
g : mesure de l'instrument
H : modèle d'acquisition
ϵ : bruit

Descente de gradient

$$J(f) = ||g - Hf||^2$$
$$f^{n+1} = f^n - \alpha \cdot \nabla J(f^n)$$
$$\nabla J(f) = -2 \cdot H^t(g - Hf)$$

Algorithme itératif



δf : **Correction du volume**

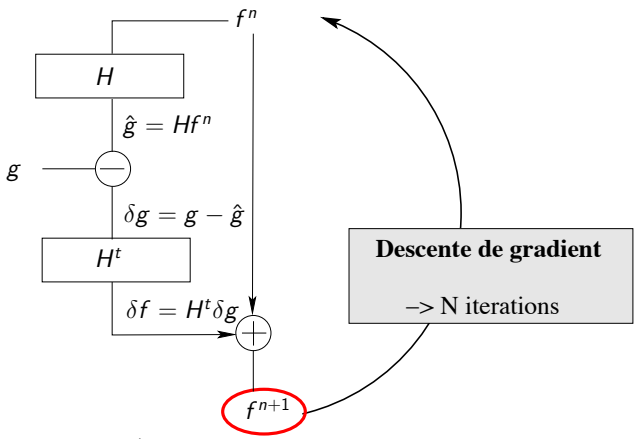
$\mathbf{g = Hf + \epsilon}$

f : objet observé
g : mesure de l'instrument
H : modèle d'acquisition
ϵ : bruit

Descente de gradient

$$J(f) = ||g - Hf||^2$$
$$f^{n+1} = f^n - \alpha \cdot \nabla J(f^n)$$
$$\nabla J(f) = -2 \cdot H^t(g - Hf)$$

Algorithme itératif



f^{n+1} : **Nouvelle estimée du volume**

$g = Hf + \epsilon$

f : objet observé
g : mesure de l'instrument
H : modèle d'acquisition
ϵ : bruit

Descente de gradient

$$J(f) = ||g - Hf||^2$$
$$f^{n+1} = f^n - \alpha \cdot \nabla J(f^n)$$
$$\nabla J(f) = -2 \cdot H^t(g - Hf)$$

Correction de vibrations mécaniques par déconvolution 1D

Collaboration avec l'IDES de l'Univ. Paris-Sud (F. Schmidt)

PhD fellowship financed by CDS (beginning autumn 2015)

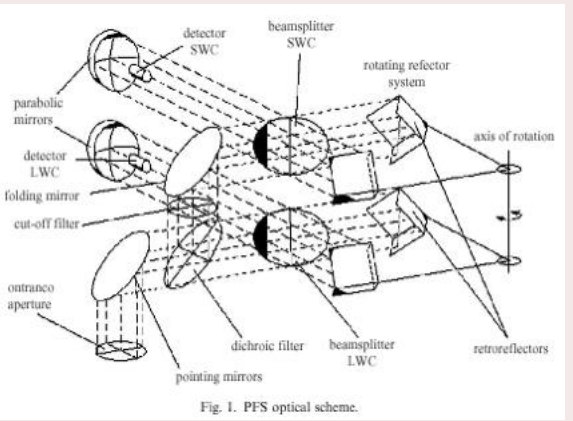
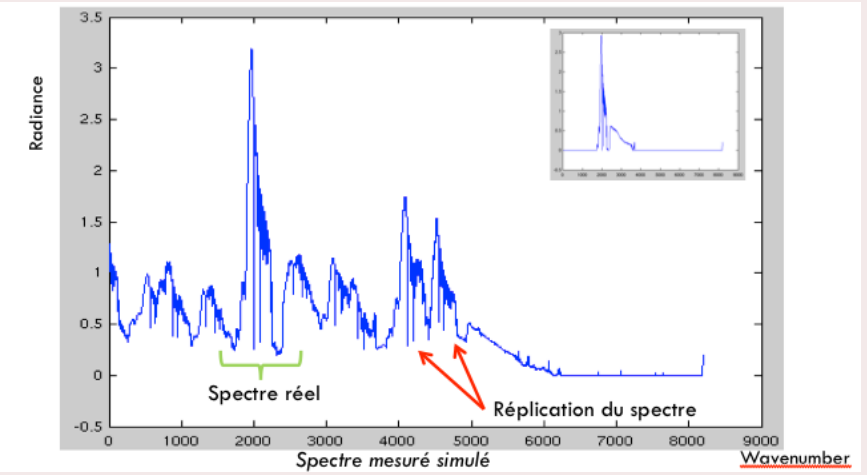


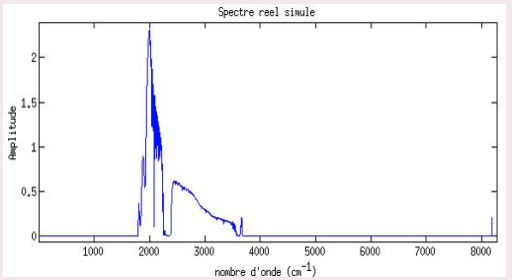
Fig. 1. PFS optical scheme.

Instrument PFS (Planetary Fourier Spectrum) de la mission MARS EXPRESS

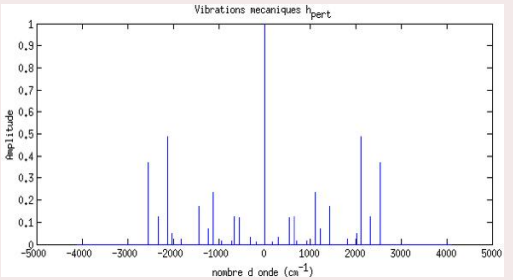


Correction de vibrations mécaniques par déconvolution 1D

Instrument modélisé par une convolution 1D

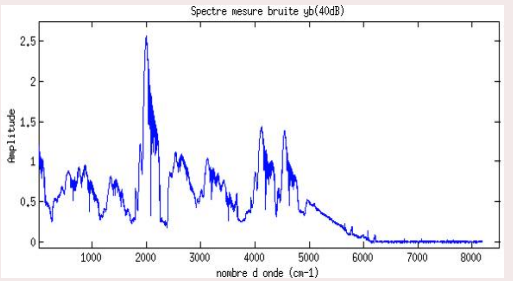


x (spectre réel) * h (instrument PFS)



=

y (spectre mesuré)



Taille gigantesque des données

Des années d'enregistrements de la mission MARS EXPRESS (2003) donc potentiellement 1 milliard de spectres (de 8192 échantillons) !

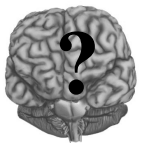
Projet ANR Magellan (OCA Nice/ENS Cachan/Telecom ParisTech)

Méthodes de reconstruction pour interféromètres du futur (très grands réseaux d'antennes)

Objectif du projet est de s'attaquer à 3 verrous :

- Complexité de la calibration
- Taille des données
- Précisions des méthodes de restauration

Algorithmes de reconstruction tomographique



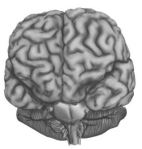
Objet à imager

APPLICATIONS
- **Controle non destructif** (réacteurs, pipelines, pièces industrielles...)
- **Imagerie medicale**
- **Transport**
- **Agro–alimentaire**
- **Micro–électronique** (controle qualité)

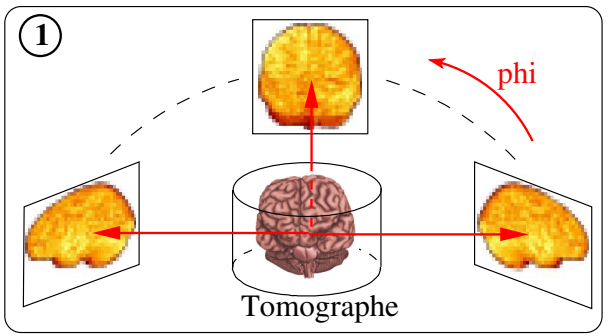
① ACQUISITION
- Scanner (X–ray, TEP..)

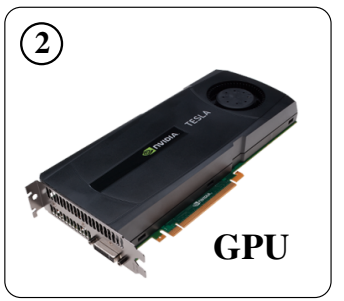
Projections

② RECONSTRUCTION
- Algorithme itératif
- Accélération sur GPU



Structure interne de l'objet





Calcul de Hf et $H^t \delta g$: choix de la méthode

❶ Calcul matriciel

➲ lecture des coefficients h_{ij} dans la mémoire SDRAM

⚠volume $2048^3 -> $ matrice H $= 1$ To !

Calcul de Hf et $H^t \delta g$: choix de la méthode

❶ Calcul matriciel

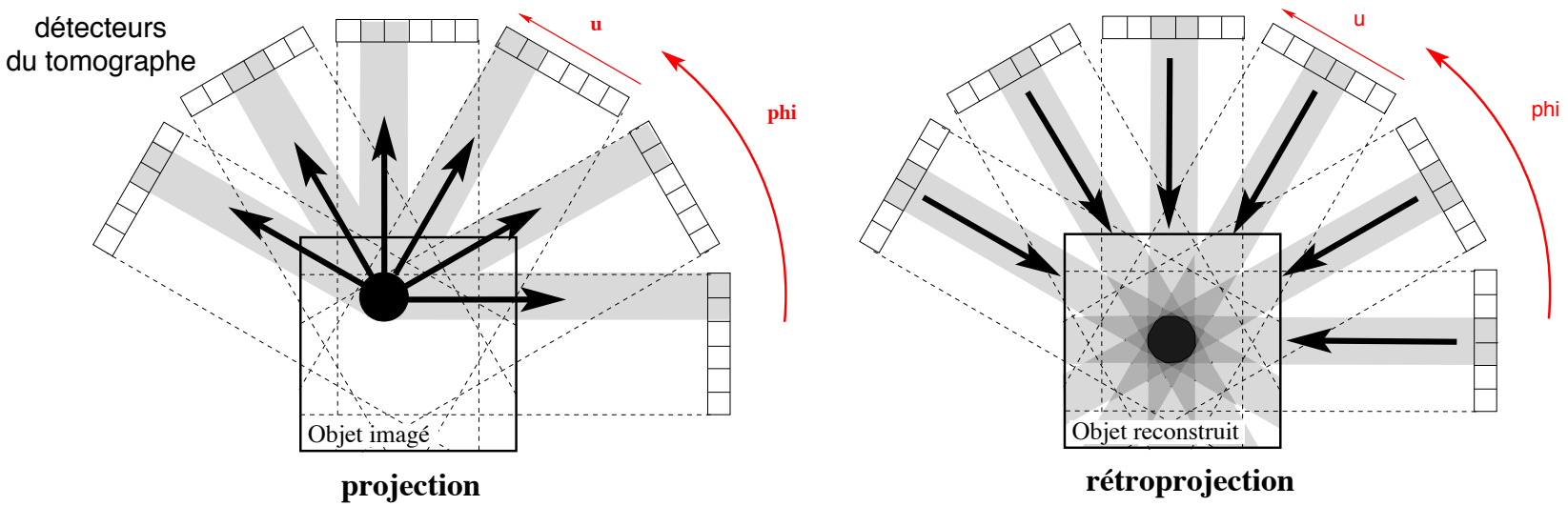
➲ lecture des coefficients h_{ij} dans la mémoire SDRAM

⚠volume $2048^3 -> $ matrice H $= 1$ To !

❷ Opérateurs géométriques

➲ calcul en ligne des coefficients h_{ij}

Paire de projection/rétroprojection en tomographie à émission (géométrie parallèle)



Temps de reconstruction mono-GPU

Opérateurs	**Temps de calcul** v1	v2
Projection $2 \times H_P$	4.1 h *(42.5 %)*	7.1 mn *(64.9 %)* $\rightarrow \times$ **35**
Rétroprojection H_{RP}^t	5.5 h *(56.9 %)*	21.8 s *(3.3 %)* $\rightarrow \times$ **908**
Convolution $3 \times D$	3.2 mn *(0.6 %)*	3.2 mn *(29.2 %)*
Autre	17 s *(0.0 %)*	17 s *(2.6 %)*
Total	9.7 h	10.9 mn $\rightarrow \times$ **53**

v1 : H_P, H_{RP}^t et D sur CPU (⚠ code "naif" non optimisé)
v2 : H_P et H_{RP}^t sur 1 GPU, D sur CPU
v3 : H_P et H_{RP}^t sur 8 GPUs, D sur CPU
v4 : H_P et H_{RP}^t sur 8 GPUs, D sur 1 GPU

Temps de reconstruction multi-GPUs

Opérateurs	**Temps de calcul**			
	v1	v2	v3	v4
Projection $2 \times H_P$	4.1 h *(42.5 %)*	7.1 mn *(64.9 %)* $\rightarrow \times$ **35**	57 s *(21.1 %)* $\rightarrow \times$ **7**	57 s *(63.3 %)*
Rétroprojection H_{RP}^t	5.5 h *(56.9 %)*	21.8 s *(3.3 %)* $\rightarrow \times$ **908**	4.0 s *(1.5 %)* $\rightarrow \times$ **5**	4.0 s *(4.4 %)*
Convolution $3 \times D$	3.2 mn *(0.6 %)*	3.2 mn *(29.2 %)*	3.2 mn *(71.1 %)*	12.1 s *(13.4 %)* $\rightarrow \times$ **16**
Autre	17 s *(0.0 %)*	17 s *(2.6 %)*	17 s *(6.3 %)*	17 s *(18.9 %)*
Total	9.7 h	10.9 mn $\rightarrow \times$ **53**	4.5 mn $\rightarrow \times$ **2.4**	1.5 mn $\rightarrow \times$ **3.0**

v1 : H_P, H_{RP}^t et D sur CPU (⚠ code "naif" non optimisé)
v2 : H_P et H_{RP}^t sur 1 GPU, D sur CPU
v3 : H_P et H_{RP}^t sur 8 GPUs, D sur CPU
v4 : H_P et H_{RP}^t sur 8 GPUs, D sur 1 GPU

Temps de transfert mémoire PC - GPU

	1 GPU	**8 GPUs**
Projecteur H_P	10 %	37.5 %
Rétroprojecteur H_{RP}^t	1.4 %	6.8 %
Convolution D	68.9 %	

Proportion du temps de traitement consacré au transfert mémoire entre le PC et la carte graphique pour chaque opérateur lors de la reconstruction d'un volume de 1024^3 à partir de 256 projections.