

Interactive Visualization and On-Demand Processing of Large Volume Data: A Fully GPU-Based Out-Of-Core Approach.

Jonathan Sarton - Nicolas Courilleau - Yannick Remion - Laurent Lucas

CReSTIC – Université de Reims Champagne-Ardenne – France ICube – Université de Strasbourg – France







Introduction

Background and motivations

Large volume data, how to

- interactively visualize them
- process them on-the-fly ?
- \rightarrow interesting to use GPUs !



Background and motivations

Large volume data, how to

- interactively visualize them
- process them on-the-fly ?
- ightarrow interesting to use **GPUs** !

Issue : memory occupation

- Large datasets
- \gg GPU and CPU physical memory !
- Interactive manipulation complicated



\rightarrow Elaborate out-of-core algorithms

Out-of-core data access



Gigavoxels



[Crassin et al., ACM SIGGRAPH i3D, 2009]



GPU data cache

Or

-

Octree

Multi-resolution Page Table







[Hadwiger et al., IEEE SciVis 2012]

4

Out-of-core data access

Better for very large volume !!



Gigavoxels



[Crassin et al., ACM SIGGRAPH i3D, 2009]



Octree

GPU data cache

Or

Multi-resolution Page Table







[Hadwiger et al., IEEE SciVis 2012]

- Multi-resolution: to choose the desired level of detail
 - \Rightarrow Reduces the amount of data



Data representation and storage

3D mipmap

- Multi-resolution: to choose the desired level of detail
 - \Rightarrow Reduces the amount of data
- **Bricking**: Volume subdivided into small bricks (e.g 32³, 64³).
 - \Rightarrow Allows the out-of-core approach

Data compression with LZ4 algorithm

- Loss less
- Good compression ratio
- Real-time decompression





Multi-resolution, multi-level page table hierarchy



Multi-resolution, multi-level page table hierarchy



Multi-resolution, multi-level page table hierarchy



- One page = 3D coordinates of the bloc in the next cache level + one flag:
 - Mapped
 - Unmapped
 - Empty
- Implementation: CUDA
 3D Textures
- Cache replacement algorithm: Least Recently Used (LRU)

Virtual addressing



Cache miss



How to allow on-demand processing of any part of a large volume during its visualization ?

- 1. Cache usage updates
- 2. Brick requests management

A GPU data structure fully managed on GPU

Advantages

- Avoids many data transfers between CPU and GPU
- Take advantage of the massively parallel environment of GPUs
- Free the CPU for other eventual processing

Brick request management on GPU



- Size = number of bricks in the multi-resolution volume
- Marked with a timestamp

CPU / GPU transfer

$\textbf{GPU} \rightarrow \textbf{CPU}$ communications A simple list with the requested brick IDs



$\begin{array}{l} \textbf{GPU} \leftarrow \textbf{CPU communications} \\ \textbf{Only the bricks ! (With CUDA Zero Copy)} \end{array}$

Model in action: interactive visualization & on-demand processing on GPU

Virtual miscroscope

2D multi-resolution visualization of a high resolution image stack. Interactive navigation:

- move and zoom in a slide
- navigate through the volume from slide to slide





Virtual miscroscope ...

2D multi-resolution visualization of a high resolution image stack. Interactive navigation:

- move and zoom in a slide
- navigate through the volume from slide to slide

+ on-demand processing

Region-growing from a voxel selected by the user in the screen space



Virtual miscroscope ...

2D multi-resolution visualization of a high resolution image stack. Interactive navigation:

- move and zoom in a slide
- navigate through the volume from slide to slide

+ on-demand processing

Region-growing from a voxel selected by the user in the screen space

Cache miss due to processing outside the screen space !

 $\begin{array}{l} \mbox{Electron micorsocpy dataset} \\ \mbox{4096} \times 3072 \times 2130 \mbox{ Bbits} \approx \mbox{27 GB} \\ \mbox{Rendering performance:} \approx \mbox{250 FPS} \end{array}$



Ray-guided approach

- Intuitive visibility selection: no additional culling calculation
- Intuitive out-of-core integration: only load visible bricks on GPU cache



Datasets



 $\begin{array}{l} \mbox{Primate hippocampus}\\ \mbox{Light sheet microscope}\\ 2160\times2560\times1072 \ 16bits\approx \mbox{12 GB} \end{array}$



 $\begin{array}{l} \mbox{Mouse brain} \\ \mbox{Histological scanner} \\ \mbox{64000} \times 50000 \times 114 \mbox{ RGBA} \approx \mbox{1.5 TB} \end{array}$

Performances – frames frequency

55 49,4 50 47,6 45 40 35 On a single workstation 30 NVidia GeForce Titan X 25 20 15 10 5 n Dataset 1 – 12 GB Dataset 2 – 1,5 TB Primate hippocampus Mouse brain

Performances – frames frequency

117,9 On a single workstation NVidia GeForce Titan X Dataset 1 – 12 GB Dataset 2 – 1,5 TB Primate hippocampus A farme A capacity Transfer Transfer



- Primate hippocampus ($2160 \times 2560 \times 1072 \approx 12$ GB)
 - Brick size: $64^3 \implies \approx 27000$ bricks (7 LOD)
 - One virtualization level
 - \rightarrow Need 1.2~MB on GPU
- Mouse brain (64000 imes 50000 imes 114 pprox 1.5 TB)
 - Brick size: $64^3 \implies 3.13$ million bricks (10 LOD)
 - One virtualization level \rightarrow \approx 63~MB needed on GPU
 - Two virtualization levels $\rightarrow \approx 13~MB$ needed on GPU

Conclusion

- Out-of-core data management: multi-resolution multi-level page table hierarchy
 - Entirely managed on GPU
 - GPU CPU communication reduced
 - Good rendering frequency even for very large volume of data (> TB)
 - Weak GPU memory and computational footprint
 - General purpose context : interactive visualization & on-demand processing

VIS2019

Interactive Visualization and On-Demand Processing of Large Volume Data: A Fully GPU-Based Out-Of-Core Approach.

Jonathan Sarton - sarton@unistra.fr Nicolas Courilleau - nicolas.courilleau@neoxia.com Yannick Remion - yannick.remion@univ-reims.fr Laurent Lucas - laurent.lucas@univ-reims.fr





