

To 3D or not to 3D?

Why GPUs Are Critical
for
3D Mass Spectrometry Imaging

Eri Rubin
SagivTech Ltd.

SagivTech Snapshot

- Established in 2009 and headquartered in Israel
- Core domain expertise: GPU Computing and Computer Vision
- What we do:
 - Technology
 - Solutions
 - Projects
 - EU Research
 - Training
- GPU expertise:
 - Hard core optimizations
 - Efficient streaming for single or multiple GPU systems
 - Mobile GPUs



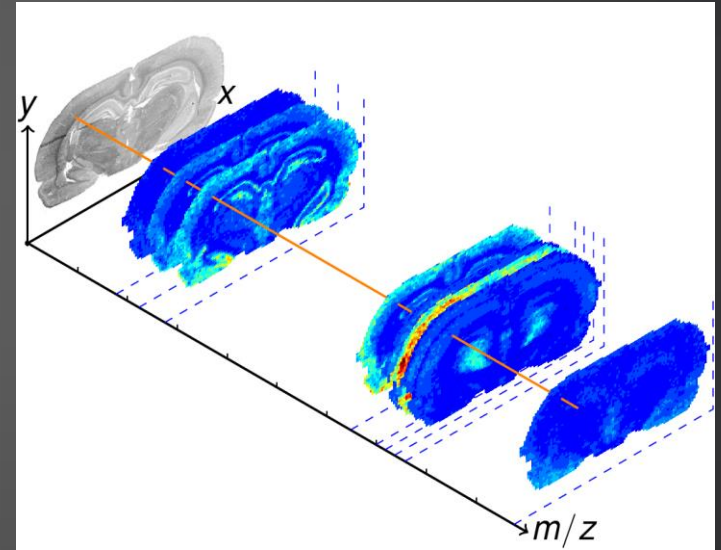
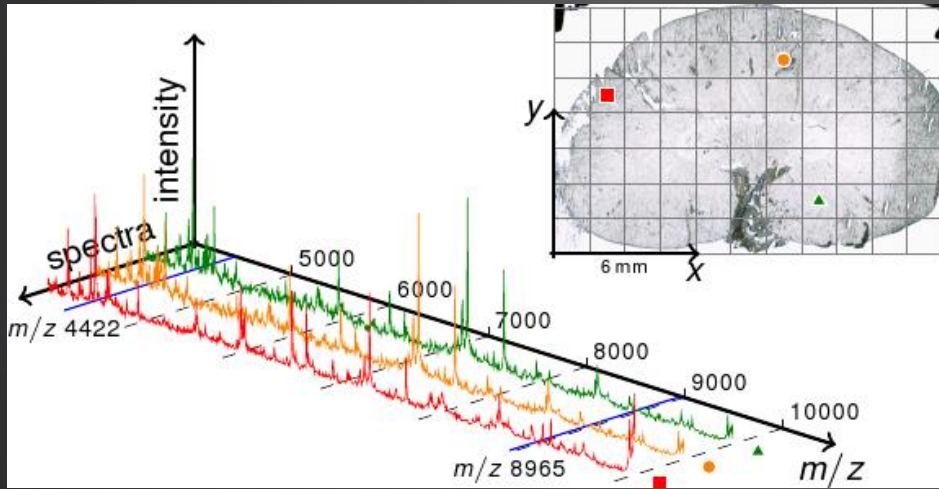
What is Mass Spectrometry ?

- A sample is ionized, for example by bombarding it with electrons.
- Then, some of the sample's molecules break into charged fragments.
- These ions are then separated according to their mass-to-charge ratio.

What is Mass Spectrometry ?

Two ways of looking at MALDI data:

- 1) Set of spectra measured at different positions
- 2) Set of images representing molecular distribution for different m/z values



MALDI imaging as a BIG DATA problem

- Big Data
 - A 2D MALDI-IMS dataset exceeds 1 gigabyte, typically comprising 5.000-50.000 spectra of approximately 10.000 bins length.
 - A 3D MALDI-IMS dataset is built of 10-50 2D datasets of serial sections, reaching up to 100 gigabytes per dataset.
- Complex Algorithms

Probabilistic Latent Semantic Analysis

- PLSA - A PCA alternative for detecting strong components
- A measure of image spatial chaos
- Used for detecting strong components in hyper spectral data (PCA alternative)
- Uses simple algebraic operations
- Algebra is a perfect fit for the GPU!

PLSA- Results

Num Channels	Num Spectra	Num Components	CPU time[sec]	GPU time[sec]	Factor
900	125	15	3.05	0.842	3.62
900	125	64	8.5	0.872	9.75
1800	250	64	36.5	1.607	22.71
3600	500	64	128.91	3.532	36.50
7200	1000	64	525.13	11.32	46.39
1800	250	128	56.4	1.85	30.49
3600	500	256	402.67	6.74	59.74

A measure of image spatial chaos

- Images can contain real objects or just noise
- Measure the “spatial chaos”
- Images with objects have less chaos.
- For hyper spectral data:
 - Each image comes from a spectra
 - Images with less chaos correspond to an interesting spectra. Peak picking!
 - Can be used to identify molecules

MOC Results

- Depends on search radius!
- Per image:
 - CPU i5 2.5GHz - 310ms per image
 - GPU k20 – 1.6ms per image ~x190

PCA acceleration via SVD acceleration

- The SVD (Singular value decomposition) and Scores calculation sections of the PCA were implemented.
- The SVD is defined by:

$$A = U * S * V^t$$

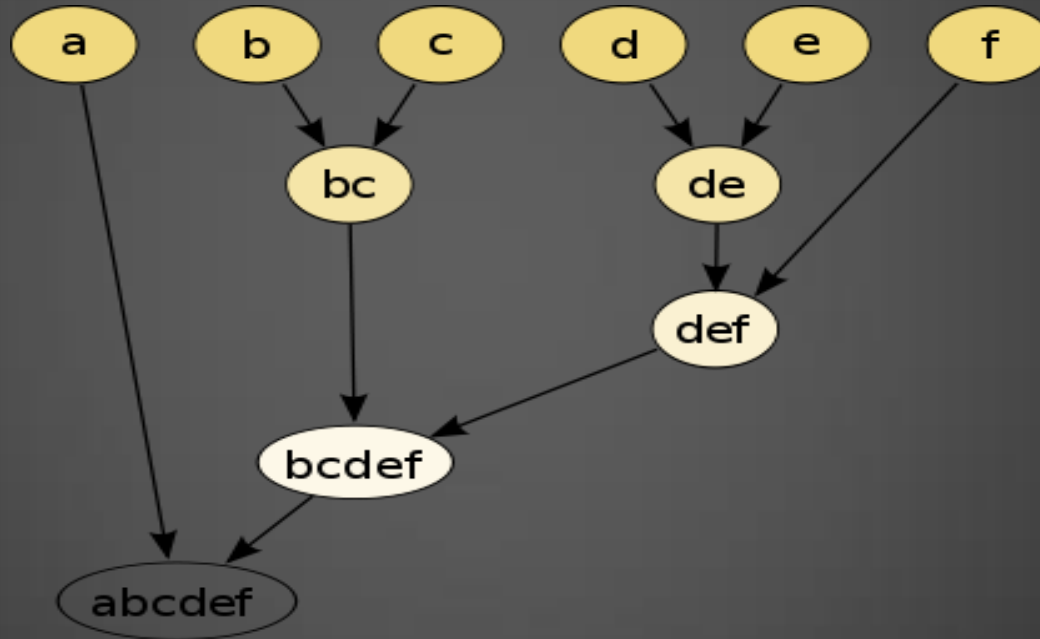
- The SVD is the most time consuming section of the PCA.
- The SVD implementation uses the CULA library.

SVD GPU Results (Kepler K20)

- The SVD computation on the GPU

Width	Height	Time in Seconds
256	256	0.0092
512	512	0.3
1024	1024	1.2
2048	2048	4.7
4096	4096	18.9
4159	6972	26.7

Hierarchical Clustering Distance



Hierarchical Clustering Distance

- The distance calculation is defined as matrix multiplication with its transposed matrix.
- CUBLAS is used to perform an optimized matrix multiplication.
- CUBLAS functionality is used also to transpose the matrix of signals in the device memory.
- GPU kernels were written to perform the final normalization and conversion to single precision.
- The Thrust library is used for sorting.
- The computation is done in blocks.

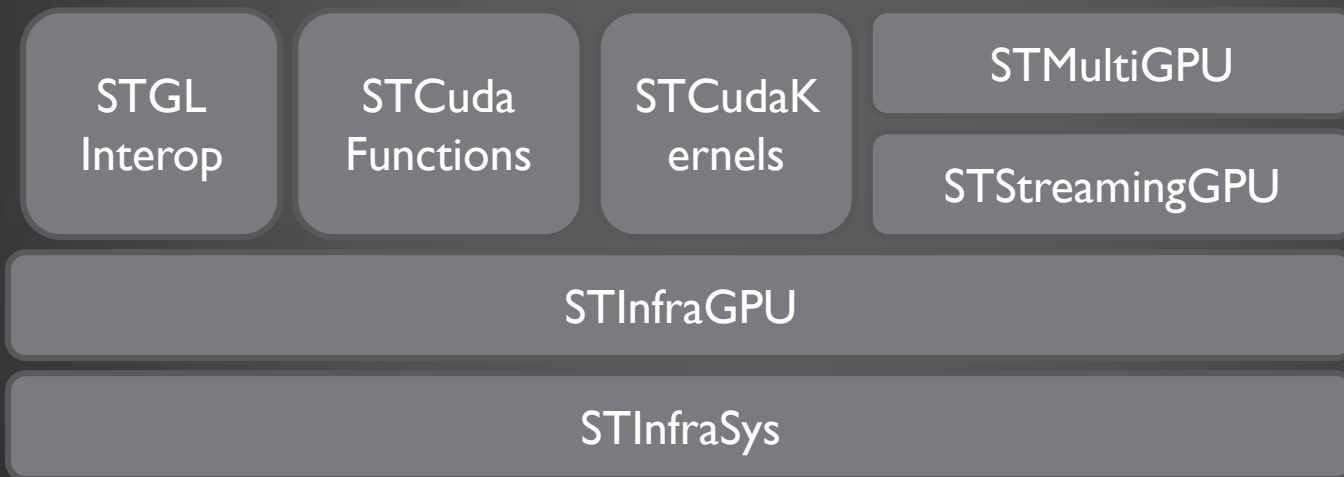
Results

Num signalsx	data per signal	Number of minimal distances	GPU Memory GB	Time (seconds)
40000	1000	10000	2.0	4.5
40000	2000	10000	2.37	6.2
40000	3000	10000	2.77	7.9

about x20 from CPU results

SagivTech Infra Stack

Our Infra is composed of a set of modules

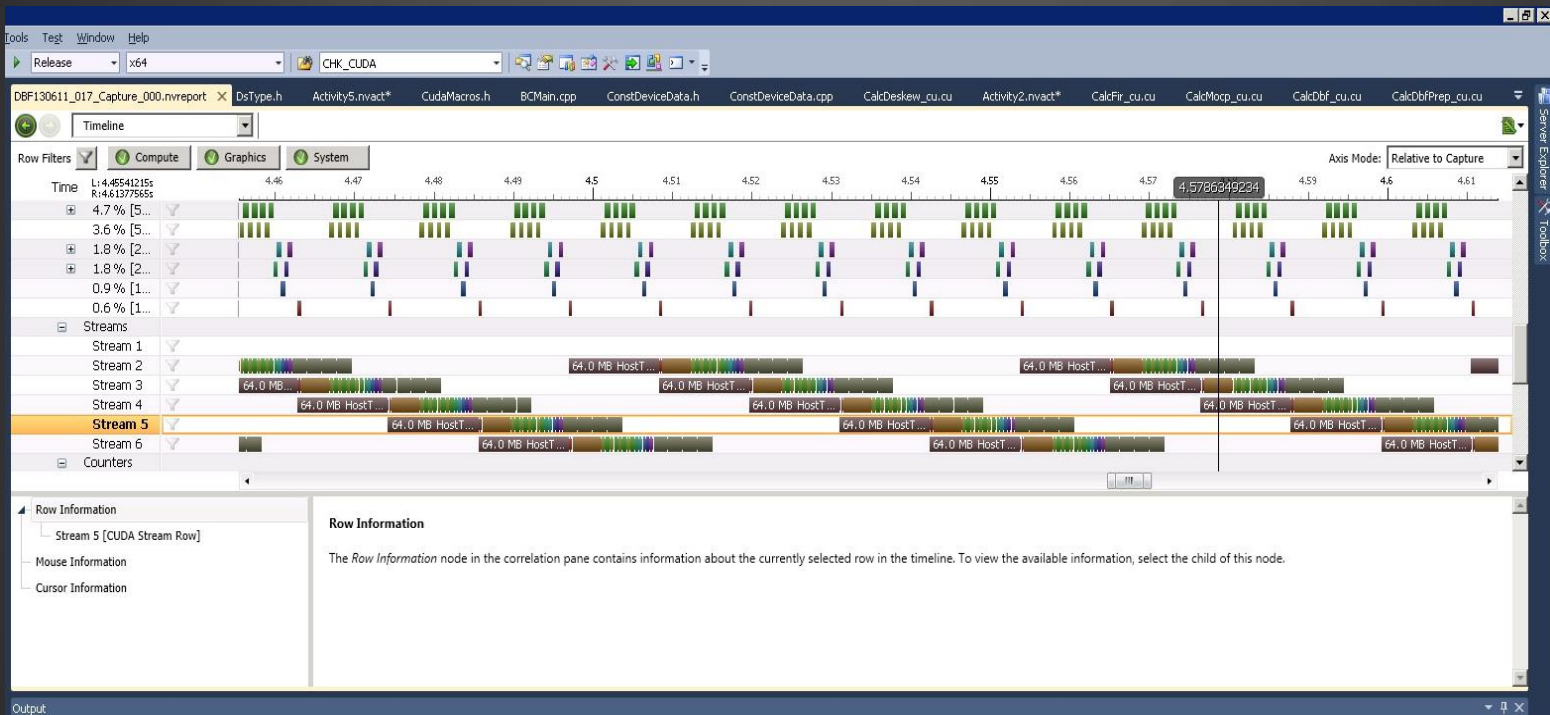


Main Attributes of SagivTech's Streaming Infrastructure

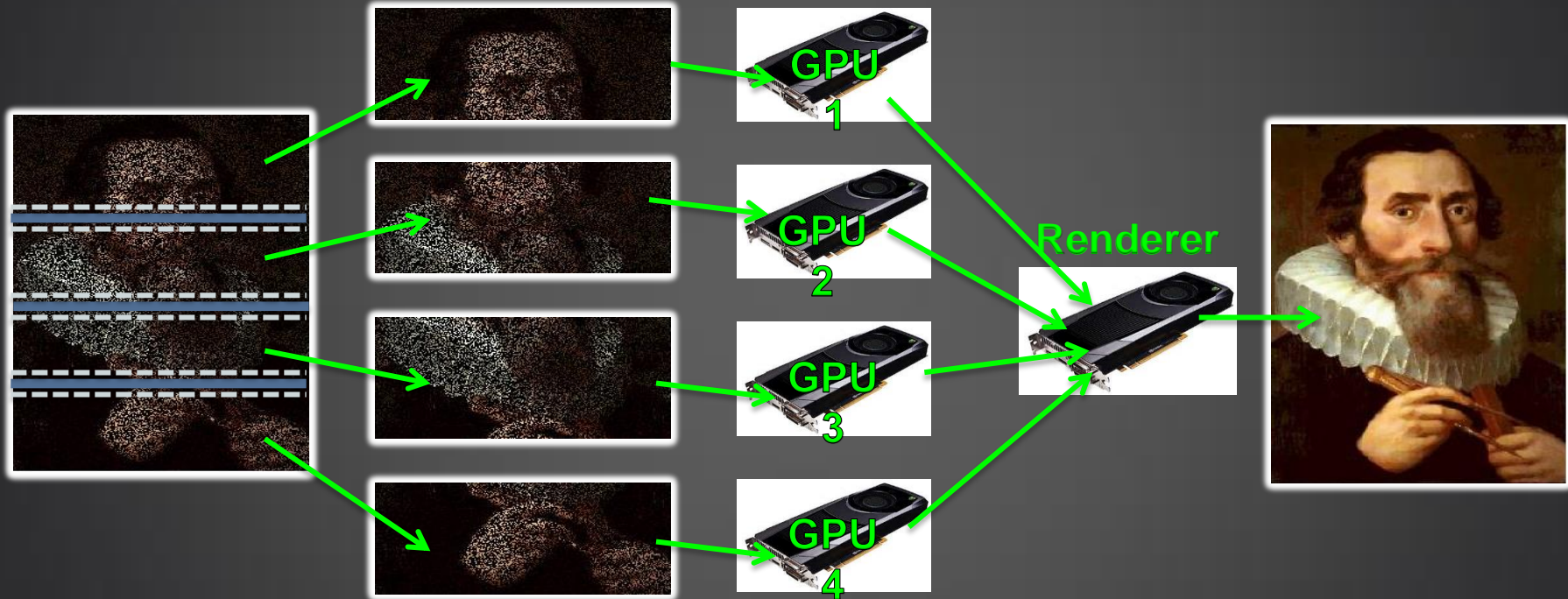
- **Pipelining:** hides memory transfer overhead between CPU and GPU
- **Asynchronous work:** allows job launch on multiple GPUs without waiting for one GPU to finish
- **Peer-to-peer communication:** enables transfer of data between multiple GPUs within the same system



GPU streaming





SagivTech Presents: A middleware for Real Time Multi GPU




ST MultiGPU Real World Use Case

SagivTech Multi-GPU Demo


Source Window:  Result Window: 

Configuration

Demo Mode: TV Full Screen Epsilon: 0.5 Lambda: 0.1
Active GPUs: 1 Inner Loops: 20 Outer Loops: 8
Pipe Size: 1 Normal Noise Apply



GPU Utilization %		Global Stats	
GPU1: 69	GPU2: 0	FPS: 4.25	Scaling (1,1): 1.00
GPU3: 0	GPU4: 0	GFlops: 574.7	Latency: 189.38





One GPU
One pipe
Utilization: ~70%

- FPS: 4.25
- Scaling: 1.00
- Note the gaps in the profiler


ST MultiGPU Real World Use Case

SagivTech Multi-GPU Demo


Source Window:  Result Window: 

Configuration

Demo Mode: TV Full Screen Epsilon: 0.5 Lambda: 0.1
Active GPUs: 1 FIT Video Inner Loops: 20 Outer Loops: 8
Pipe Size: 4 Pause Normal Noise Apply



GPU Utilization %		Global Stats	
GPU1: 98	GPU2: 0	FPS: 5.41	Scaling (1,4): 1.00
GPU3: 0	GPU4: 0	GFlops: 730.9	Latency: 517.95



One GPU



4 pipes

Utilization: 98%

- FPS: 5.41
- Scaling: 1.27
- Better utilization using pipes

ST MultiGPU Real World Use Case

SagivTech Multi-GPU Demo


Source Window:  Result Window: 

Configuration

Demo Mode: TV Epsilon: 0.5 Lambda: 0.1
Inner Loops: 20 Outer Loops: 8 Normal Noise

Active GPUs: 4 Pipe Size: 4

GPU Utilization %		Global State	
GPU1: 98	GPU2: 96	FPS: 20.46	Scaling (4,4): 3.79
GPU3: 98	GPU4: 98	GFlops: 2765.9	Latency: 173.63



SAGIVTECH

Four GPUs
Four pipes
Utilization: 96%+

- FPS: 20.46
- Scaling: 3.79 – Near linear Scaling!
- Note NO gaps in the profiler

3D Massomics

- This project is funded by the European Union, FP7 HEALTH programme Grant agreement no. 305259.



GTC 2015, San Jose



Thank You

For more information please contact

Nizan Sagiv

nizan@sagivtech.com

+972 52 811 3456



SAGIVTECH

