Karlsruhe Institute of Technology
European Synchrotron Radiation Facility

A GPU-based Architecture for Real-Time Data

S. Chilingaryan[1], M. Vogelgesang[1],

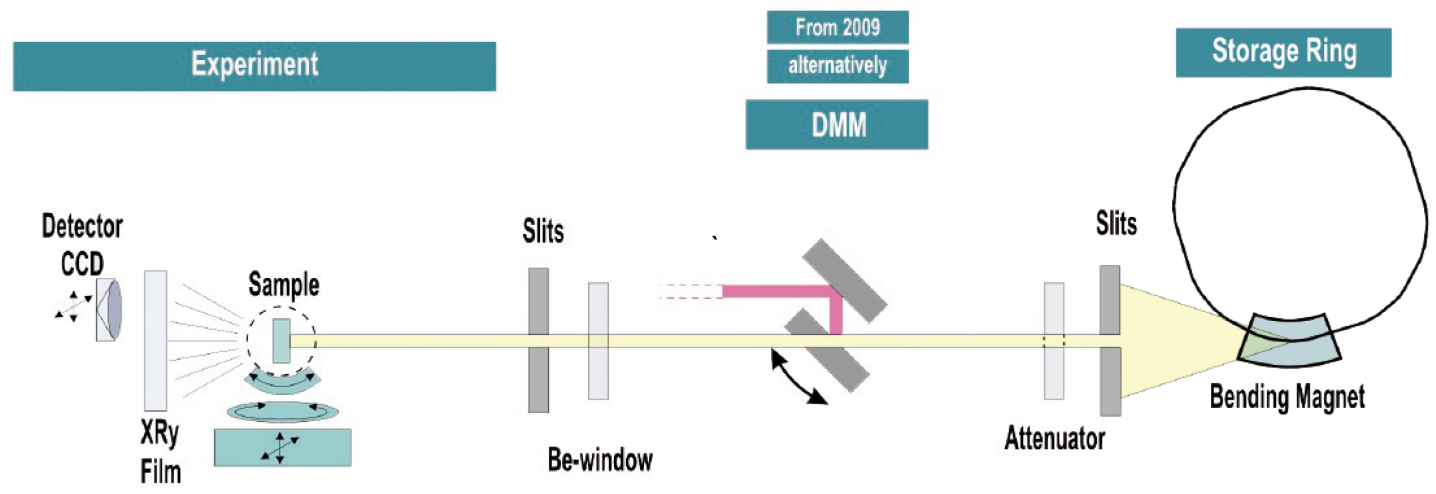
[1] *Karlsruhe Institute of Technology, Karlsruhe, Germany*

High Speed X-Ray Imaging

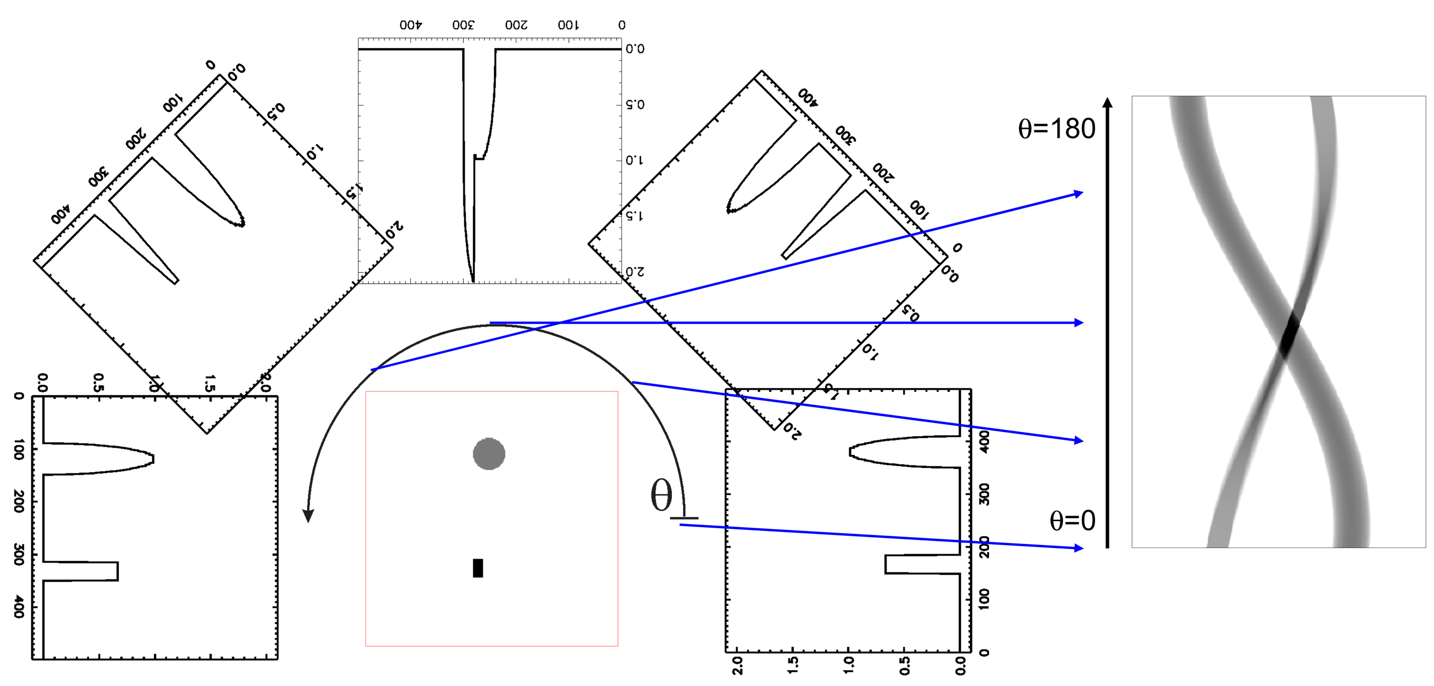
X-ray imaging permits spatially resolved visualization of 2D and 3D structures in materials and organisms which is crucial for understanding their properties. Furthermore, it allows one to recognize defects in devices from the macro- down to the nano-scale. Providing millions of pixels, each with a digitization depth of 12 bits or more, and several thousand frames per second, modern synchrotron can produce data sets of gigabytes in a few seconds. We have developed a high performance imaging station based on NVIDIA GPUs and parallelized the reconstruction software employed at the micro-tomography beamline at KIT and ESRF. Using the built setup, we were able to reduce reconstruction time of typical data-set below one minute.

Example for 3D X-Ray imaging. The functional groups of a flightless weevil are colored

Tomography at Synchrotron Light Sources



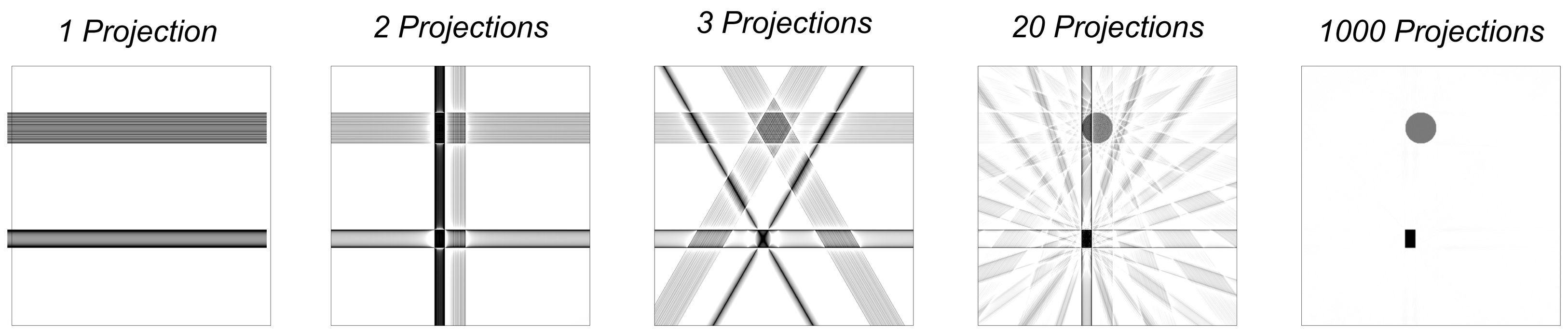
The sample, evenly rotating in the front of a pixel detector, is penetrated by X-rays produced in the synchrotron



The pixel detector registers series of parallel 2D projections of the sample density at different angles.

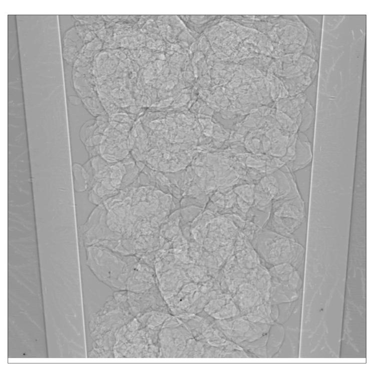
3D Image Reconstruction

According to the Back Projection algorithm, the pixel at position *(x,y,z)* is computed by $\sum_{p=1}^{P} I_p(x\cdot\cos(p\alpha)-y\cdot\sin(p\alpha), z)$, where *P* is the number of projections α is the angle between projections, and I_p is the image of *p*-th projection.



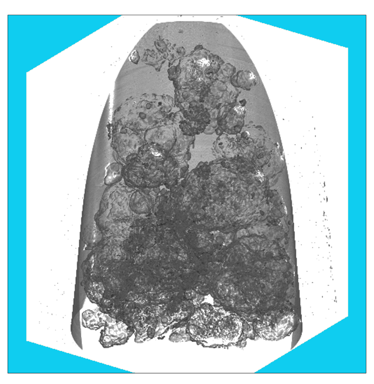
Filtered back-projection is used to reconstruct 3D images from a manifold of 2D projections. The projection values are smeared back over 2D cross sections and integrated over all projection angles. To reduce blurring effect the projections are filtered in the Fourier space before being back projected.

Typical Setup



Sample: Plastic holder with porose polyethylene grains
Source data: 24GB (2000 projections, 3 Mpix, 32 bits)
3D Image: 11GB (3 Gpix, 32 bits)
Complexity: 53 Tflop back-projection + 0.6 Tflop filtering

Goal: Reconstruct 3D image in 1 minute



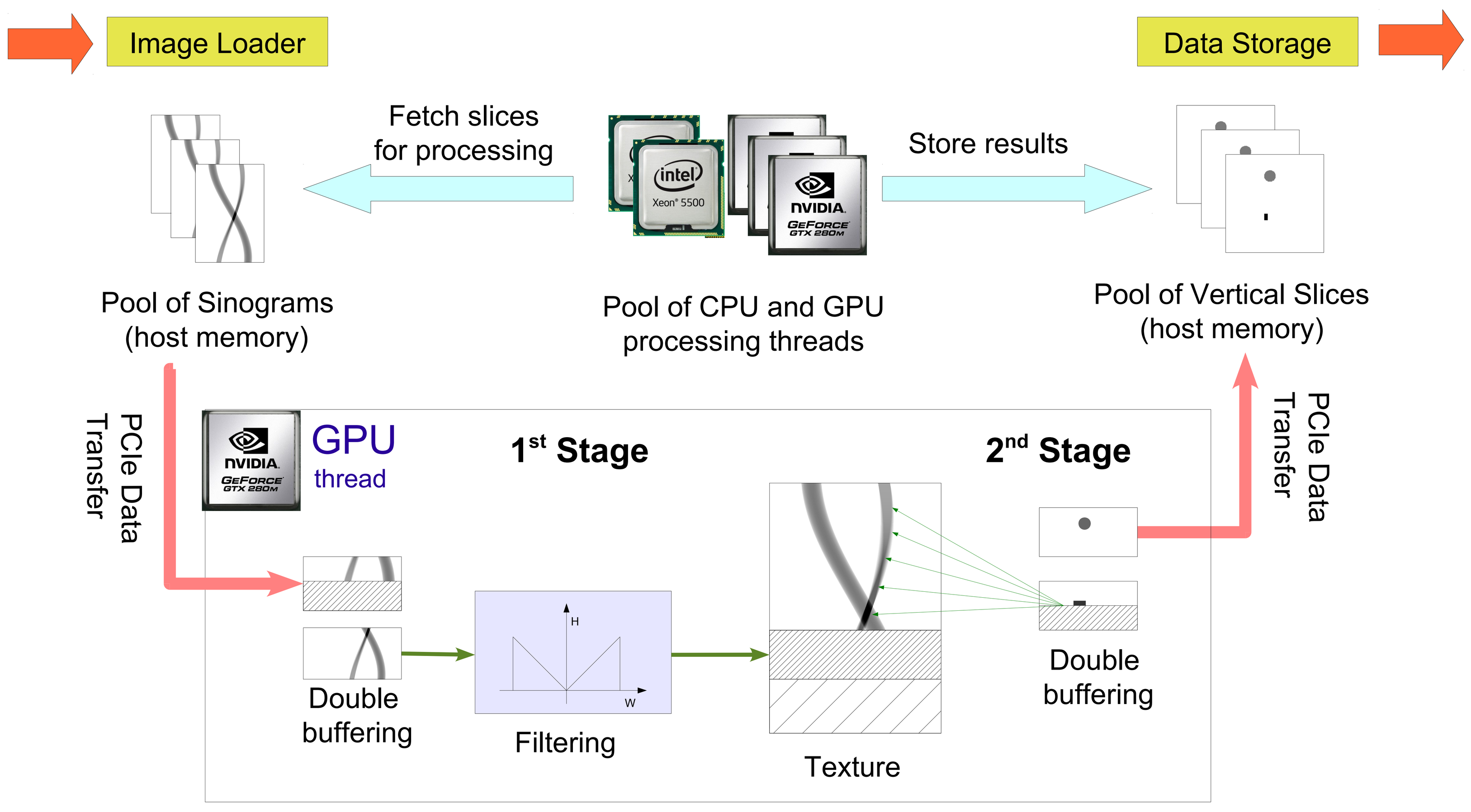
Software Optimizations

Architecture
- All GPUs are used for reconstruction and all CPUs are used to preprocess projections
- The data is prefetched from disk while CPUs and GPUs are crunching loaded data
- Both system and GPU memory are allocated once at application startup

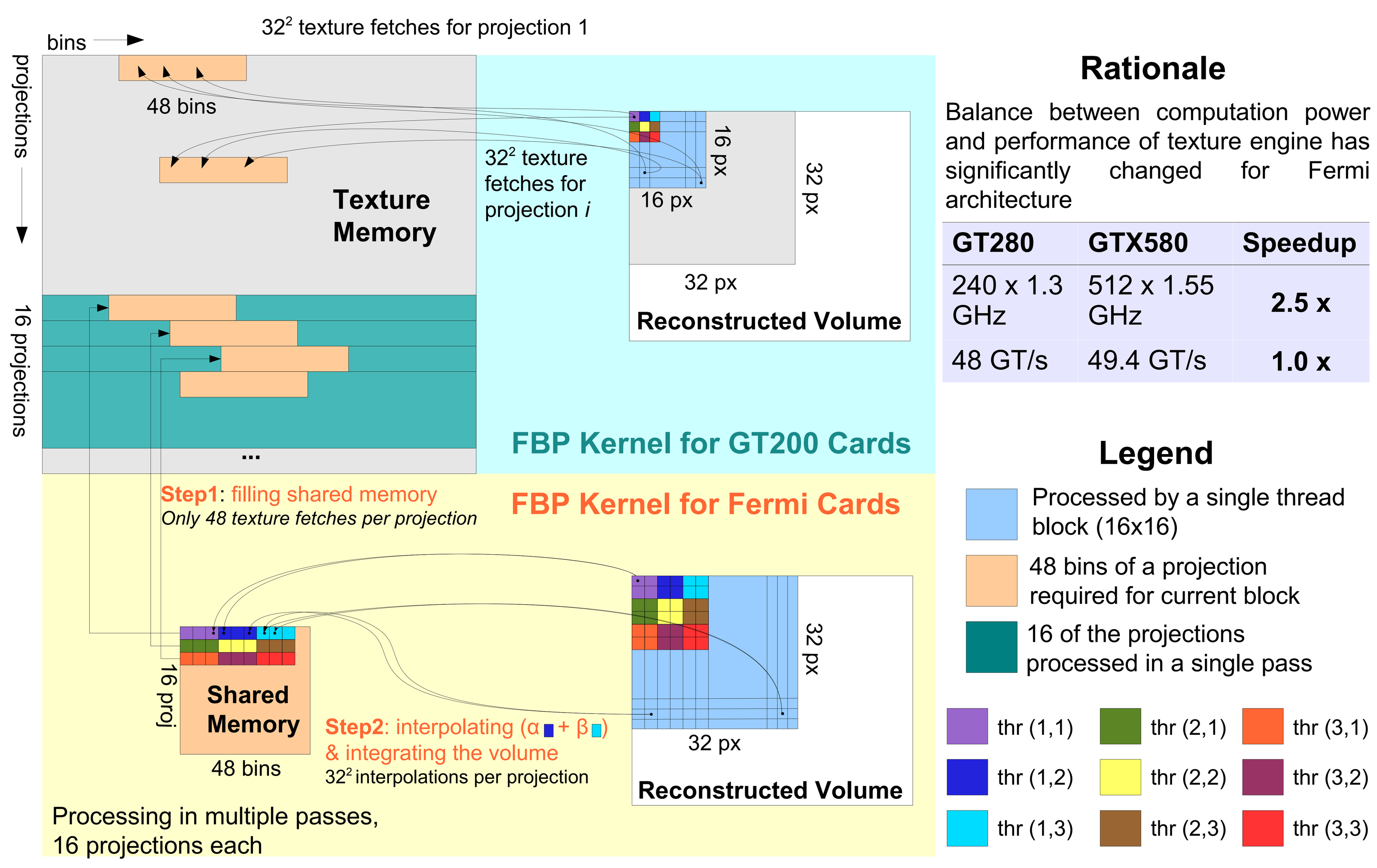
Data Transfer
- Pinned (unswappable) memory buffers are used to exchange the data with GPU
- The slice is split in blocks and the data transfer of next block is interleaved with computation of current one
- The blocks are still big enough to fully utilize GPU multiprocessors

Filtering
- Data is padded to a size equal to the power of 2
- Batched processing is used to filter slices
- Two real convolutions are computed using a single complex cuFFT transform

Back Projection
- On GT200 cards, texture engine is used to accelerate random access and linear interpolation
- For Fermi cards, a shared memory is used to reduce number of texture fetches
- To hide memory latencies caused by low occupancy due to high register usage, Fermi kernel processes four pixels per thread



Fermi-specific Optimizations



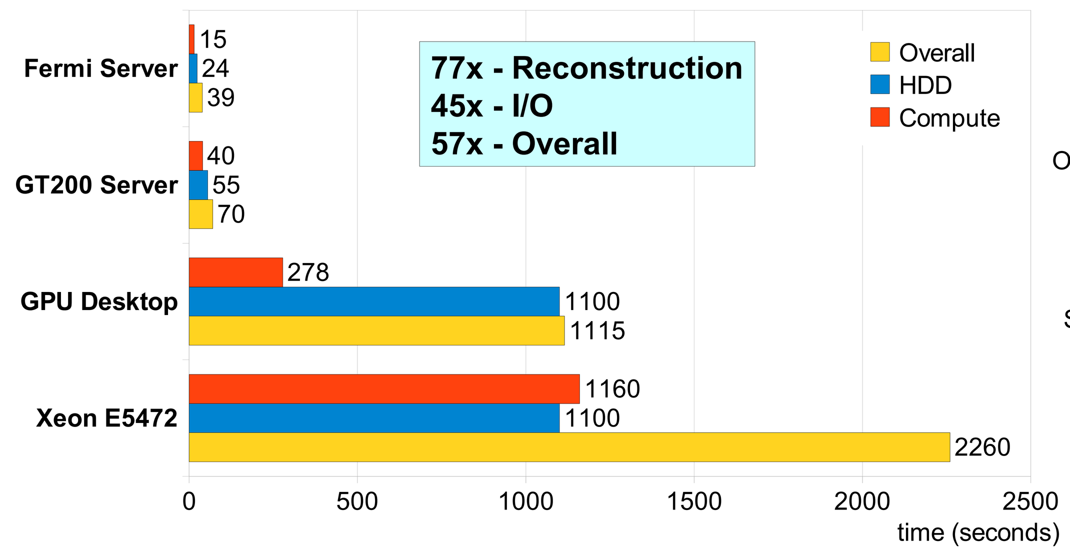
Rationale

Balance between computation power and performance of texture engine has significantly changed for Fermi architecture

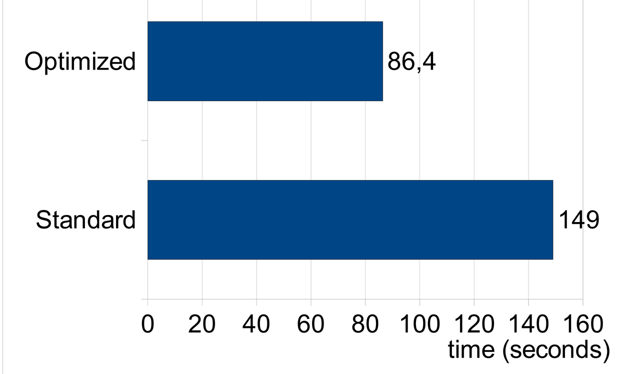
GT280	GTX580	Speedup
240 x 1.3 GHz	512 x 1.55 GHz	2.5 x
48 GT/s	49.4 GT/s	1.0 x

Performance Evaluation

	Xeon Server	GPU Desktop	GT200 Server	Fermi Server
Type of Computation	CPU / Xeon E5472 8 core, 3 GHz	GeForce GTX 280 1 core	2 x GTX295 + 2 x GTX280 6 cores	6 x GTX580 6 cores
CPU	2 x Xeon E5472	Core2 E6300	2 x Xeon E5540	2 x Xeon E5540
Memory	16GB DDR3	4GB DDR2	96GB DDR3	96GB DDR3
HDD	WDC5000AACS	WDC5000AACS	2 x Intel X25-E / Raid-0	4 x Crucial C300 / Raid-0
Software	OpenSuSe 11.4, CUDA 3.2, Intel MKL 10.2.1, gcc4.4 -O3 -march=nocona -mfpmath=sse			



Comparison of Fermi Kernels



Using single GTX580

S. Chilingaryan, A. Mirone, A. Hammersley, C. Ferrero, L. Helfen, A. Kopmann, T. dos Santos Rolo, P. Vagovic
A GPU-Based Architecture for Real-Time Data Assessment at Synchrotron Experiments
Published at IEEE Transactions on Nuclear Science, Volume 58, Issue 4, pp. 1447-1455, 2011