OpenPOWER:

First Results for Scientific Applications

2015 July 1st | Paul F Baumeister, Thorsten Hater, Dirk Pleiter | JSC

Outline

- OpenPOWER
    - HPC roadmap
    - POWER architecture
- Applications
    - Electronic structure
    - Electrodynamics
    - Image registration
- Summary

OpenPOWER

Who is involved?

and many others

Roadmap forsees a tighter integration of GPU and CPU:
faster and coherent attachment via NVLINK

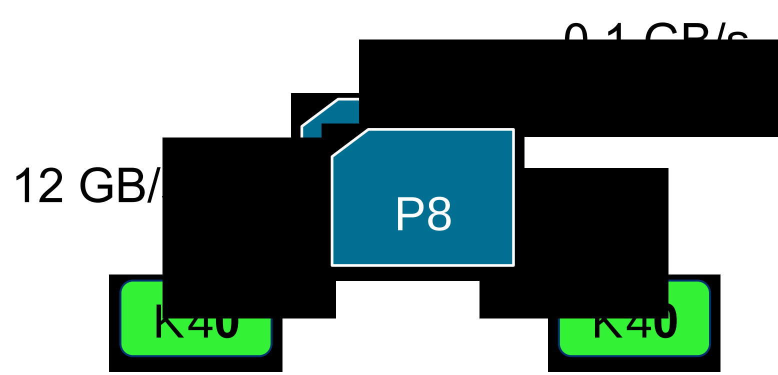
- usual attachment CPU-GPU 16x PCIe 3.0 ~16 GB/s
- starting from Pascal GPU-GPU NVlink ~80 GB/s
- starting from Volta NVlink 2.0 ~200 GB/s

OpenPOWER test system at JSC

4 nodes connected via 1Gb Ethernet

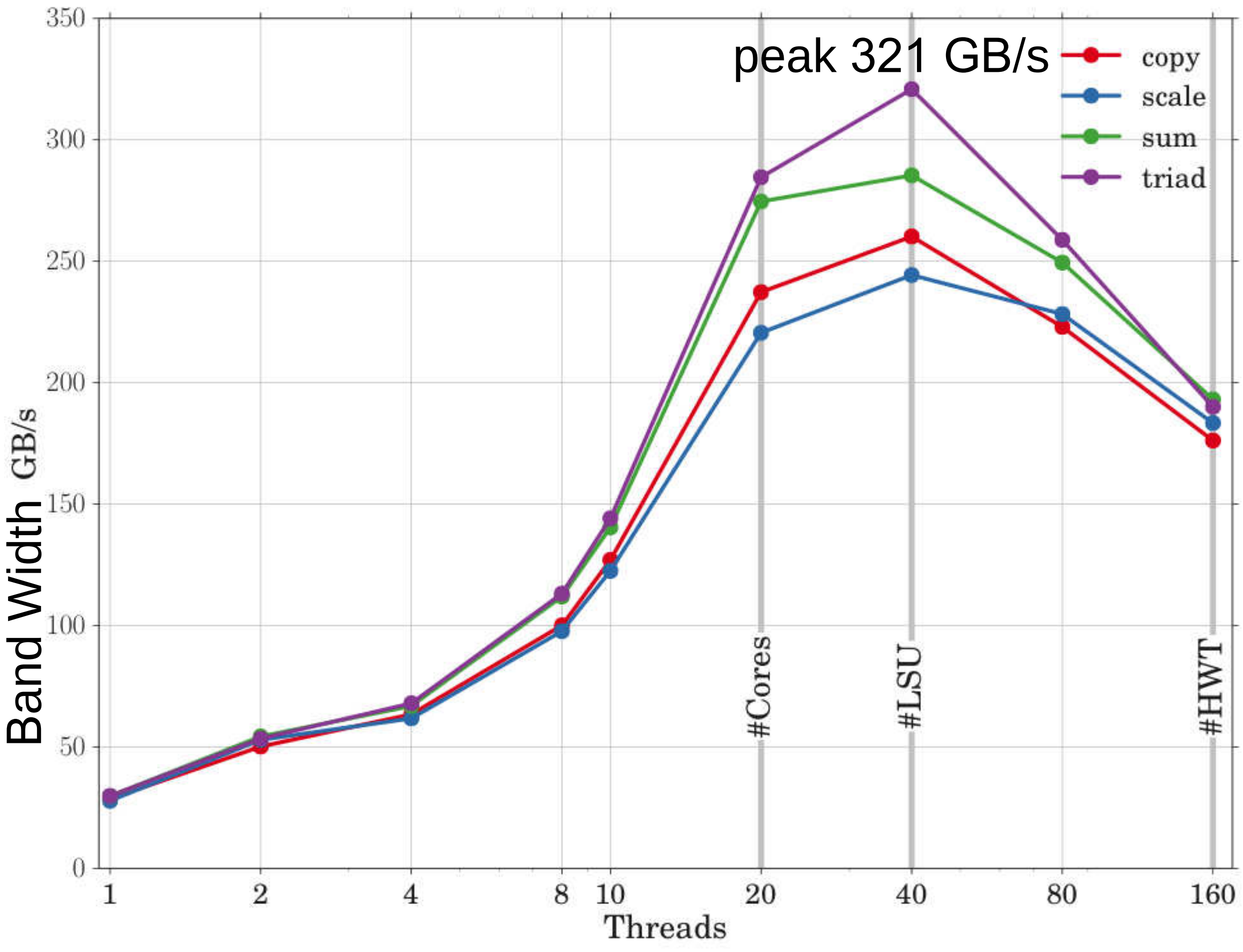
each node	= dual socket POWER8 824 47L with 256 GByte DDR3 + 2! K40 with 12 GByte GDDR5 each
each K40	= 15 SMX ! 192 CUDA cores @745--875 MHz
each P8 socket	= 10 cores @2.0--3.7 GHz
each P8 core	= 64 kB L1d$, 32 kB L1i$, 512 kB L2$, 8 MB L3$

OS: ubuntu 14.10

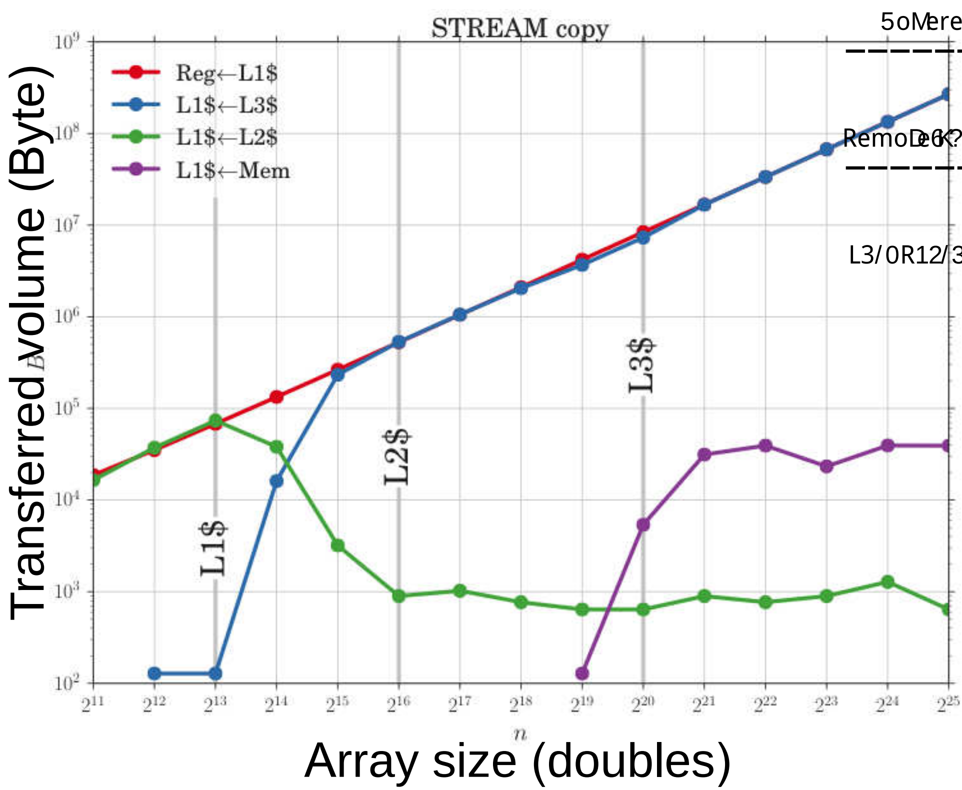


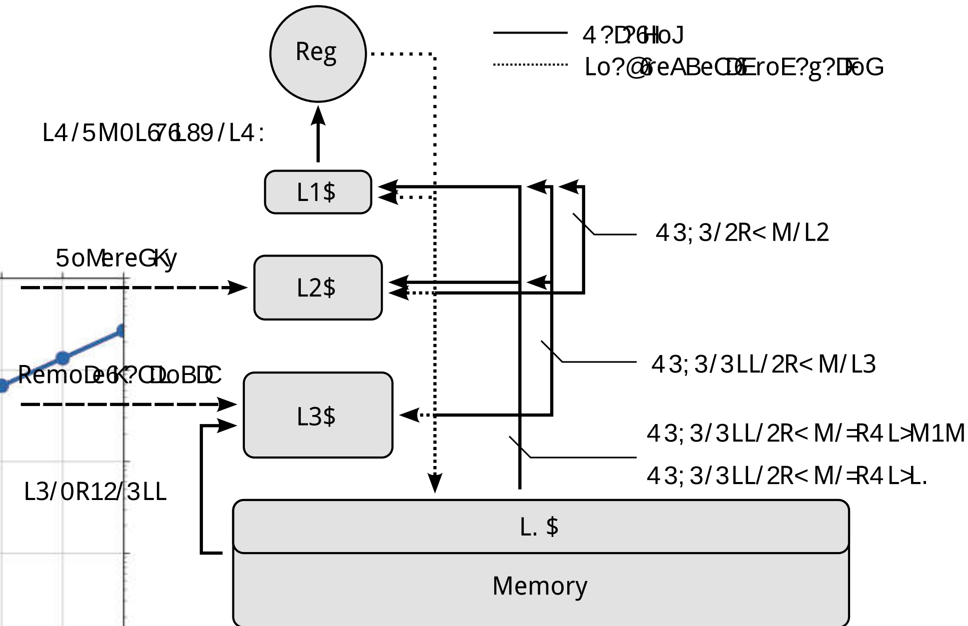
STREAM Micro-Benchmark on POWER8

Median values of the measured effective memory BandWidth



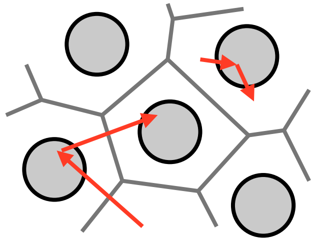
POWER8 memory model analyzed by hardware counters





Most memory processes can be monitored by the corresponding hardware counters accessible by PAPI 5.3.2 natives on POWER8

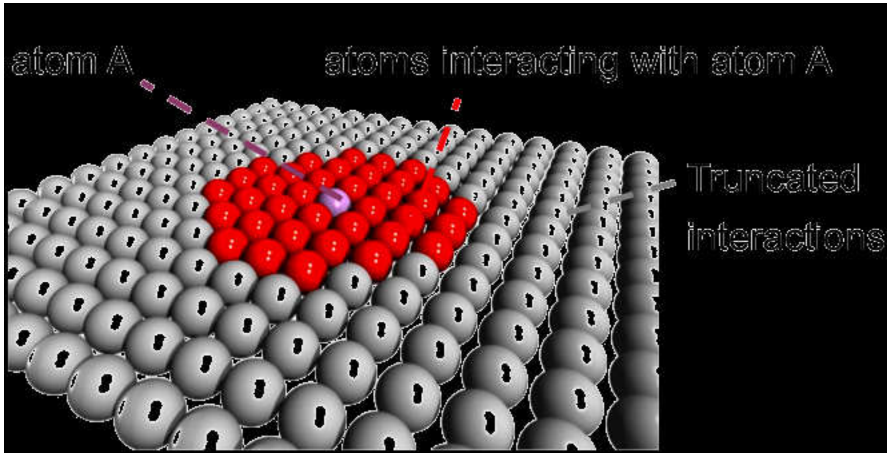
Application 1: KKRnano

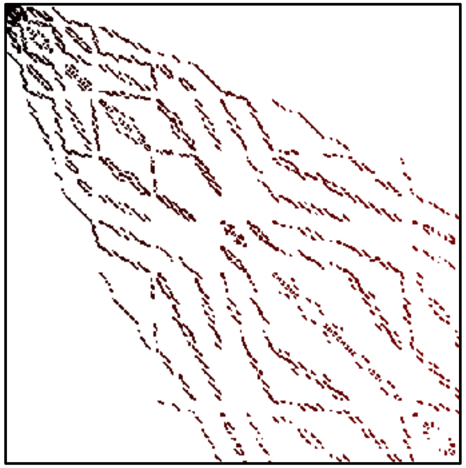


- •! Material science simulation based on density functional theory
- •! Solves the Helmholtz equation with an iterative QMR solver
- •! Block-sparse matrix times block vector

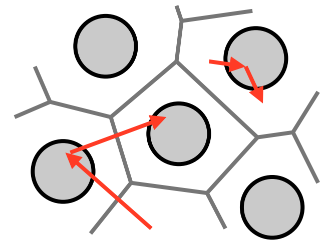
  each block is a complex 16! 16 matrix

  arithmetic intensity ~ 4.0 Flop/Byte (good for CPU, low for GPU)





Porting KKRnano to GPUs



- Solver calls sparse matrix action $O(10^3)$ times

- Porting the full QMR solver to the GPU to avoid memory transfers

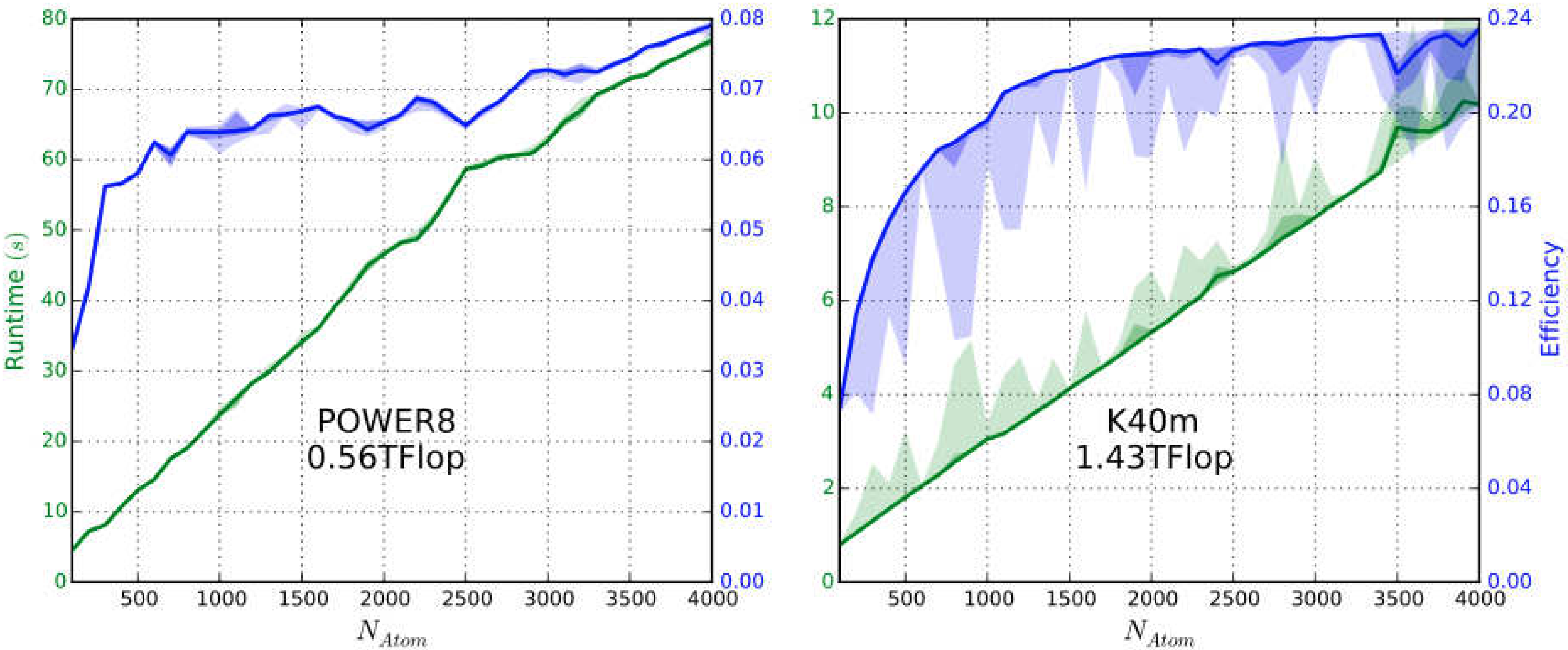
- Block-matrix and Block-vector operations
  expressed in cuBLAS and cuSPARSE calls

- Data transfer CPU-GPU is ~ 80 MiByte ! N_{atoms} / N_{MPI}
  needed only once per solver invocation

JÜLICH
FORSCHUNGSZENTRUM

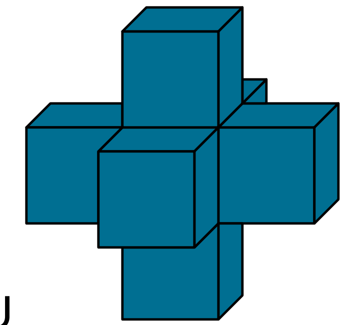
Results on KKRnano

POWER8 (20 cores)	320 GByte/s	560 GFlop/s
NVIDIA K40m	288 GByte/s	1430 GFlop/s

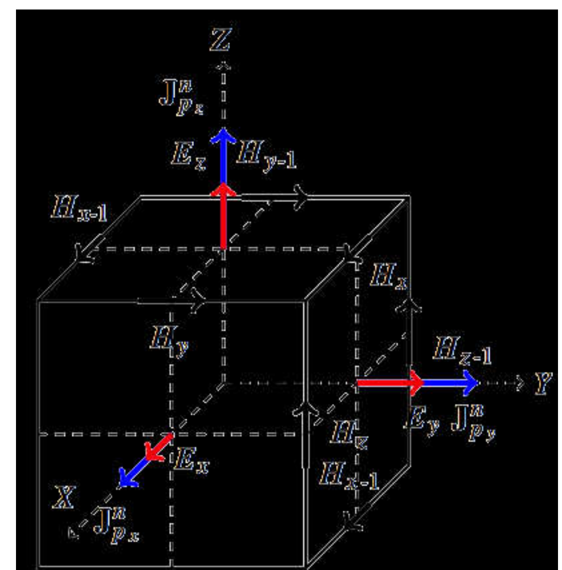


Application 2: B-CALM

- Solves Maxwell equations in dispersive media
- Finite Difference in the Time Domain

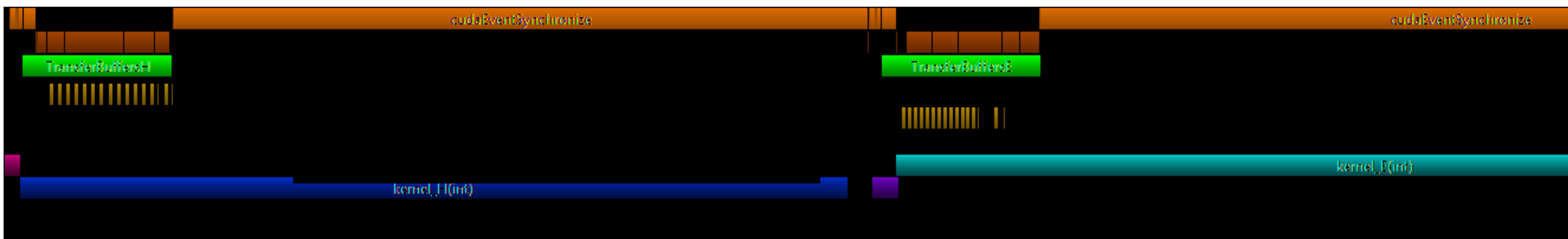


- 1D or 2D spatial domain decomposition for multi-GPU
- Memory BandWidth limited due to low order stencil



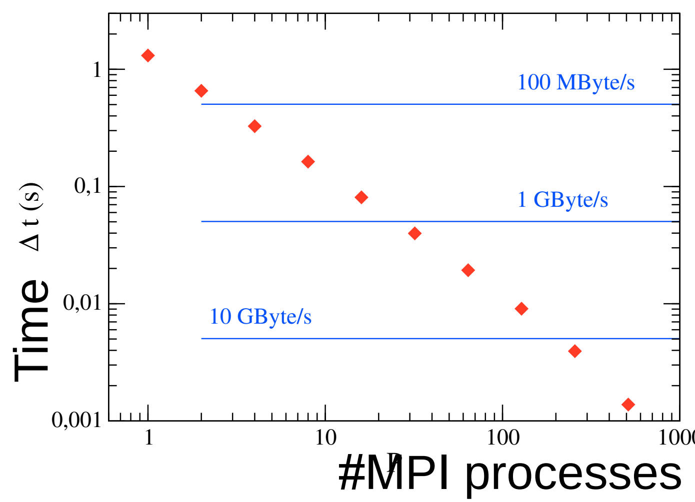
B-CALM: communication hiding

- Overlap MPI communication and bulk-computation
- GPU-GPU transfer time determines the minimum bulk work per node



nvvp trace

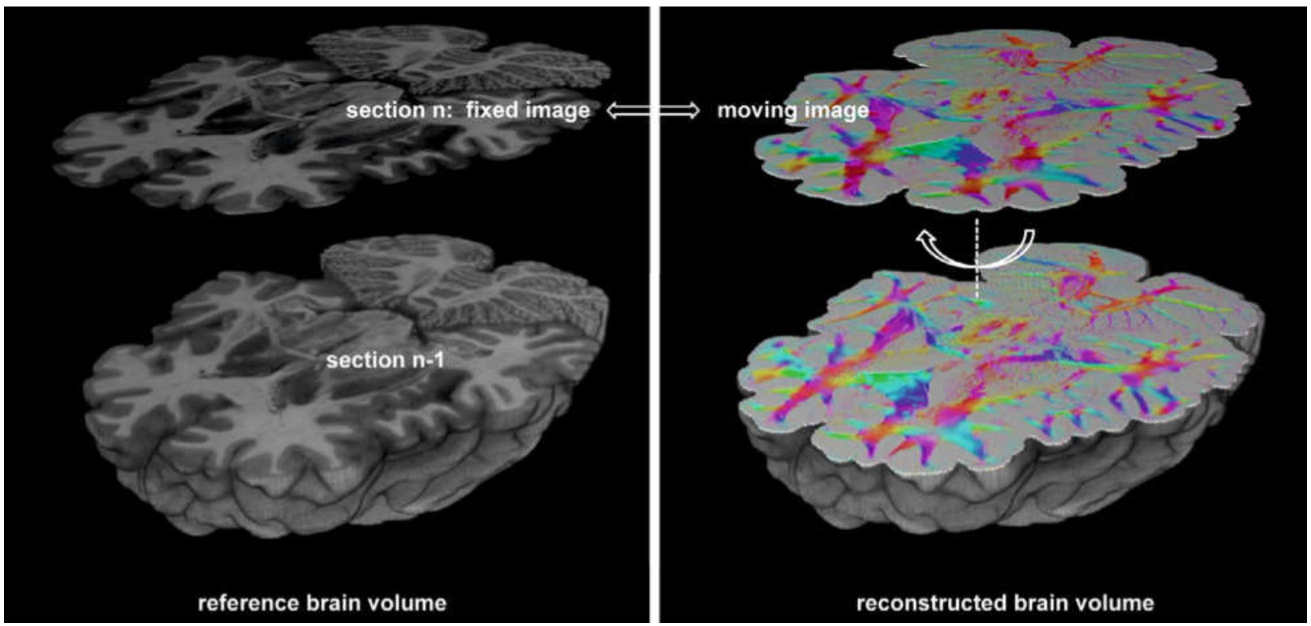
Simple performance model
(1D domain decomposition)



Application 3: juBrain

Reconstruction of neural networks in a postmortem brain

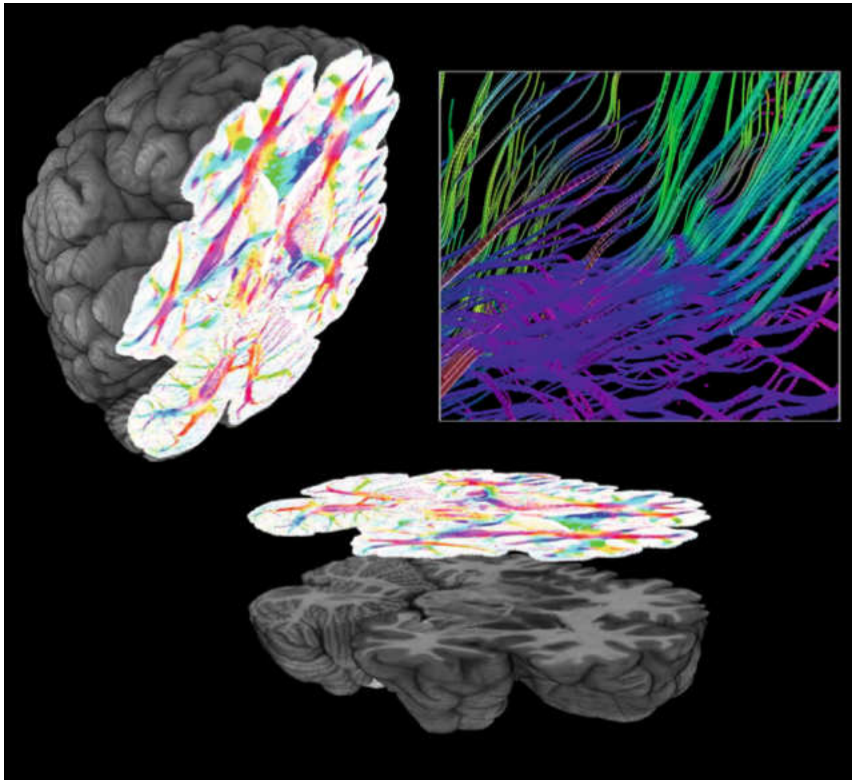
Polarized Light Imaging resolves axon orientation



Images much larger than GPU memory ! 2D domain decomposition

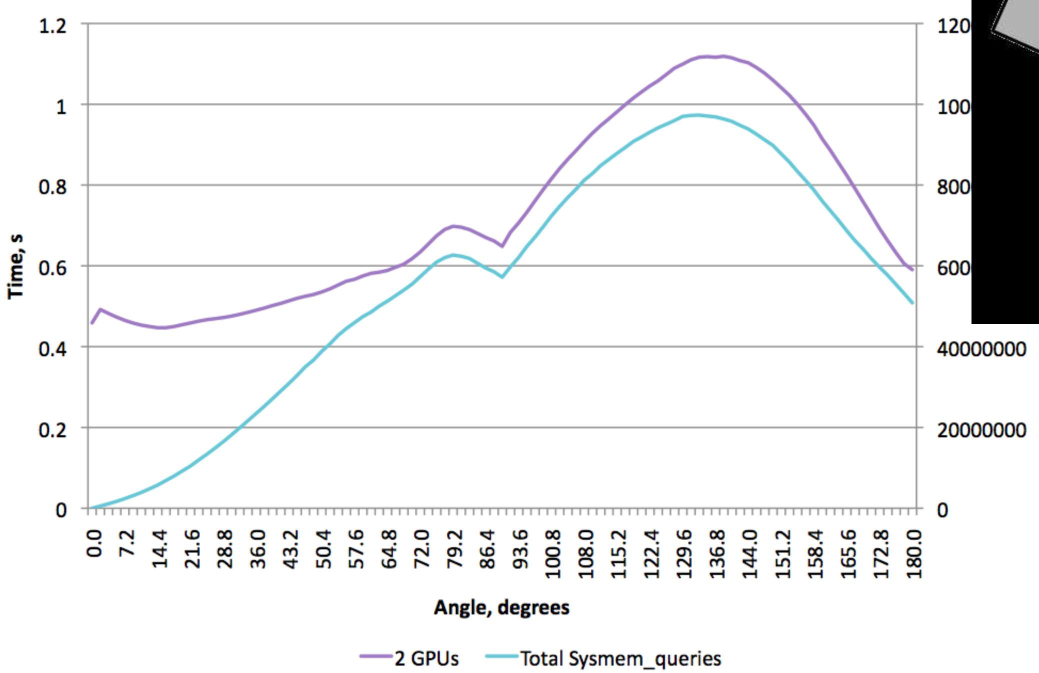
juBrain: streaming

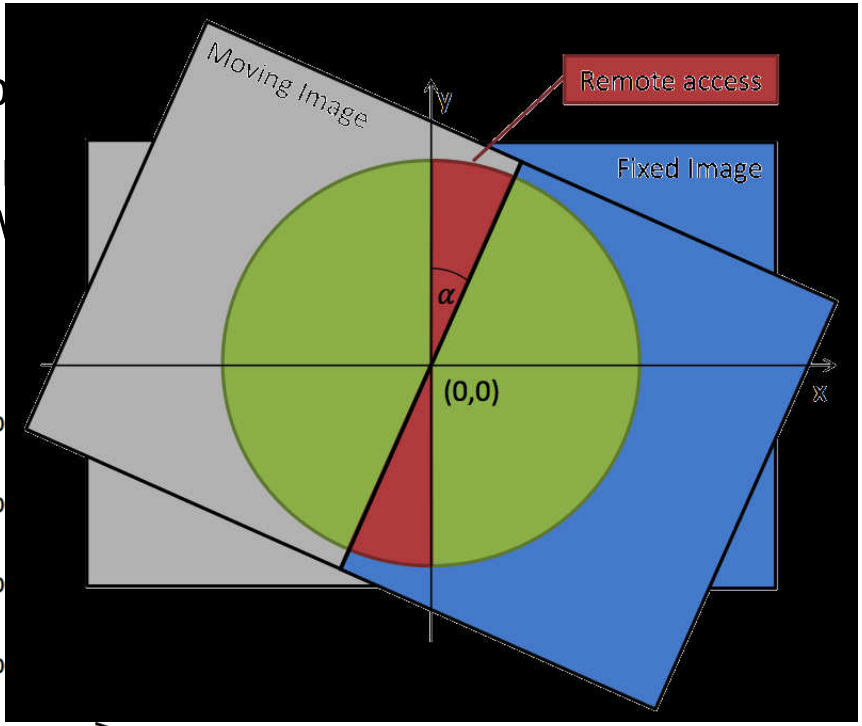
Reconstruct brain volume
with fiber trajectory from slices


reconstructed brain volume with subsequent fiber tractography

juBrain: Link speed matters

- Iterative optimizer invokes image interp
- Metric performance depends on rotatio angle and host-to-device transfer BW
- Runtime increases up to 2.2x





Summary

Three different cases of scientific applications:

- CPU-GPU transfers are negligible compared to compute
- CPU-GPU link speed determines minimum local work
- Tighter integration of CPU and GPU necessary

ToDo

- understand CPU performance of KKRnano

Acknowledgment

Andrew Adinets (NVIDIA Application Lab), Jiri Kraus (NVIDIA)

References

www.nvidia.com/

A. Adinetz, J. Kraus, et al., EuroPar2013: LNCS Vol. 8374 (2014) pp208-217

A. V. Adinetz , Paul F. Baumeister, et al., PMBS14, New Orleans, (Nov2014)

Thiess, A. and Zeller, R. et al., PhysRevB 85 (Jun2012) 235103