High-performance Data Analytics (HPDA) at the MPCDF

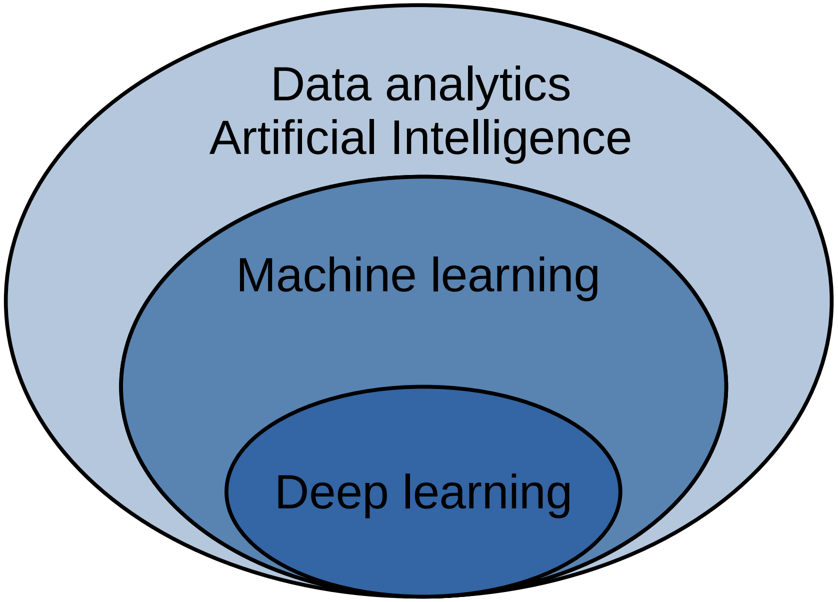


Image adapted from: arXiv:1903.11314

Markus Rampp (markus.rampp@mpcdf.mpg.de)

Andreas Marek (andreas.marek@mpcdf.mpg.de)

Max Planck Computing and Data Facility (MPCDF)

DV-Treffen der MPG, Göttingen, Sep 17-19, 2019

Overview

HPDA services and activities at the MPCDF, with some illustrative examples

MPCDF team lead (cross-division project teams): Andreas Marek

- **hosting and consulting for systems** (hardware design, BAR, procurement, …)
  - ML-Cluster Talos (MPI for Polymer Research, MPI for Iron Research, FHI)
  - Tiered large-scale storage solution for MPI Neurobiology
- **provisioning of platform-optimized software stack**
  - available to all users on MPCDF HPC systems + institute clusters
- **project-specific application support and development**
  - MPCDF is member of BiGmax, the Max Planck research network on big-data-driven materials science, www.bigmax.mpg.de
  - MPCDF supports several MP Institutes in various HPDA projects
- **training, consulting and knowledge transfer** (similar to HPC)

Overview

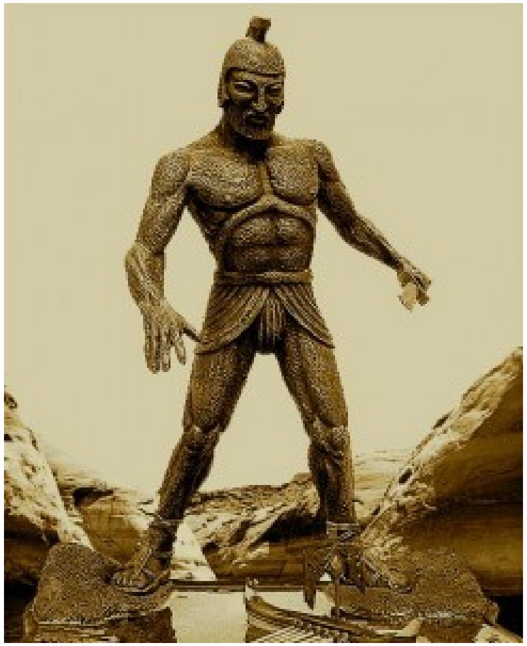
HPDA services and activities at the MPCDF, with some illustrative examples

MPCDF team lead (cross-division project teams): Andreas Marek

- **hosting and consulting for systems** (hardware design, BAR, procurement, …)
  - ML-Cluster Talos (MPI for Polymer Research, MPI for Iron Research, FHI)
  - Tiered large-scale storage solution for MPI Neurobiology
- provisioning of platform-optimized software stack
  - available to all users on MPCDF HPC systems + institute clusters
- project-specific application support and development
  - MPCDF is member of BiGmax, the Max Planck research network on big-data-driven materials science, www.bigmax.mpg.de
  - MPCDF supports several MP Institutes in various HPDA projects
- training, consulting and knowledge transfer (similar to HPC)

Talos: “A Compute Cluster for *Machine Learning* Applications”

- approved and supported by the BAR 2018
- “co-”designed and operated by the MPCDF
- operational since Feb 2019



The world’s first robot: ***Talos***
(http://www.wondersandmarvels.com)

Owners:

- Fritz-Haber Institute (Dept. M. Scheffler)
- MPI of Polymer Research (Dept. K. Kremer)
- MPI for Iron Research (Dept. J. Neugebauer, Dept. D. Raabe)

ML cluster Talos

Hardware

- Compute: 85 GPU-accelerated nodes, Interconnect: Intel OmniPath 100Gbit/s, GPFS
- node characteristics
  - 2x 20-core CPUs Intel Xeon 6138 Skylake @ 2.0 GHz , 192 GB RAM
  - 2 GPUs Nvidia Tesla V100 (2x 32 GB)
  - 1TB SSD (OS + ca. 700 GB)

Design considerations

- Talos was *not* specifically designed for deep learning only:
  - neural networks (NN), generative adversarial networks (GAN) → GPU
  - large-scale applications of SISSO → CPU
  - data normalization, prototyping etc. → CPU
- we have *scalable* machine learning in mind
  - some standard deep learning applications *might* run better on (expensive) multi-GPU complexes like Nvidia DGX, Nvswitch
  - scalable ML is discussed everywhere (meanwhile)

Connectomics of the Mouse Brain

Robotic imaging of the entire mouse brain (~1 cm^3 , 100 PB) using Serial Block Face Scanning Microscopy @MPI for Neurobiology (Dept. W. Denk)

Scientific goal: spatial reconstruction of the connectome

Role of MPCDF: Design, deployment and operation of the storage infrastructure

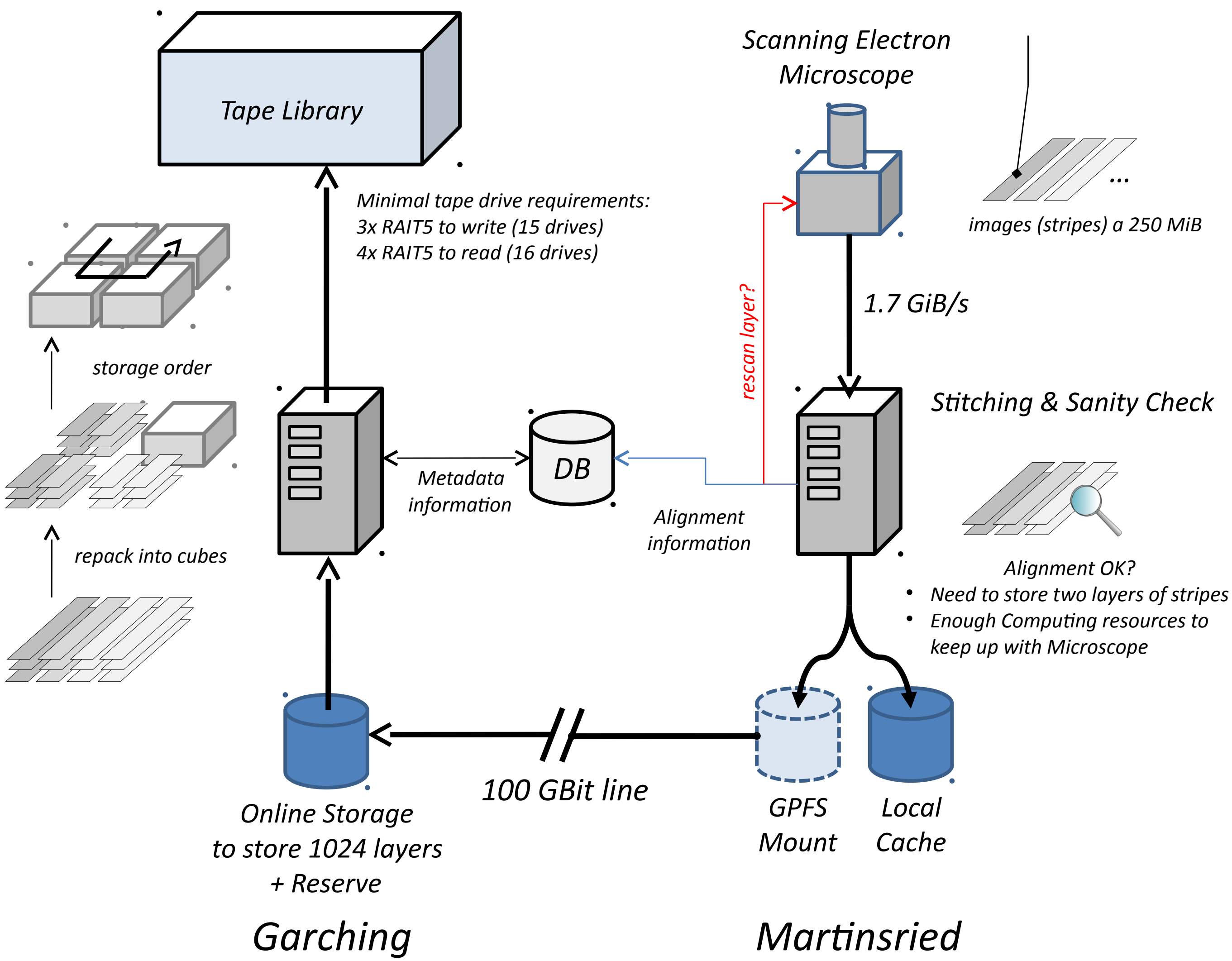
Technical specifications:

- Microscope output bandwidth 1.7 GiB/s (91 beams at 20 MHz, 8 bit pixel)
- Images are stripes of 256x1M pixels, so about 250 MiB each
- 10 - 20 minutes scan time per layer
- Intention of 30s (typical) - 3min (max) to cut down to next layer
- About 1 TiB image data in 4000 images per layer
- O(100 PB) per specimen (or year)
- Human analysis of 3D volume at mm/hour traversal speed

→ a multi-tiered storage solution based on HPSS

→ crucial: data-locality optimization for future analysis (cost vs. latency vs. bandwidth)

Status: BAR proposal (2016), deployment started



Overview

HPDA services and activities at the MPCDF, with some illustrative examples

MPCDF team lead (cross-division project teams): Andreas Marek

- hosting and consulting for systems (hardware design, BAR, procurement, …)
  - ML-Cluster Talos (MPI for Polymer Research, MPI for Iron Research, FHI)
  - Tiered large-scale storage solution for MPI Neurobiology
- **provisioning of platform-optimized software stack**
  - available to all users on MPCDF HPC systems + institute clusters
- project-specific application support and development
  - MPCDF is member of BiGmax, the Max Planck research network on big-data-driven materials science, www.bigmax.mpg.de
  - MPCDF supports several MP Institutes in various HPDA projects
- training, consulting and knowledge transfer (similar to HPC)

Welcome to the jungle ...

theano

TensorFlow

RAPIDS

HOROVOD

scikit learn

APACHE Spark™

Caffe

Software for machine learning

Software for machine learning (ML) provided by the MPCDF

low-level ML libraries:

- mkl-dnn, mlsl (CPU)
- cudnn, nccl, tensorflowRT (GPU) + soon: Nvidia rapids, ...

high-level ML frameworks:

- scikit-learn, tensorflow, horovod, apache spark, keras (CPU)
- tensorflow, horovod, keras (GPU)
- Hyperopt:Distributed Asynchronous Hyper-parameter Optimization (CPU)

Documentation (including comprehensive “howtos”) at:

- www.mpcdf.mpg.de/services/computing/software/data-analytics/machine-learning-software

Access:

- MPG’s HPC systems Cobra and Draco
- Institute clusters (on request), e.g. Talos

Software for machine learning (ML) provided by the MPCDF

low-level ML libraries:

- mkl-dnn, mlsl (CPU)
- cudnn, nccl, tensorflowRT (GPU) + soon: Nvidia rapids, ...

platform-optimized ML software stack, e.g.

- TensorFlow with AVX512 CPU + GPU optimization
- Pytorch with AVX512 CPU + GPU optimization
- TensorFlow with native MPI support by Horovod (developed at Uber)
- CPU SW built with: MKL, DAAL (Intel)
- GPU SW built with: cuBLAS, cuDNN, NCCL, tensorflowRT,… (Nvidia)

high-level ML fr

- scikit-learn, t
- tensorflow, h
- Hyperopt:Di

Documentation (including comprehensive "howtos") at:

- www.mpcdf.mpg.de/services/computing/software/data-analytics/machine-learning-software

Access:

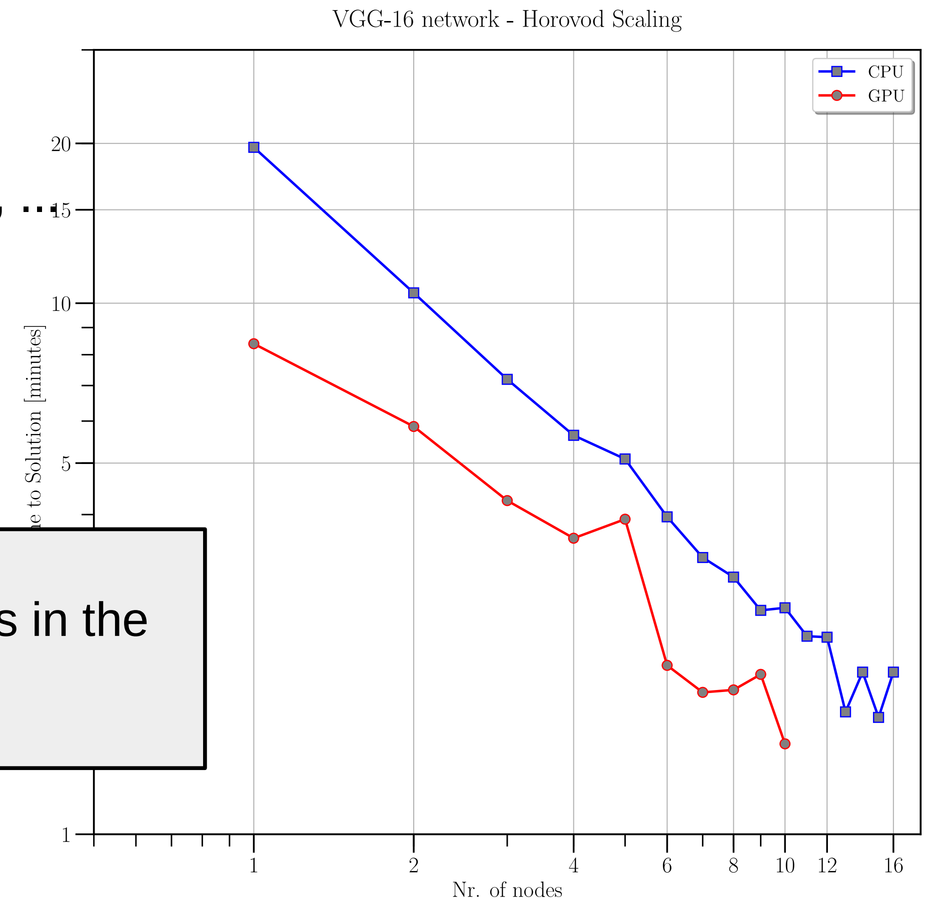
- MPG's HPC systems Cobra and Draco
- Institute clusters (on request), e.g. Talos

Distributed deep learning

Distributed training of a CNN with Keras/Tensorflow/Horovod (based on MPI communication)

The deep-learning community (at Facebook, Uber !) has discovered/established HPC concepts: optimized collective MPI communication, ...

(Ben-Nun & Hoefler, arXiv:1802.09941)



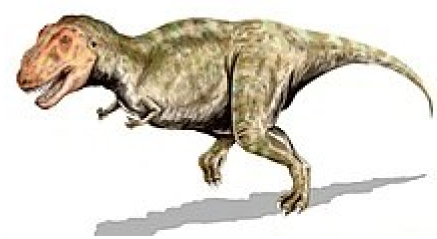
Application-specific benchmarking and optimization in collaboration with partners in the MPG

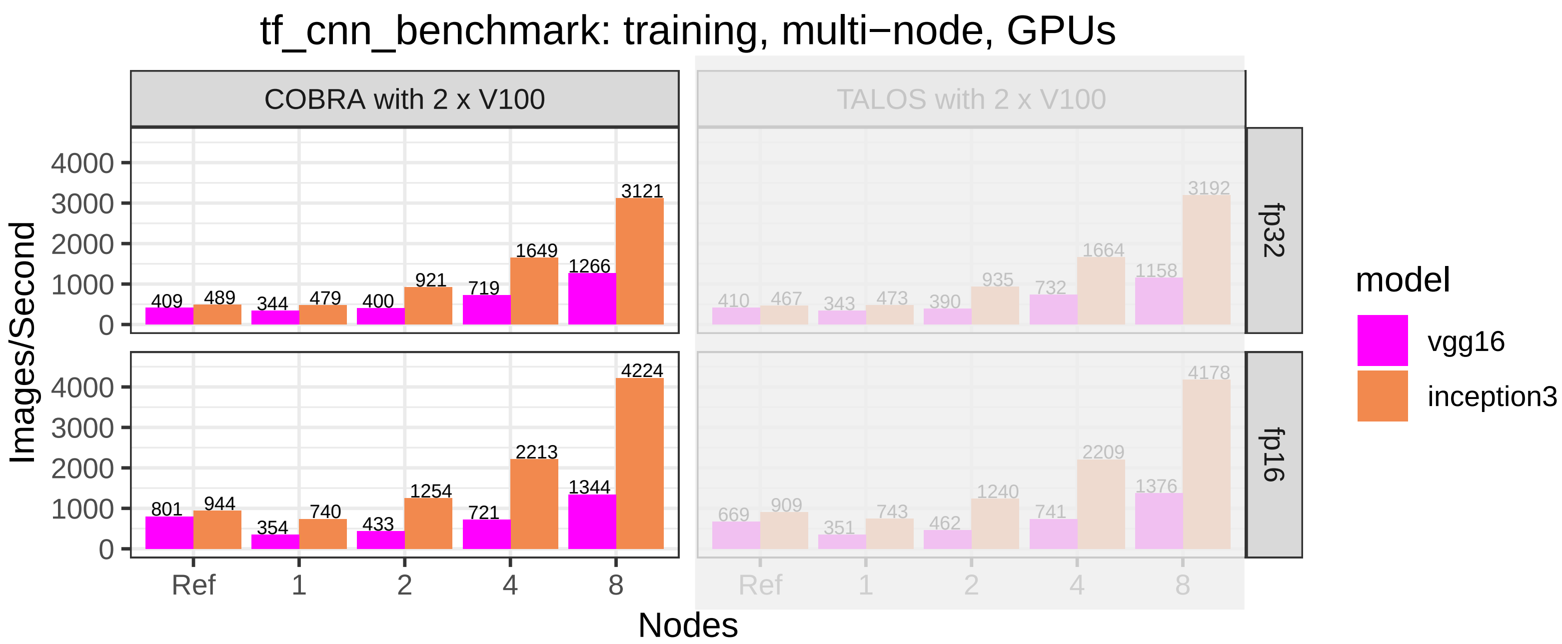
notes:

- Horovod default strategy: the mini-batch size per process is kept constant
- the scaling study is performed on a pre-trained network → GPU not saturated
- benchmarks were performed on MPCDF Cobra (HW/SW setup similar to Talos)

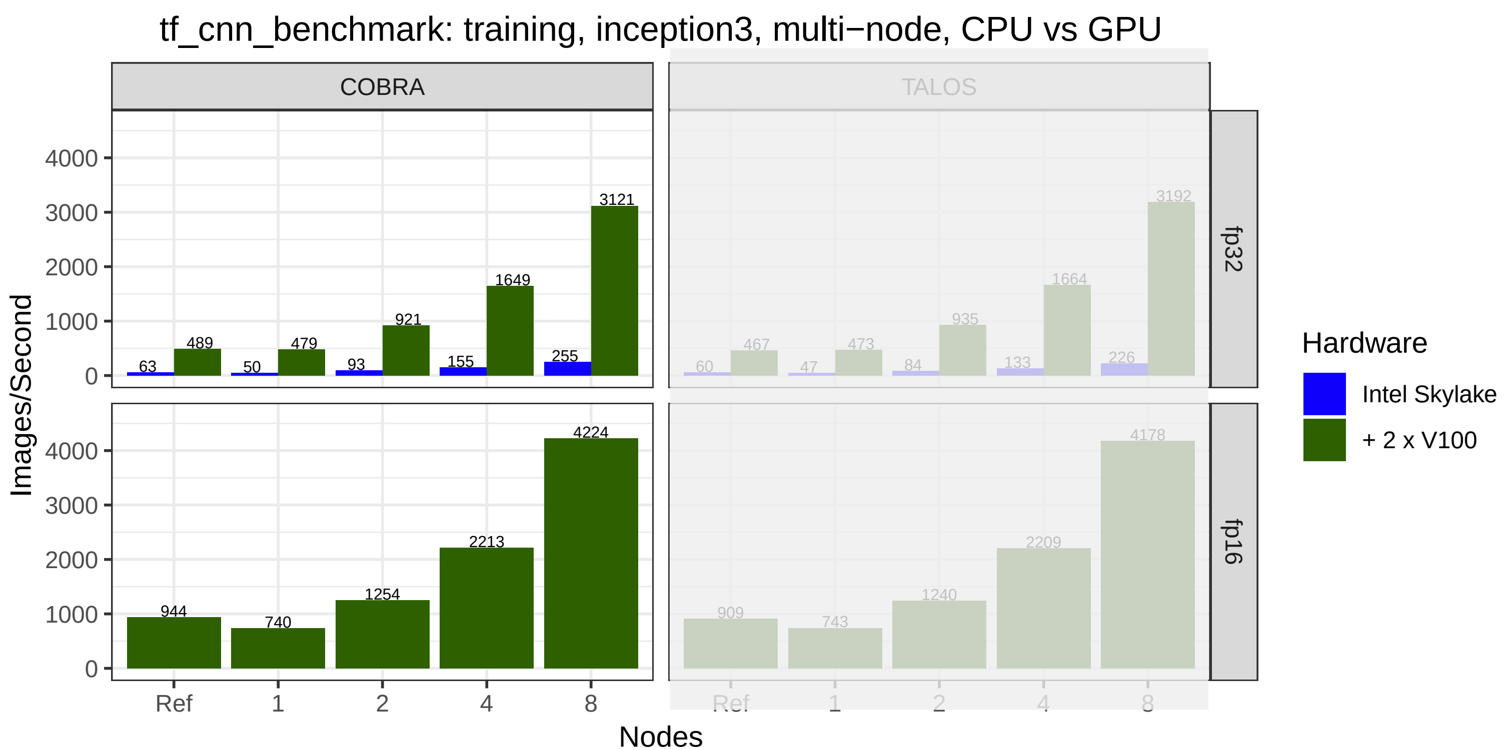
Benchmarking TensorFlow/Horovod

https://github.com/tensorflow/benchmarks





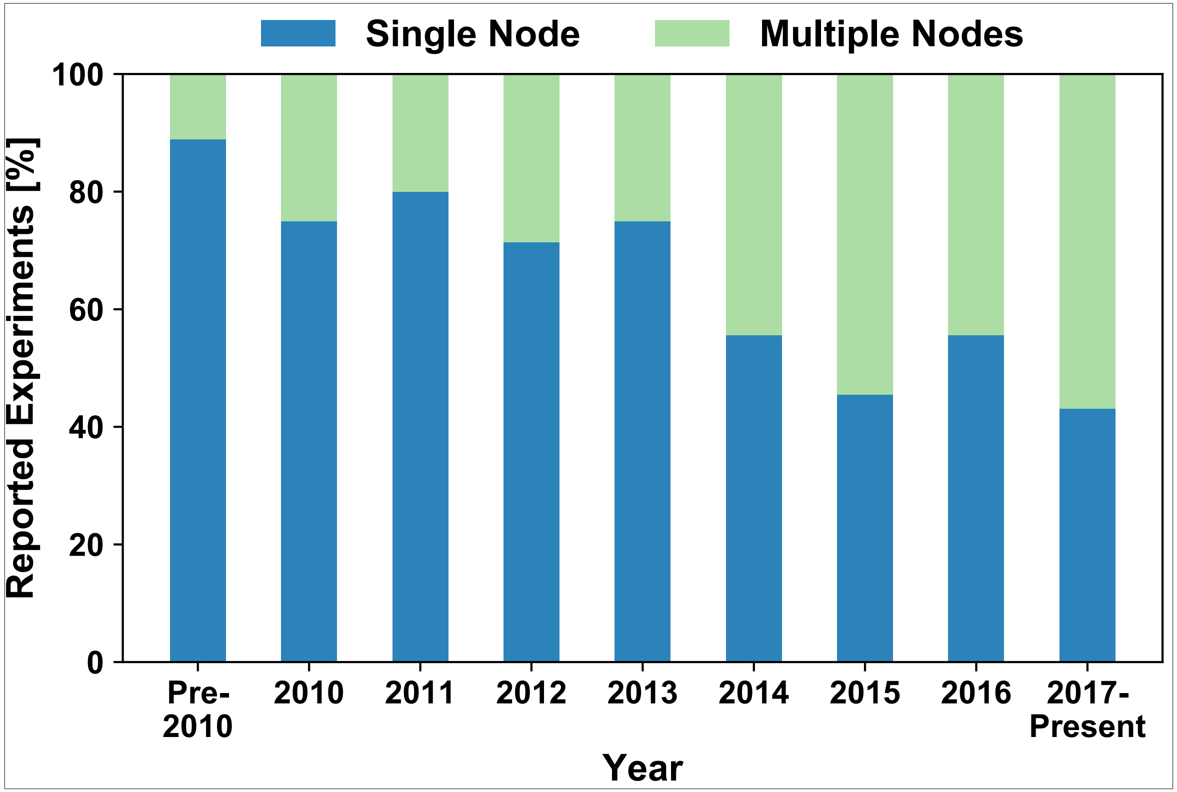
→ scaling across nodes works efficiently



→ scaling across nodes works efficiently

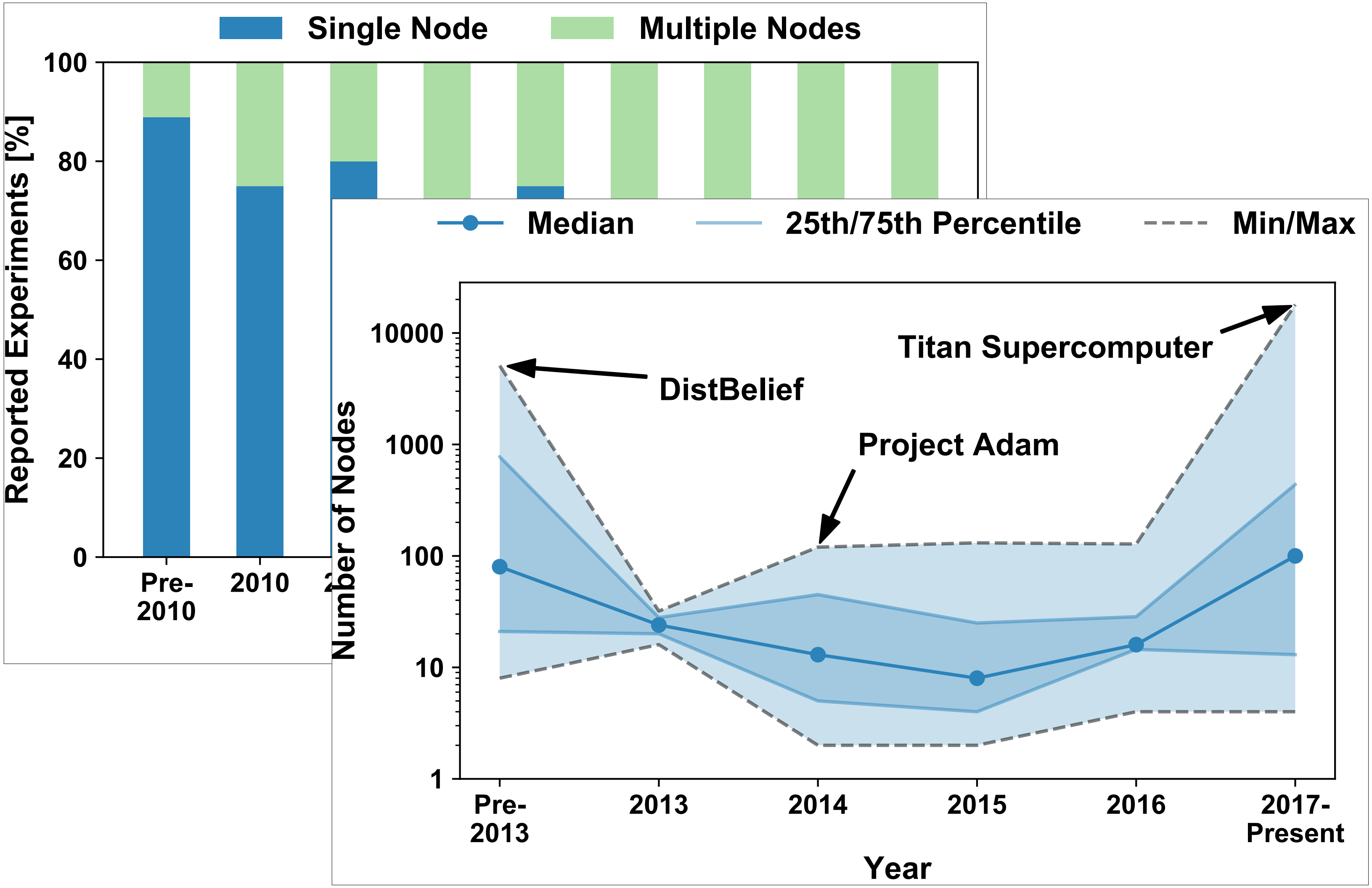
→ GPUs provide significant speedup (wrt. CPU-only)

Relevance of distributed ANN computation



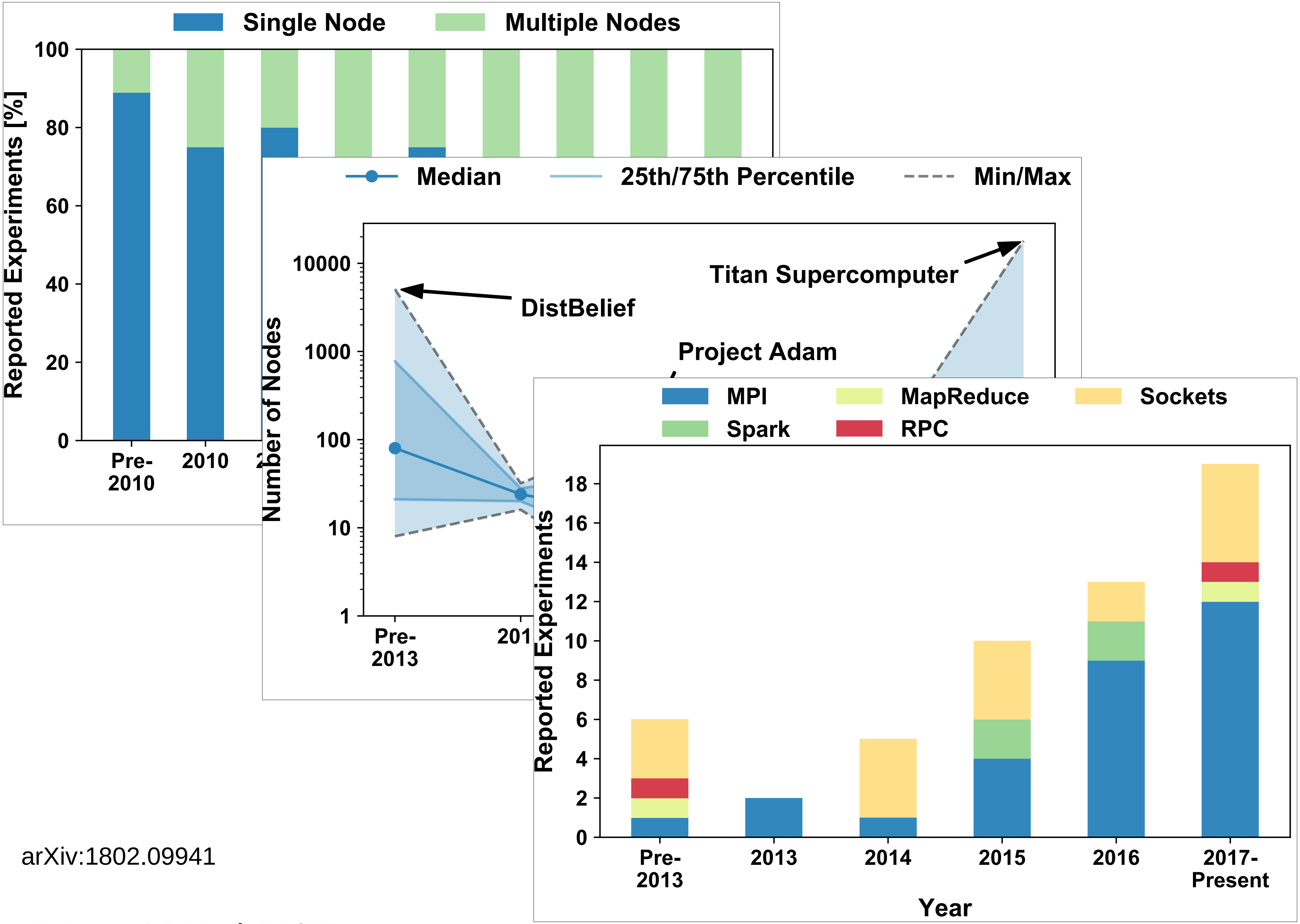
arXiv:1802.09941

Relevance of distributed ANN computation



arXiv:1802.09941

Relevance of distributed ANN computation



arXiv:1802.09941

M. Rampp & A. Marek, MPCDF

Relevance of distributed ANN computation

Benchmarking ANN: what is the right metric?

→ time to solution ! = time to reach a specified accuracy (validation loss)

→ commonly used: images/second (= throughput)

→ opens up many opportunities to cheat (ourselves)

→ watch out !

https://htor.inf.ethz.ch/blog/index.php/2018/11/08/twelve-ways-to-fool-the-masses-when-reporting-performance-of-deep-learning-workloads/

Overview

HPDA services and activities at the MPCDF, with some illustrative examples

MPCDF team lead (cross-division project teams): Andreas Marek

- hosting and consulting for systems (hardware design, BAR, procurement, …)
  - ML-Cluster Talos (MPI for Polymer Research, MPI for Iron Research, FHI)
  - Tiered large-scale storage solution for MPI Neurobiology
- provisioning of platform-optimized software stack
  - available to all users on MPCDF HPC systems + institute clusters
- **project-specific application support and development**
  - MPCDF is member of BiGmax, the Max Planck research network on big-data-driven materials science, www.bigmax.mpg.de
  - MPCDF supports several MP Institutes in various HPDA projects
- **training, consulting and knowledge transfer** (similar to HPC)

MPG research network BiGmax

Members:

- Fritz Haber Institute of the Max Planck Society, Berlin
- Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg
- Max Planck Institute for Iron Research, Düsseldorf
- Max Planck Institute for the Physics of Complex Systems, Dresden
- Max Planck Institute for Informatics, Saarbrücken
- Max Planck Institute for Structure and Dynamics of Matter, Hamburg
- Max Planck Institute for Intelligent Systems, Stuttgart
- Max Planck Institute for Polymer Research, Mainz
- Max Planck Institute of Colloids and Interfaces, Potsdam
- Max Planck Computing and Data Facility, Garching
- Humboldt University, Berlin

www.bigmax.mpg.de

Coordinators: P. Benner, M. Scheffler

Topics:

1. Structure and plasticity of materials
2. Data diagnostics in imaging
3. Discovering interpretable patterns, correlations, and causality
4. Learning thermodynamic properties of materials
5. Materials Encyclopedia (incl. metadata for experimental samples and methods)

MPG research network BiGmax

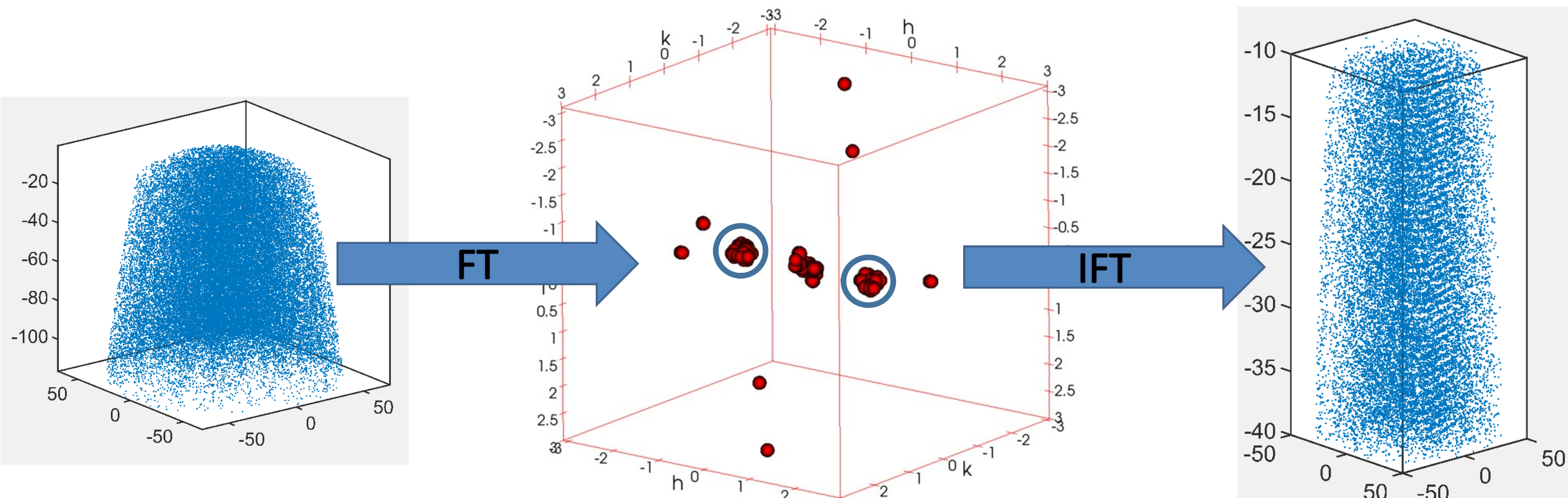
MPCDF Max Planck Computing & Data Facility

BiGmax MAX PLANCK RESEARCH NETWORK on big-data-driven materials science

Development of efficient computational and visualization methods
PIs: Hermann Lederer, Markus Rampp (MPCDF)

Example project (MPCDF together with MPI Iron Research, Dept. D. Raabe):

Development of a Python based and GPU-accelerated analysis workflow for identifying crystalline sub-volumes of large Atom Probe Tomography (APT) samples (millions to billions of atoms). *Accurate direct Fourier summation* is used to transform between real space of atom coordinates and reciprocal space.



M. Rampp & A. Marek, MPCDF

MPG research network BiGmax

1) APT scattering maps (forward transform from atom positions to reciprocal space)

- requires ***direct summation, O(N²),*** for 10^9 atoms in 3 dimensions
- accomplished by adopting the PyNX package (GPU acceleration)

$$A(\mathbf{s})=\sum_i f_i(\mathbf{s})e^{2\pi i\mathbf{s}\cdot\mathbf{r}_i}$$

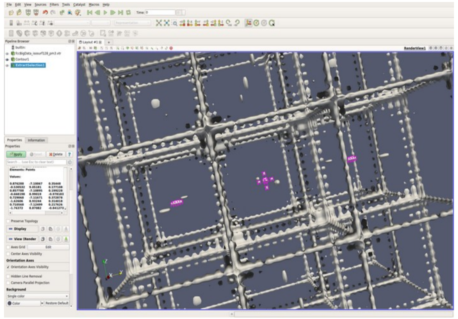
The PyNX package

Python interface: **pynx.gpu** - the main module - returns a complex vector (created by Vincent Favre-Nicolin **http://pynx.sf.net**)

```
$from pynx import gpu
$gpu.Fhkl_thread(h,k,l,x,y,z,gpu_name="GeForce GTX 980")
```

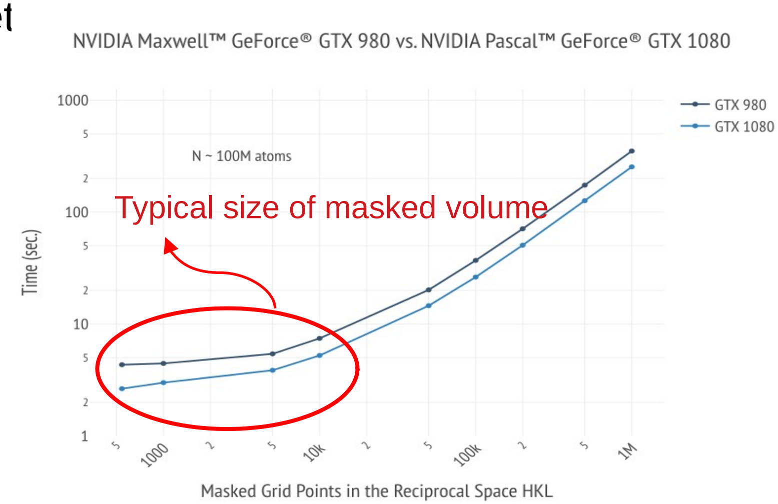
2) visual inspection and masking regions of interest in reciprocal space

- accomplished using ParaView

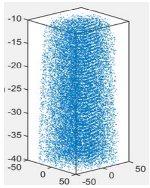


3) backward transformation of masked region to real space

- requires direct summation (h,k,l) → (x,y,z) over selected subset
- implementation in PyNX, GPU: O(10 s) vs. CPU: O(10 min)



$$A^{-1}(\mathbf{r})=\sum_i \ldots$$

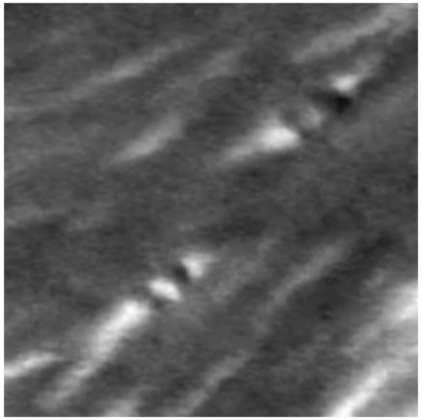


M. Rampp & A. Marek, MPCDF

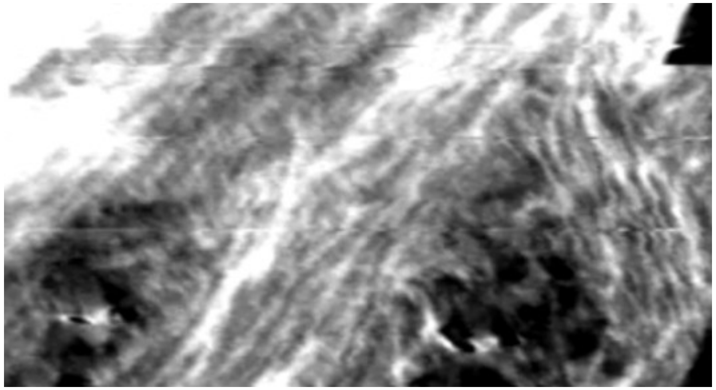
Bone Tissue segmentation

**Automatic Bone Tissue segmentation in 2d/3d images
@MPI for Colloids and Interfaces (Dept. P. Fratzl, group: L. Bertinetti)**

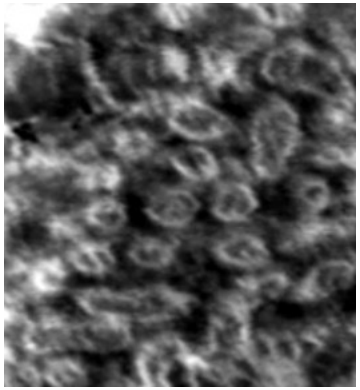
- Goal: use CNN to segment 2d/3d data sets into different classes of bone tissue



class: pore canals



class: fibers



class: villi

Challenges:

→ data (preparation for NN): very unbalanced data sets (a lot of empty pixels)
→ choosing best NN architecture: currently U-NETs are explored
→ memory requirement during training (2d and 3d images are large)

M. Rampp & A. Marek, MPCDF

Automatic segmentation of 3D medical images
@MPI for Human Cognitive and Brain Sciences (Dept. N. Weiskopf)

- Goal: use a (deep) CNN to segment 3D data from histology samples of brain tissue

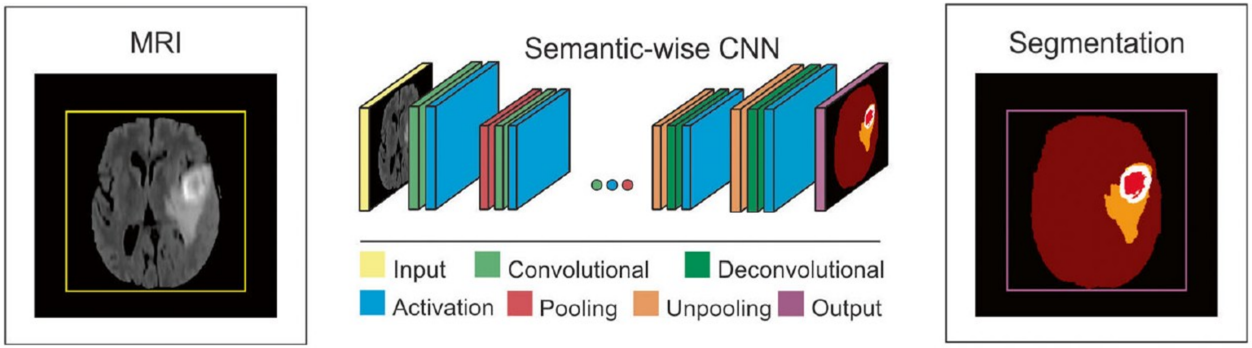


Figure from Z. Akkus et al. 2017: Deep Learning for Brain MRI Segmentation: State of the Art and Future Directions

Challenges:

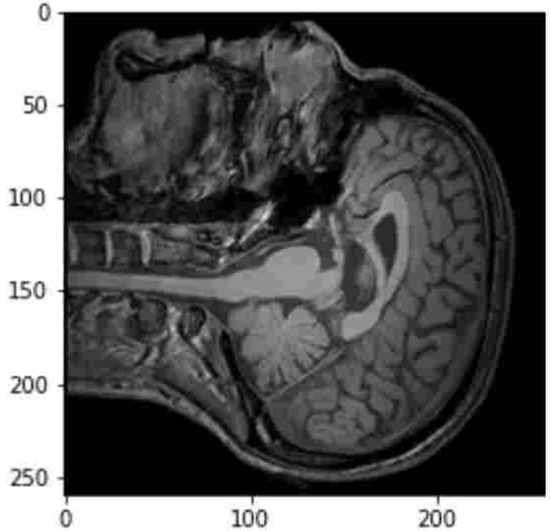
→ memory requirement during inference: ~24 TB in Tensorflow

Solution:

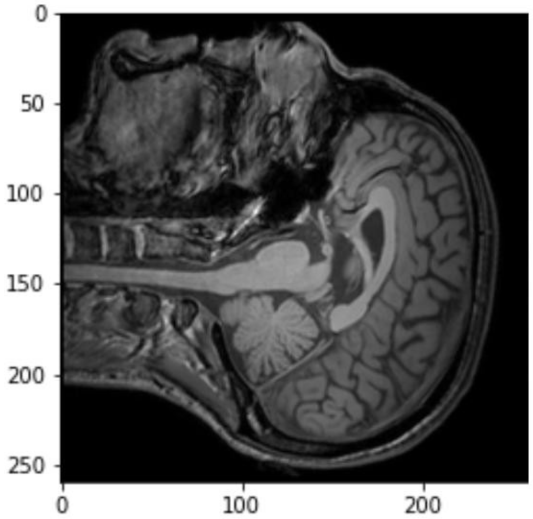
→ implementation of classical “domain-decomposition” approach to parallelize and reduce the memory footprint per node

Automatic Denoising of 3D neuro images @MPI for Human Cognitive and Brain Sciences (Dept. N. Weiskopf)

- Goal: filter out noise from large 3D MRI data sets to achieve “super-resolution”



Use a decoder-encoder NN architecture



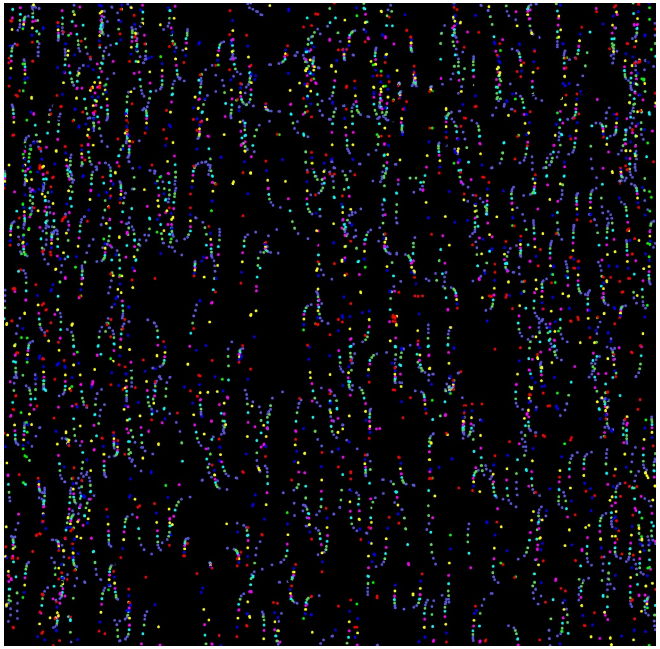
Challenges:

→ each “denoising” of an image implies the training of the NN for ~5000 epochs (no classical “train once, apply often” approach)
→ since memory requirements are too large for GPU, training is very slow (~4d)
→ implement a parallelized approach to distribute training over multiple GPUs until total memory requirement can be satisfied

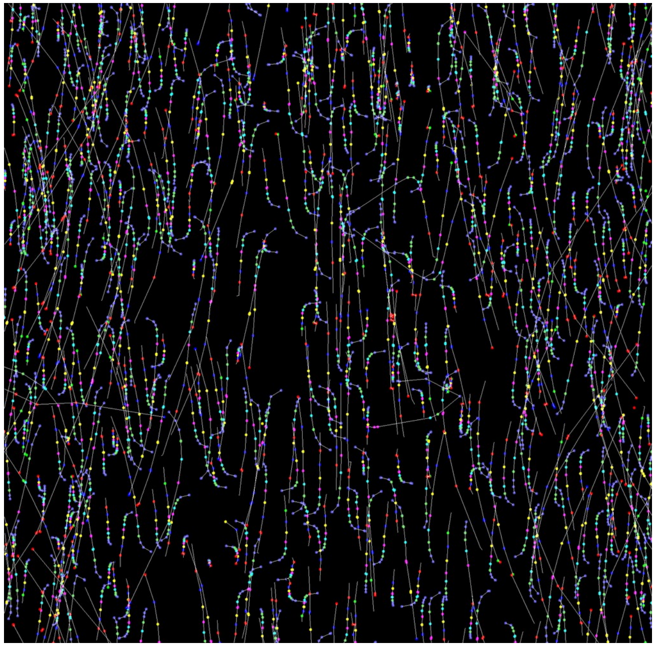
Particle Track Reconstruction

Particle Track Reconstruction of Detector Events in HEP physics @MPI for Physics (group S. Stonjek)

- Goal: extract from the raw events in a detector the (causal) connected points which belong to a particle track



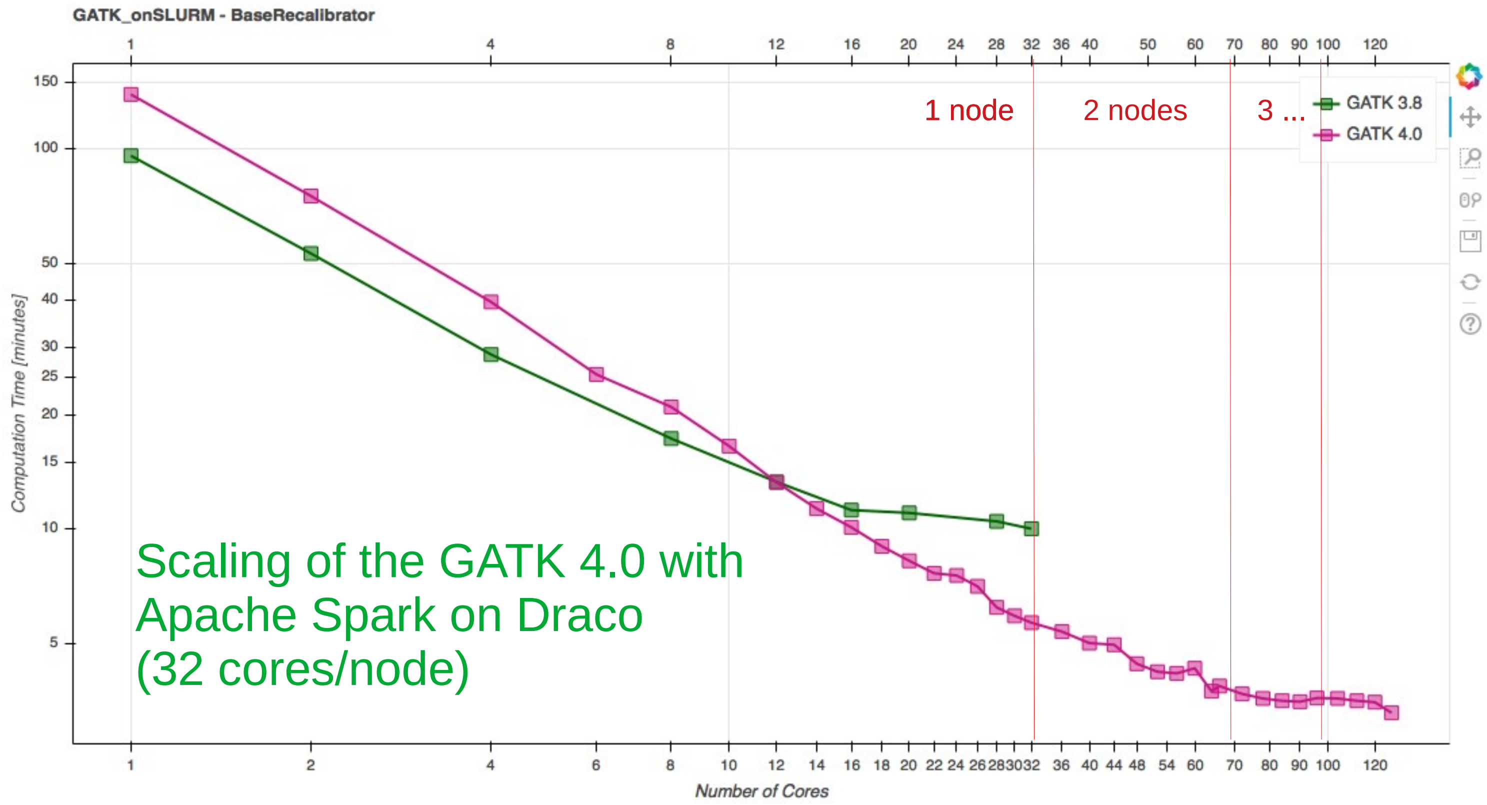
Use a CNN, possibly in combination with a RNN (time information)



This is an exploratory project in order to investigate whether DL methods are compatible with the classical track reconstruction algorithms (very time consuming) of the particle physics community.

Next generation sequencing data at scale @MPI for Biology of Aging (J. Boucas)

- Goal: speed-up an (open-source software), which could previously only run on one node, but now supports Apache Spark parallelization, by providing Apache Spark on MPCDF HPC-systems



=> Significant improvement of turn-around time for the users

Training & Education

- Deep learning workshop (MPCDF, Garching, Sep. 2017)
- MPCDF is partner of MUDS (est. 2018, w/IPP)
- NOMAD summer school (Lausanne/Switzerland, Sep 2018)
- BiGmax summer school (Platja d'Aro/Spain, Sep 2019)