Dr. Andreas Knüpfer
ZIH / TU Dresden

Data-intensive High Performance Computing at TU Dresden: Combining Computer Science Research with Support for Computational Sciences

11th Dresden Probabilistic Workshop

2019-10-10

Overview

— Overview about ZIH
— HPC Hardware and Infrastucture
— Selected Computer Science Challenges
— Selected Computational Science Challenges

Overview about ZIH and HPC

ZIH Structure

- Director Prof. Dr. Wolfgang E. Nagel, deputy director Dr. Andreas Knüpfer
- 7 departments, >160 staff including apprentices, at 6 locations on the campus

- Networking and Communication Services (NK)
- Operational Processes and Systems (OPS)
- Systems Design and Engineering (SDE)
- Service Desk

Service oriented

- Interdisciplinary Application Development and Coordination (IAK)
- Distributed and Data Intensive Computing (VDR)
- Innovative Methods of Computing (IMC)

Research oriented

ZIH Topics

IT Services for TU Dresden

- Campus network, internet uplink, phone network
- E-Mail, groupware, data exchange, backup etc.
- Software procurement

Computational Science Services

- Virtual machines, hosting
- Supercomputing
- Big Data methods

Research and Development

- Parallel programming and algorithms
- Performance optimization and scaling, energy efficiency, …
- Data Analytics applications, machine learning, …

ZIH Topics

IT Services for TU Dresden

- Campus network, internet uplink, phone network
- E-Mail, groupware, data exchange, backup etc.
- Software procurement

Computational Science Services

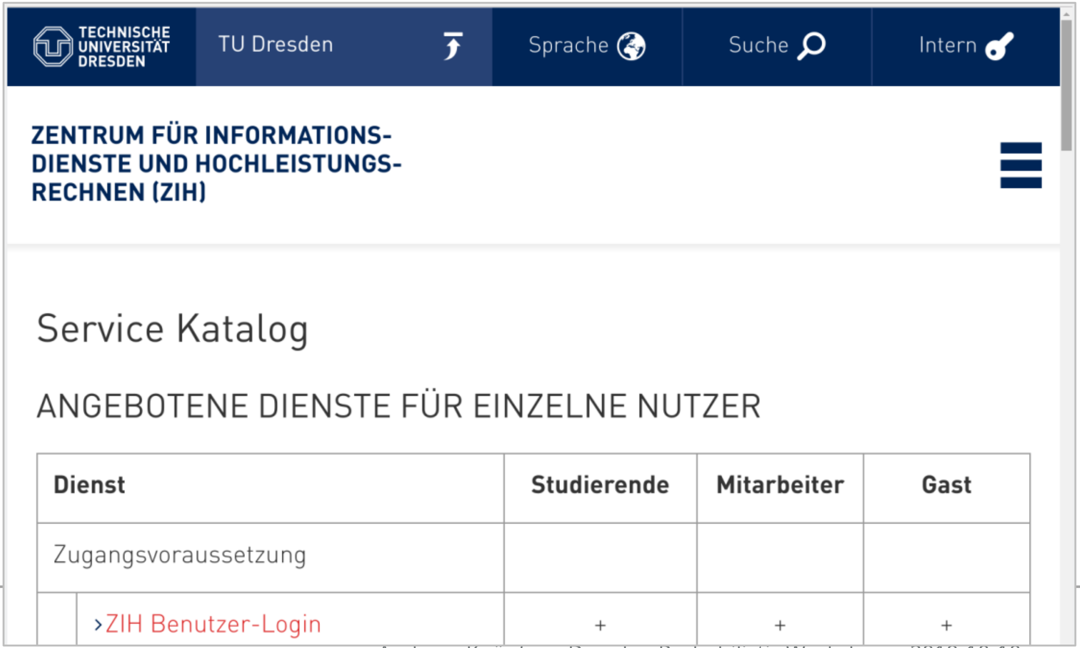
- Virtual machines, hosting
- Supercomputing
- Big Data methods

Research and Development

- Parallel programming and algorithms
- Performance optimization and scaling, energy efficiency, …
- Data Analytics applications, machine learning, …

IT Services and the Service Catalog

- https://tu-dresden.de/zih/dienste/service-katalog
- Topic lists and service descriptions

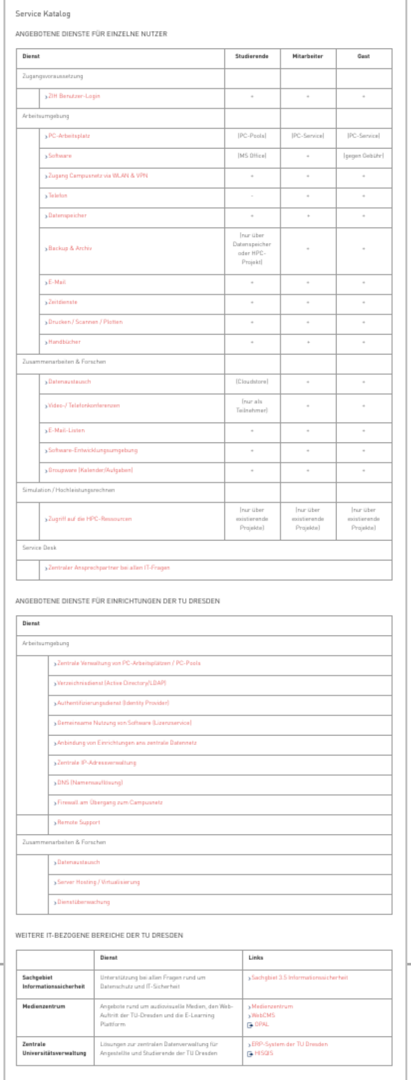

TECHNISCHE UNIVERSITÄT DRESDEN | TU Dresden | Sprache | Suche | Intern

ZENTRUM FÜR INFORMATIONS-DIENSTE UND HOCHLEISTUNGS-RECHNEN (ZIH)

Service Katalog

ANGEBOTENE DIENSTE FÜR EINZELNE NUTZER

Dienst	Studierende	Mitarbeiter	Gast
Zugangsvoraussetzung			
›ZIH Benutzer-Login	+	+	+


Service Katalog

ANGEBOTENE DIENSTE FÜR EINZELNE NUTZER

Dienst	Studierende	Mitarbeiter	Gast
Zugangsvoraussetzung			
›ZIH Benutzer-Login	+	+	+
Arbeitsumgebung			
›PC-Arbeitsplatz	(PC-Pools)	(PC-Service)	(PC-Service)
›Software	(MS Office)	+	(gegen Gebühr)
›Zugang Campusnetz via WLAN & VPN	+	+	+
›Telefon	-	+	+
›Datenspeicher	+	+	+
›Backup & Archiv	(nur über Datenspeicher oder HPC-Projekt)	+	+
›E-Mail	+	+	+
›Zeitdienste	+	+	+
›Drucken / Scannen / Plotten	+	+	+
›Handbücher	+	+	+
Zusammenarbeiten & Forschen			
›Datenaustausch	(Cloudstore)	+	+
›Video-/ Telefonkonferenzen	(nur als Teilnehmer)	+	+
›E-Mail-Listen	+	+	+
›Software-Entwicklungsumgebung	+	+	+
›Groupware (Kalender/Aufgaben)	+	+	+
Simulation / Hochleistungsrechnen			
›Zugriff auf die HPC-Ressourcen	(nur über existierende Projekte)	(nur über existierende Projekte)	(nur über existierende Projekte)
Service Desk			
›Zentraler Ansprechpartner bei allen IT-Fragen			

ANGEBOTENE DIENSTE FÜR EINRICHTUNGEN DER TU DRESDEN

Dienst
Arbeitsumgebung
›Zentrale Verwaltung von PC-Arbeitsplätzen / PC-Pools
›Verzeichnisdienst (Active Directory/LDAP)
›Authentifizierungsdienst (Identity Provider)
›Gemeinsame Nutzung von Software (Lizenzservice)
›Anbindung von Einrichtungen ans zentrale Datennetz
›Zentrale IP-Adressverwaltung
›DNS (Namensauflösung)
›Firewall am Übergang zum Campusnetz
›Remote Support
Zusammenarbeiten & Forschen
›Datenaustausch
›Server Hosting / Virtualisierung
›Dienstüberwachung

WEITERE IT-BEZOGENE BEREICHE DER TU DRESDEN

	Dienst	Links
Sachgebiet Informationssicherheit	Unterstützung bei allen Fragen rund um Datenschutz und IT-Sicherheit	›Sachgebiet 3.5 Informationssicherheit
Medienzentrum	Angebote rund um audiovisuelle Medien, den Web-Auftritt der TU-Dresden und die E-Learning Plattform	›Medienzentrum ›WebCMS OPAL
Zentrale Universitätsverwaltung	Lösungen zur zentralen Datenverwaltung für Angestellte und Studierende der TU Dresden	›ERP-System der TU Dresden HISQIS

Topics overview

IT Services for TU Dresden	• Campus network, internet uplink, phone network • E-Mail, groupware, data exchange, backup etc. • Software procurement
Computational Science Services	• Virtual machines, hosting • Supercomputing • Big Data methods
Research and Development	• Parallel programming and algorithms • Performance optimization and scaling, energy efficiency, … • Data Analytics applications, machine learning, …

HPC Resources at ZIH

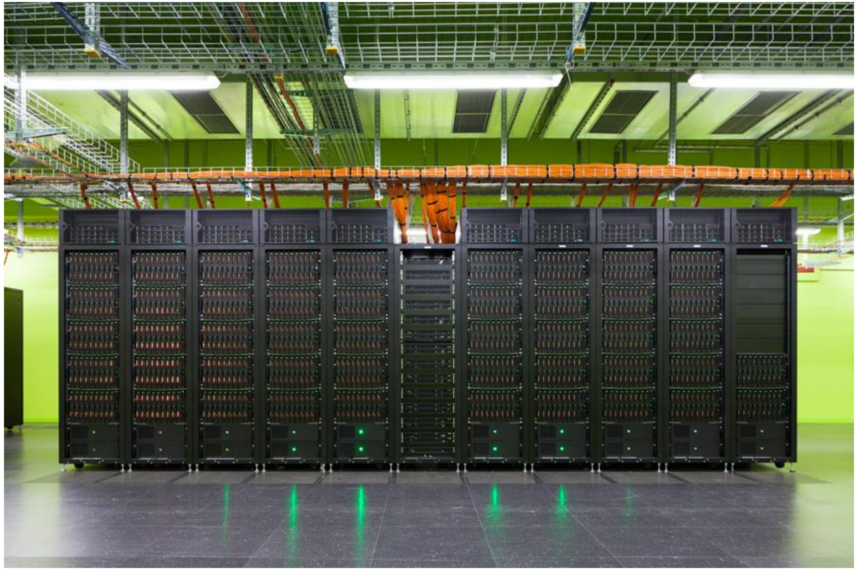
HRSK II 2015

- ~ 44,000 cores Intel (mostly Haswell)
- 256 GPUs Nvidia Tesla K80 +
- 44 GPUs: Nvidia Tesla K20
- 136 TB RAM, >5 PB scratch file system

HPC-DA extension 2018

- 22 Machine Learning nodes IBM AC922
- 2 PB NVME storage (90 nodes, NVMEoF) with 2 TB/s bandwidth in total
- 10 PB Object Storage

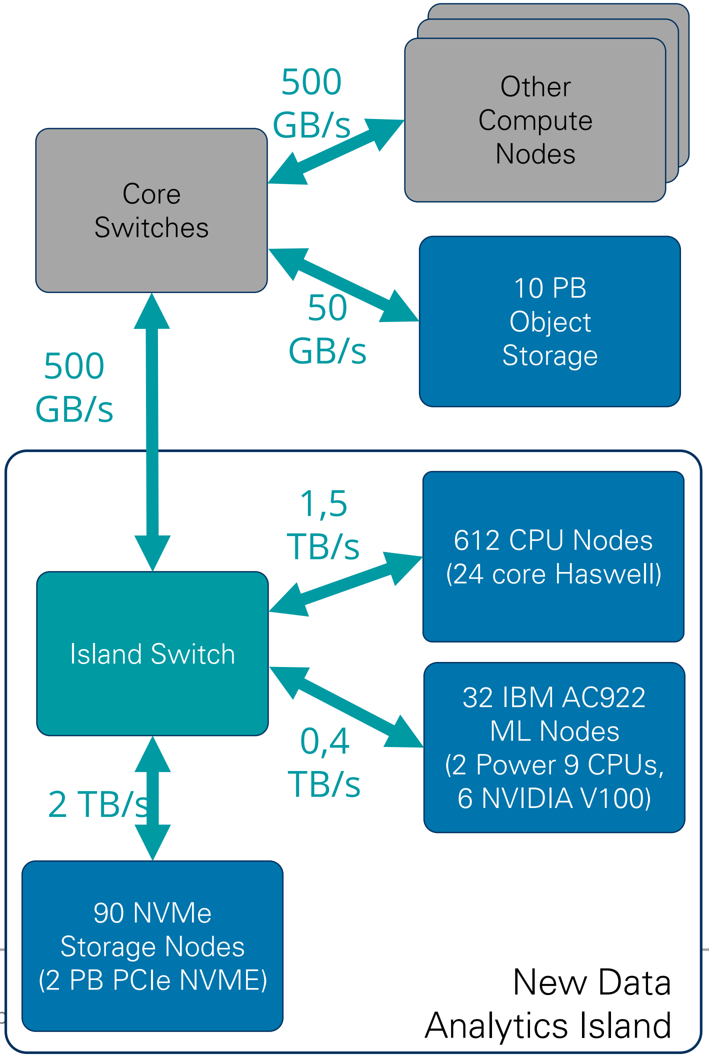
Follow-up extension in 2019 (approx. 4 M€)



HPC-DA Extension 2018

HPC-DA extension towards extremely fast I/O

- Redesigned one compute island of HRSK II
- Strong focus on highest bandwith and low latency
- 612 existing CPU compute nodes
- 32 Machine Learning Knoten IBM AC922
  - 2x Power-9 CPUs, 6x NVIDIA V100 GPUs, NVLink
- 90 NVME storage nodes
  - Each node with 8 3,2 TB PCIe x4 NVME cards
  - Dual-link EDR IB, NVME over fabric
- 10 TB Object Storage with 50 GB/s



HPC-DA ML Nodes

Hardware

— 32 IBM AC922 nodes
— 2x POWER9 CPU, 22 core, 4-way HT
(176 threads per node in total)
— 2.80 GHz, 3.10 GHz boost
— 256 GB RAM DDR4 2666MHz
— 6x NVIDIA VOLTA V100 with 32GB HBM2
— NVLINK with 150 GB/s between GPUs
and between host and GPUs
— CPUs and GPUs direct water cooled
— 0.4 TB/s aggregated bandwidth to NVME nodes

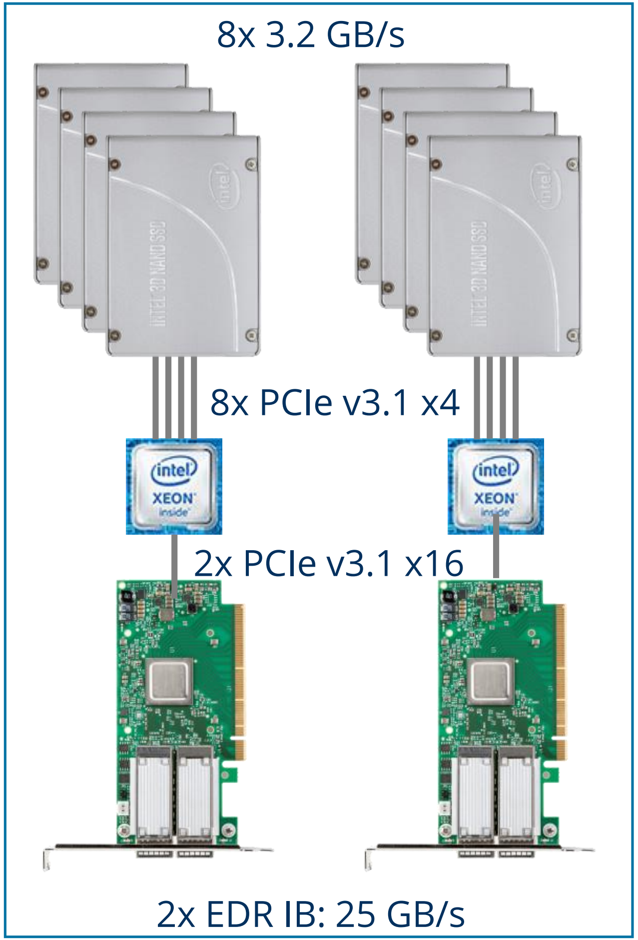
Image: https://www.ibm.com/it-infrastructure/power/accelerated-computing

HPC-DA NVME Storage Nodes

Hardware

- 90 NVME storage nodes
- 2 sockets Intel Xeon E5-2620 v4 (16 cores each, 2.10GHz)
- 64 GB RAM

- 8x Intel NVMe Datacenter SSD P4610, 3.2 TB, PCIe 3.1 x4 3DNAND ME 2.5" U.2, 3,2 GB/s (8x 3.2 ==25.6 GB/s)
- 2 Infiniband EDR links, Mellanox MT27800, ConnectX-5, PCIe x16, 100 Gbit/s each



90 x

90 NVME storage nodes

in total

2 TB/s

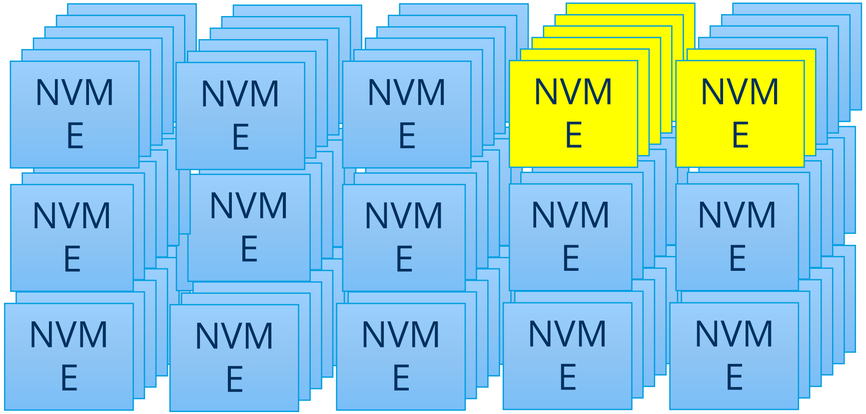
peak bandwidth

HPC-DA NVME Usage Models

Allocation strategy

— NVME shares allocated as long-term "NVME leases" (weeks to months)

— Granularity of 1/8th node (1 NVME card) or full NVME nodes



Separate BeeGFSes in own NVME lease

— Instantiate separate BeeGFS

— Granularity of ½ or 1 NVME nodes, including the EDR links and the CPU cores

— Separate MDSs an OSTs for this FS

— No meta-data interference with everyone else

— Full nominal bandwidth per NVME node

— Have it mounted automatically to compute jobs of your HPC project

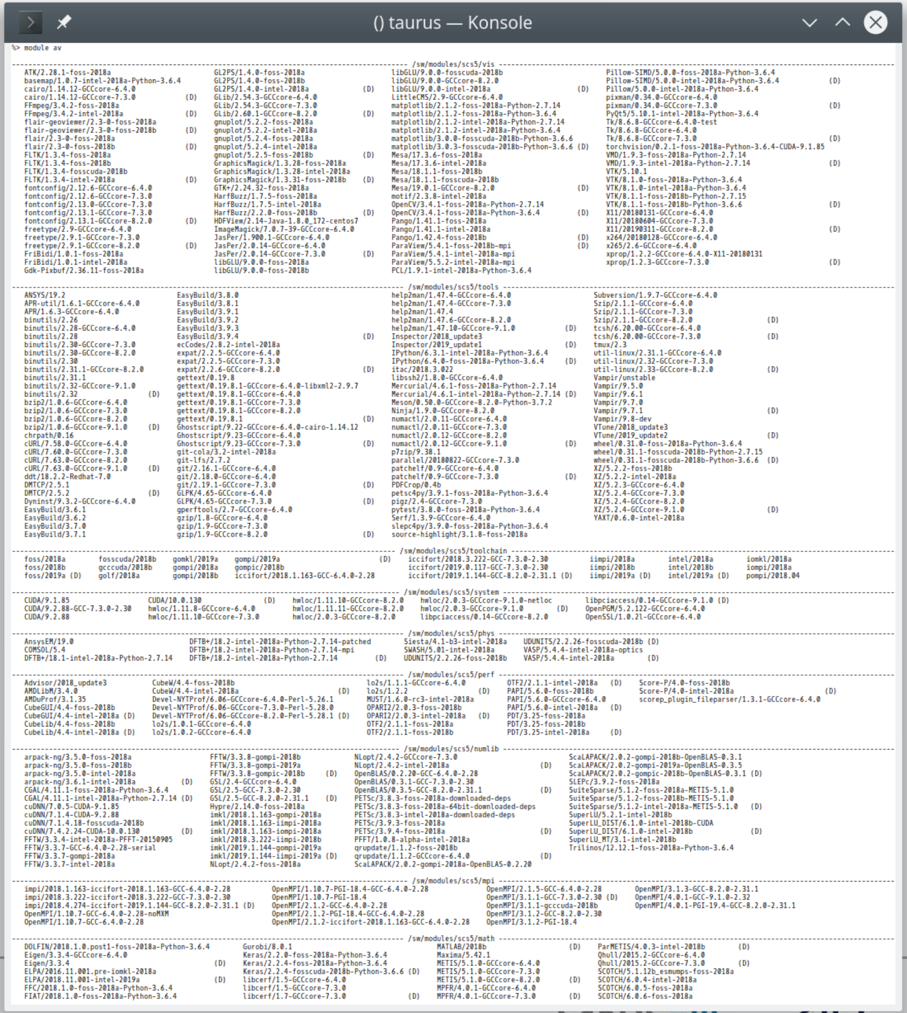
Software for HPC-DA: Modules vs. Containers

A “basic” software installation

- Classical HPC software packages and libraries
- Big Data frameworks
- AI frameworks

Software modules

- Long list of software packages, multiple versions each, dependencies
- Open Source SW, scientific community software packages, commercial SW
- Application software, libraries, software tools
- Compiled vs. interpreted (Python etc.)
- Central and individually per project/user



Software for HPC-DA: Modules and Containers

Singularity containers

- Tailored software environments that you can take with you or share with others
- Can be defined/built on top of each other
- Independent (mostly) from central software installation, updates, version changes, …
- Cannot combine two existing containers, though

- Automatic building from definition files
- Combine with version control for definition files

- Some challenge to build containers for Power9
  … because few have Power9 laptops yet ;)
- Automatic solution with Singularity definition files

- Designed for HPC
- Works well with MPI, Infiniband, GPUs, …
- Thin run-time layer
- No danger of privilege escalations

- Huge supply of prebuilt containers available
- Also compatible with Docker images

See https://sylabs.io and https://sylabs.io/docs/

Topics overview

IT Services for TU Dresden

- Campus network, internet uplink, phone network
- E-Mail, groupware, data exchange, backup etc.
- Software procurement

Computational Science Services

- Virtual machines, hosting
- Supercomputing
- Big Data methods

Research and Development

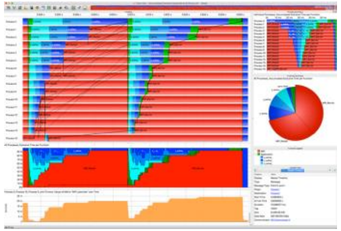
- Parallel programming and algorithms
- Performance optimization and scaling, energy efficiency, …
- Data Analytics applications, machine learning, …

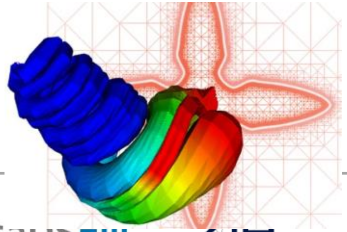
Selected Computer Science Challenges

Center for Information Services &
High Performance Computing

ZIH Computer Science Research Topics

- Scalable software tools
to support the optimization
of applications for HPC systems
- Data Intensive Computing
and Data Life Cycle
- Performance and energy efficiency
analysis for innovative
computer architectures
- Distributed Computing and Cloud Computing
- Data analysis, methods, and
modeling in life sciences
- Parallel programming,
algorithms and methods





Parallel Performance Analysis Tools

Parallel Performance Analysis

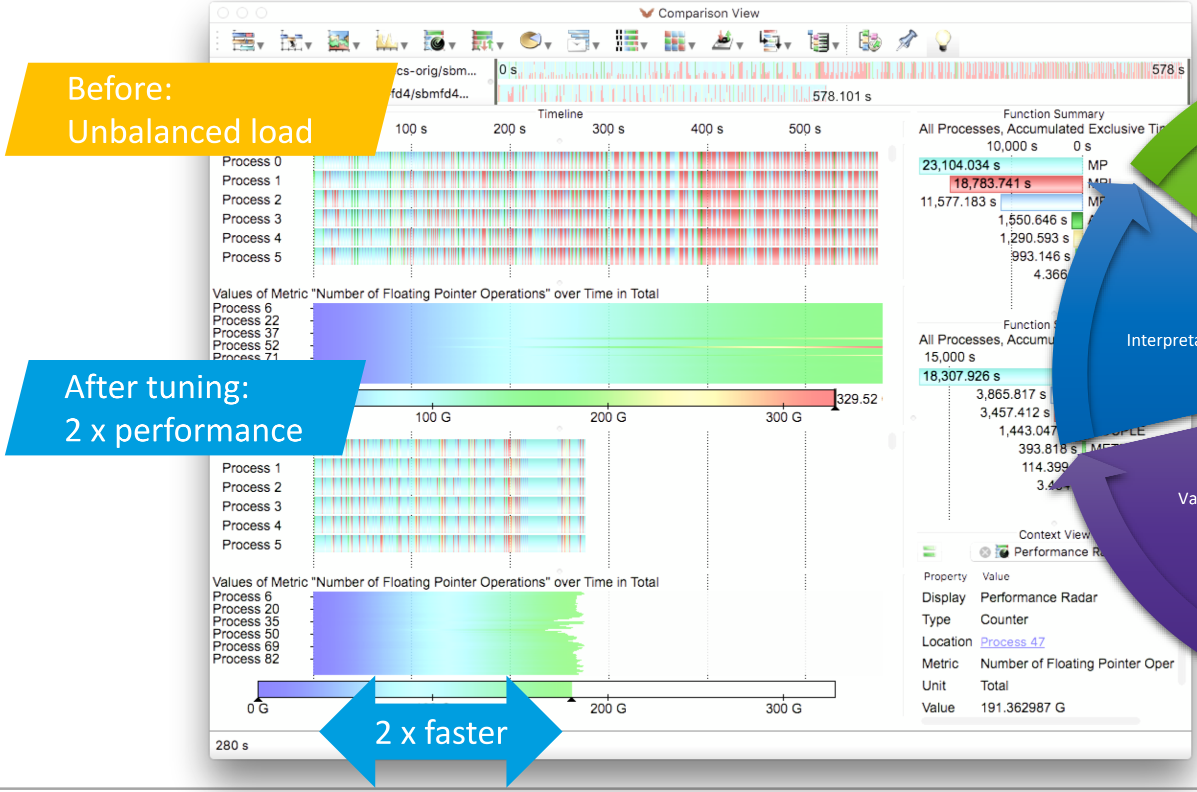
Enable or improve computational sciences

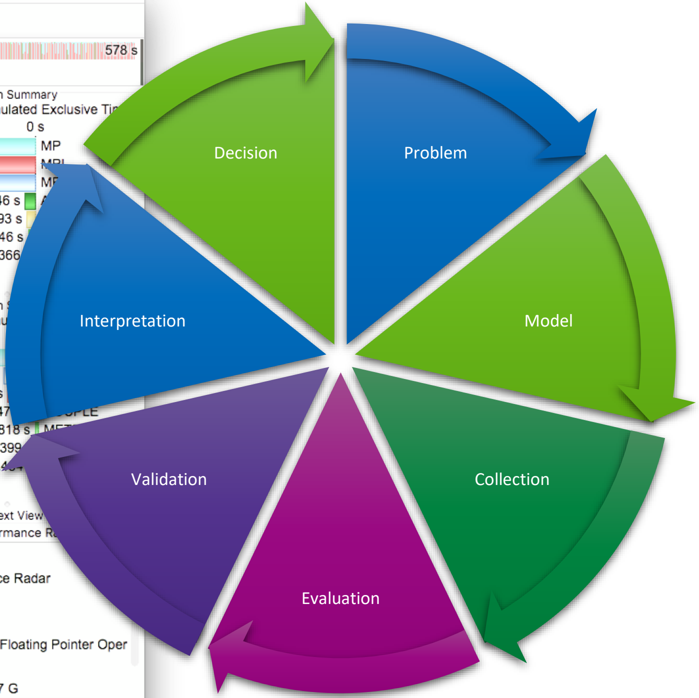
- Throughput and Response time
- Strong and weak Scaling
- Capability for larger / more complex scenarios
- Additional functionality / novel aspects
- Reliability
- Development Cycle

Score-P
Scalable performance measurement infrastructure for parallel codes

VAMPIR

Performance Consulting is a Cyclic Process





TECHNISCHE UNIVERSITÄT DRESDEN

ScaDS DRESDEN LEIPZIG

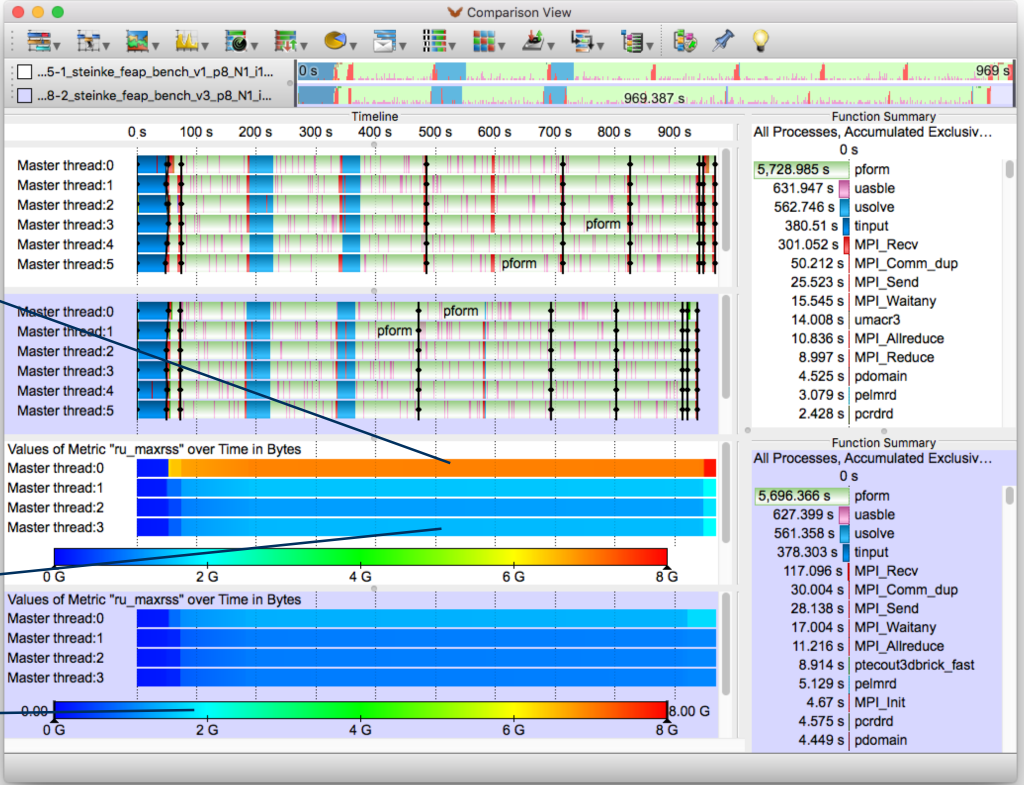
ZIH

Analyzing Unexpected Memory Demand

Memory demand on rank 0 explicably high (~ 8 GB)

Memory demand on rank 1 – (N-1) also too high (~1.8 GB)

Prior to new module (output) demand was ~1GB per rank

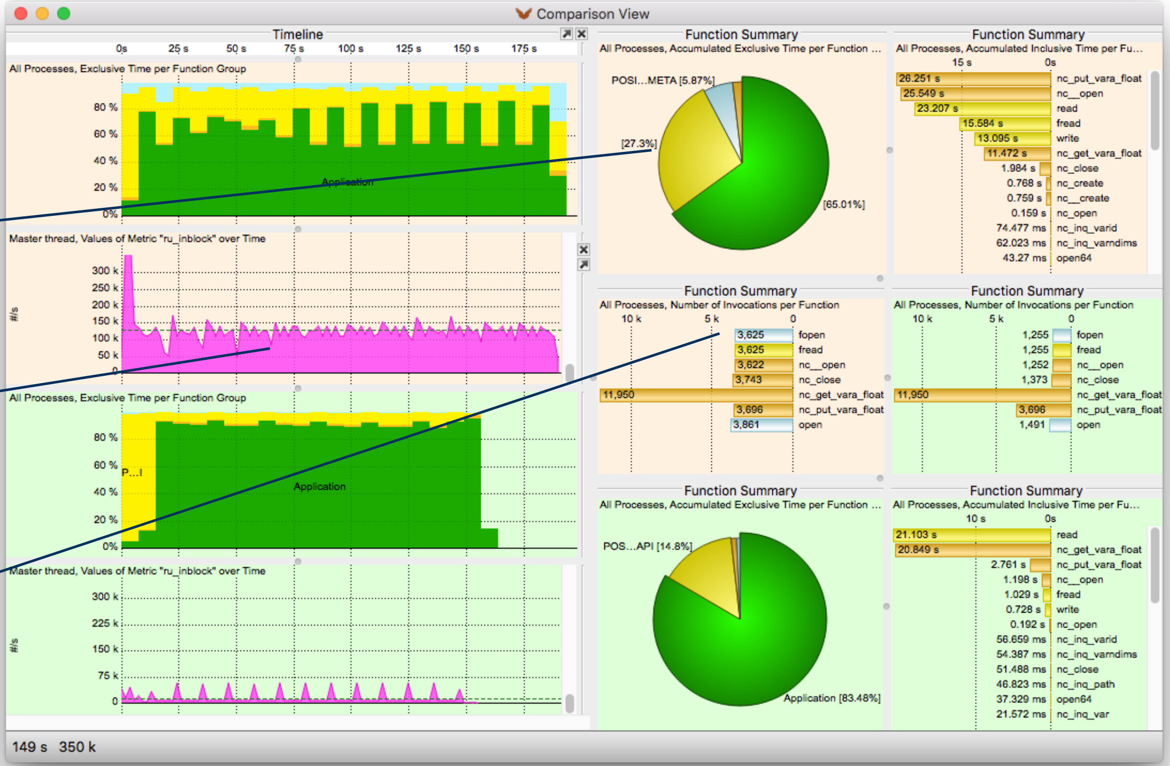


Analyzing Bottleneck in Multi-Stage-I/O

High I/O load on a single rank

~ 260 MB/s per rank !

High number of I/O open AND read operations

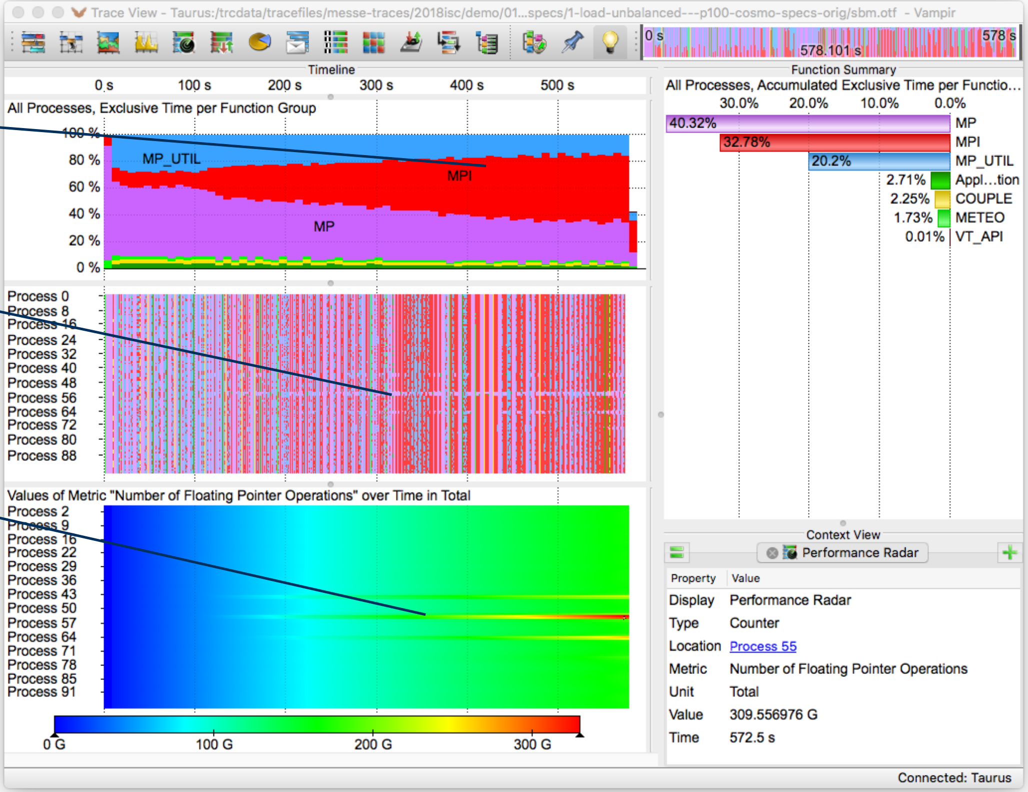


Load Balance Analysis of Weather Forecast Model

Message Passing share increases over time. Happens uniformly?

No, MPI Share does not increases on Process 56. Why not?

Process 56 is loaded heavily with FLOPs starting at t=130s



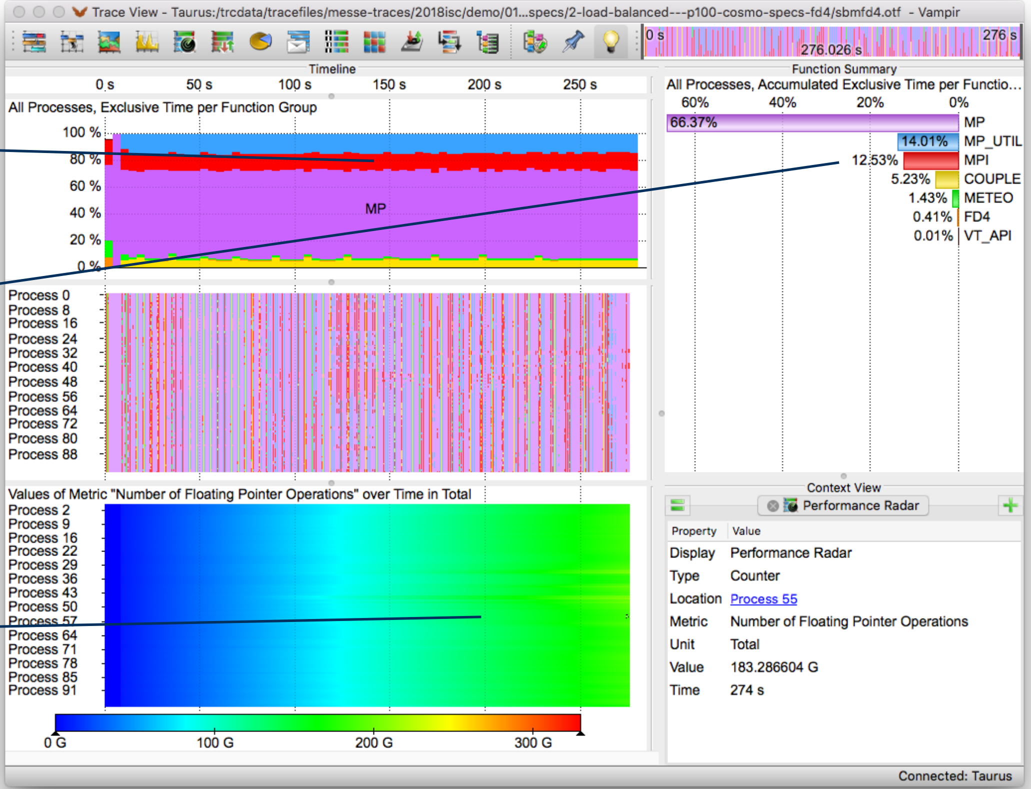
Load Balance Analysis of Weather Forecast Model

After Tuning

No increase in MPI share anymore

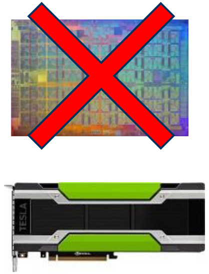
MPI consumes 12.5 % of the total time (down from 33%)

FLOP load is equally distributed now



Parallel Programming Abstractions with C++

Spectrum of Computing Architectures


Supercomputer
100 000+ cores

Cluster
1000s of cores

Manycore

Server
10s of cores

Notebook
2-4 cores

Mobile
2-8 cores

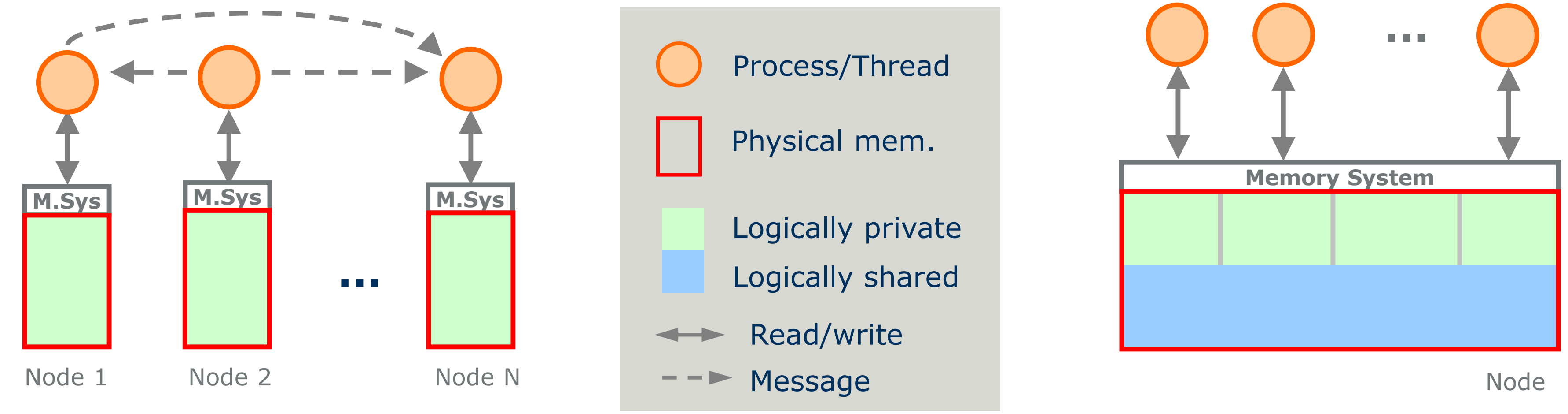
Distributed Memory (DM)

Shared Memory (SM)

DM programming: MPI, Charm++, …

SM programming: OpenMP, Pthreads, Cilk, TBB, …

Shared Memory vs. Distributed Memory Programming



Message Passing

- (+) Performance, runs everywhere
- (−) Productivity

Threading

- (+) Productivity
- (−) Locality control, limited to SM hardware

PGAS – Combining the Advantages of both Approaches

PGAS: Partitioned Global Address Space

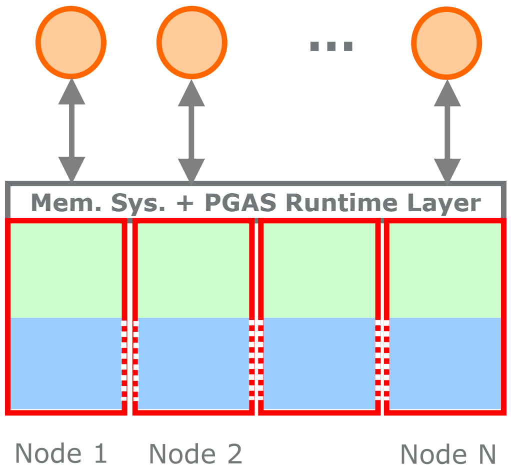
Process/Thread

Physical mem.

Logically private

Logically shared

Read/write (local)
Put/get (remote)



PGAS *Languages*

Chapel, CoArray Fortran, UPC, ...

PGAS *Libraries*

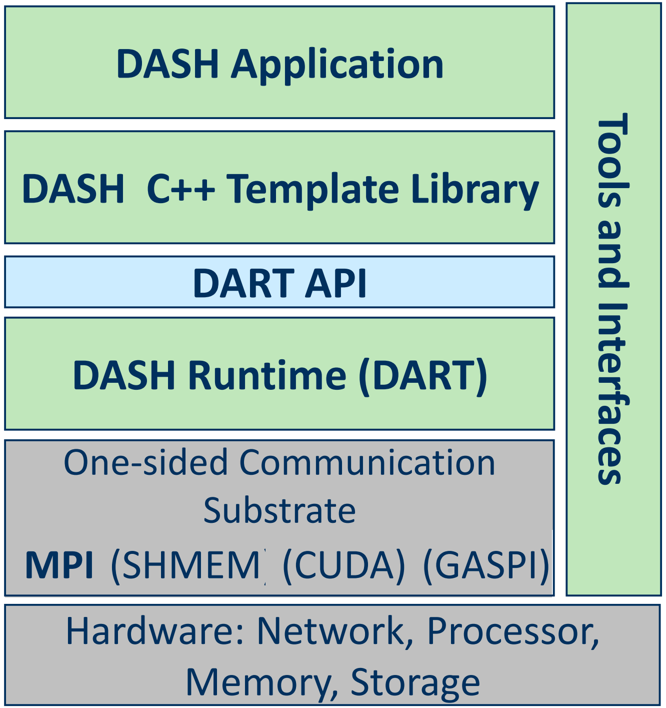
Global Arrays (GA), GASPI, OpenShmem, MPI3.0 RMA

Locality control, runs everywhere, performance and productivity

DASH C++ Template Library for Parallel Programming

— C++ template library for application programmers

— Distributed data container classes
— Similar to the C++ STL container classes, compatible
— Built-in knowledge about distribution
— Algorithms similar to STL on distributed containers

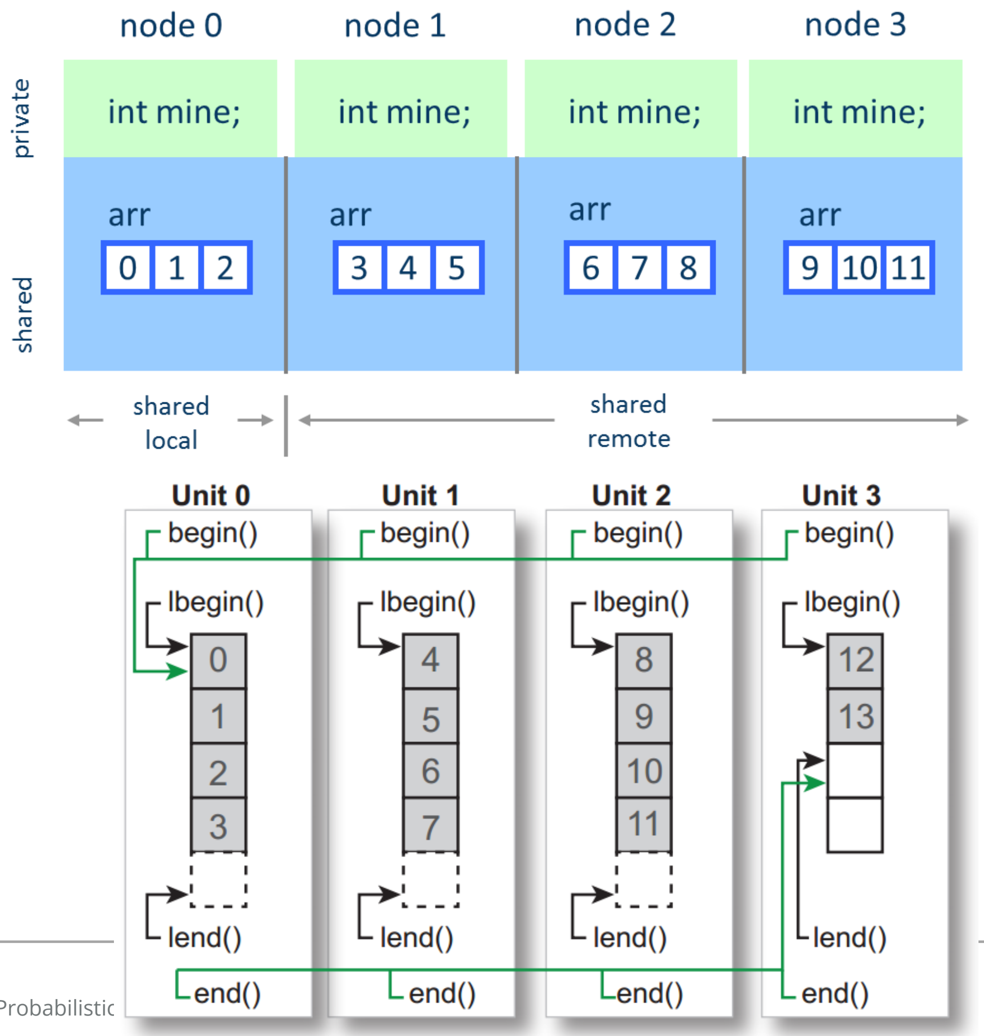


dash
www.dash-project.org
Distributed Data Structures and Parallel Algorithms

DASH Array

DASH n-dimensional array

- Global random access with **`begin(),end()`** and **`[]`** … via slow element-wise get
- Dedicated local access with **`myarray.local.begin() / .end()`** and **`.local[]`** … direct and fast
- Configurable data distribution patterns in n dimensions
- STL-like algorithms considering actual data distribution patterns



Selected Results from Computational Science Projects at ZIH

Center for Information Services &
High Performance Computing

Prestige Project

Gefördert durch:

Bundesministerium für Wirtschaft und Technologie

aufgrund eines Beschlusses des Deutschen Bundestages

Ongoing LuFo V project PRESTIGE "Interdisciplinary probabilistic analysis of compressor blades accounting for real geometry effects"

TU Dresden partners:

- Chair of Turbomachinery and Flight Propulsion (TFA)
- Institute of Lightweight Engineering and Polymer Technology (ILK)
- ZIH

- Coordinator Rolls-Royce Deutschland Ltd & Co KG
- DLR Braunschweig and Köln
- TU Darmstadt
- University of Surrey (UK)
- Kitware
- Intelligent Light
- T-Systems Solutions for Research GmbH

… not the focus today

Which one next?

Laser Particle Acceleration
or
Cartography of the Milky Way

Courtesy of Dr. Michael Bussmann et.al.

Novel Particle Accelerators and Highly Scalable GPU-based Simulation

Center for Information Services & High Performance Computing

Laser Wakefield Acceleration Experiments at HZDR

J.P. Couperus et al.: Demonstration of a beam loaded nanocoulomb-class laser wakefield accelerator. **Nature Comunications** 8.1 (2017)
A. Irman et al.: Improved performance of laser wakefield acceleration by tailored self-truncated ionization injection. **LPAW proceeding** pp.1-13 (2017)

HELMHOLTZ ZENTRUM DRESDEN ROSSENDORF

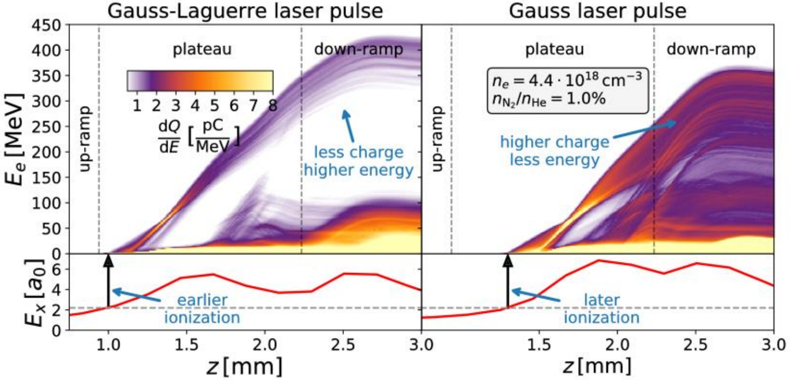
- Advanced method of electron acceleration
- Based on highly non-linear laser plasma interaction
- Requires large scale particle-in-cell simulation for modeling
- Hundreds of simulations on up to 146 K80 GPUs performed with PIConGPU at Taurus/ZIH

PIConGPU

Tuned with VAMPIR

Video

Simulations accompanying experiment at HZDR



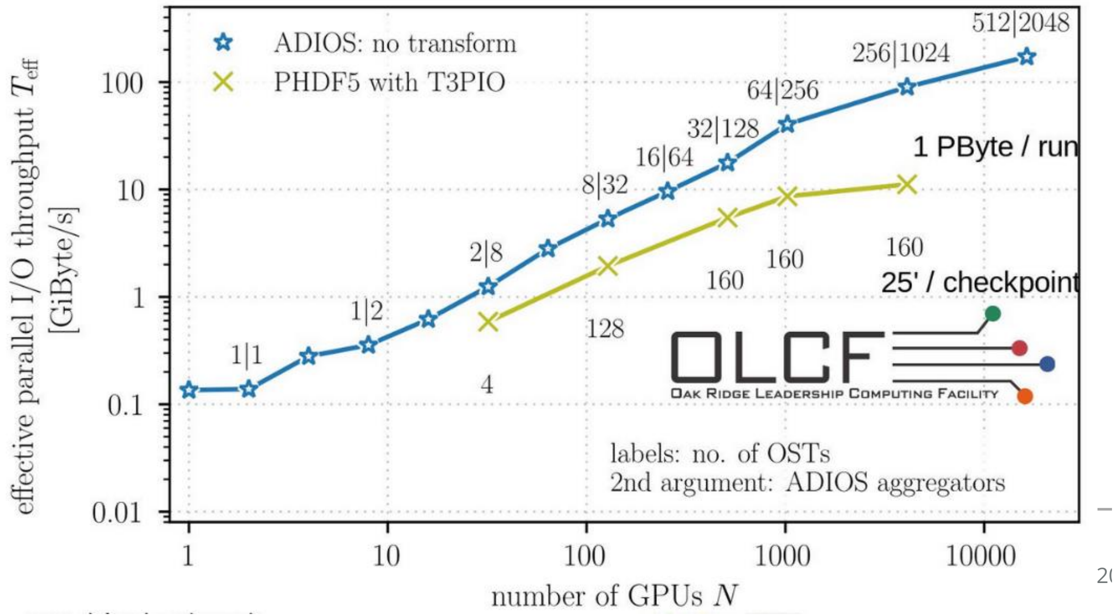
Studying the influence of higher order laser modes

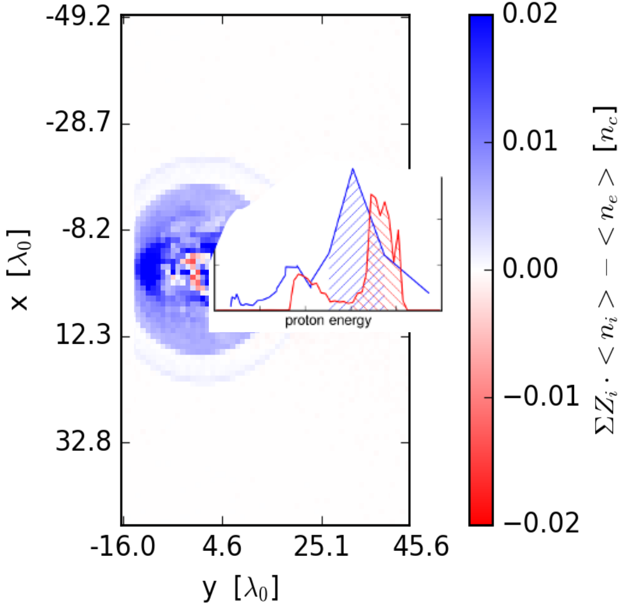
Laser-Ion Acceleration with Mass-Limited Targets

P. Hilz, T.M. Ostermayr, A. Huebl et al.: Isolated proton bunch acceleration by a petawatt laser pulse. **Nature Comunications** 9.423 (2018)
A. Huebl et al.: On the Scalability of Data Reduction Techniques in Current and Upcoming HPC Systems from an Application Perspective, **ISC'17**, LNCS 10524

HZDR HELMHOLTZ ZENTRUM DRESDEN ROSSENDORF

— 3D simulation of novel, fully isolated target for laser-ion acceleration
— 15 M CPUhrs (½ MGPUhrs), INCITE Award Highlight
— PByte-Scale I/O through ADIOS at Titan/OLCF

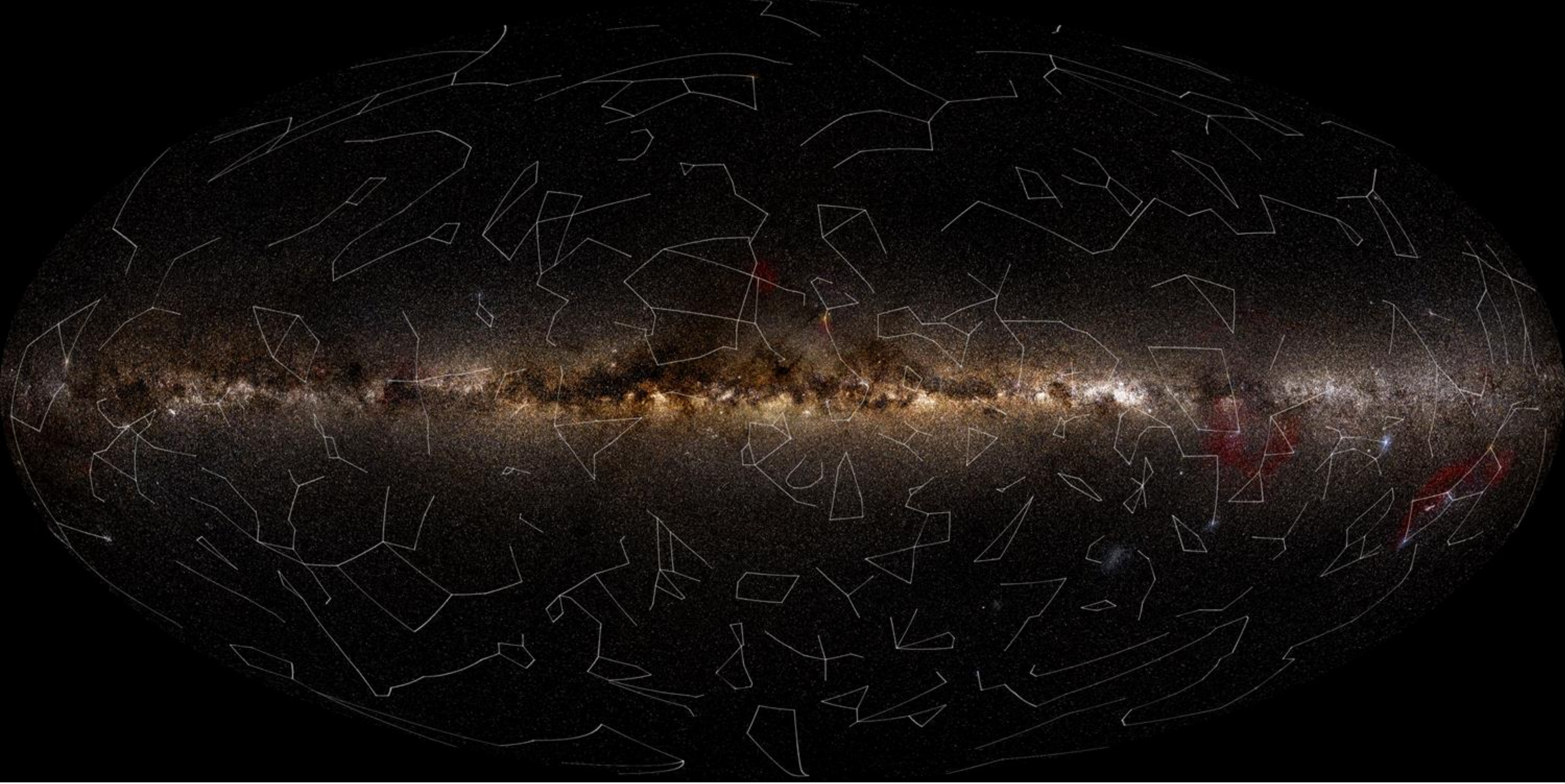




Courtesy of Prof. Sergei A. Klioner et.al.

Astrophysics: The Gaia Project

Center for Information Services &
High Performance Computing

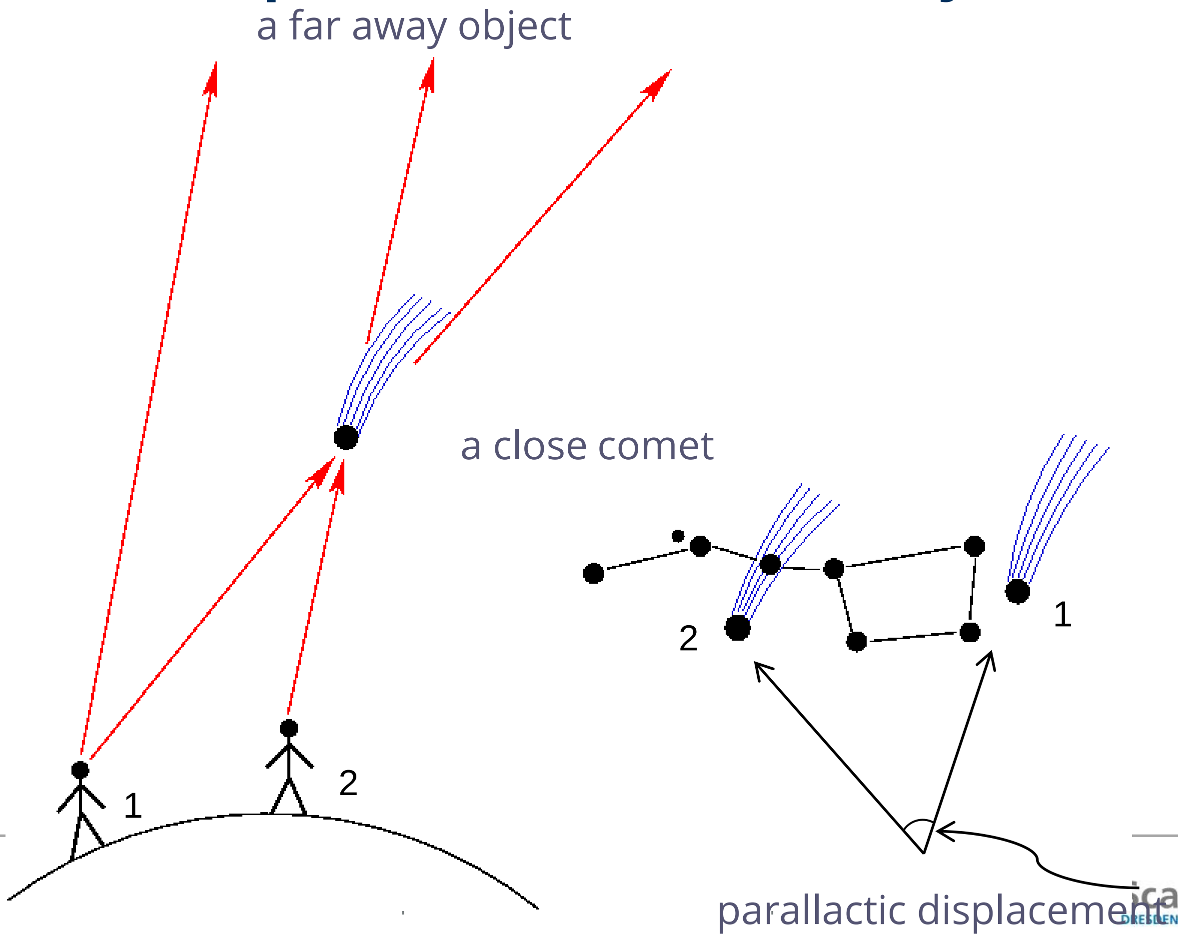


One of the main problems of astronomy: distance

Without knowing how far the object is, physical understanding of that object is impossible…

A comet: far away and very big or inside the Earth atmosphere and rather small?
Tycho Brahe, 1577

One of the main problems of astronomy: distance



Astrometry: the art of measuring stellar positions

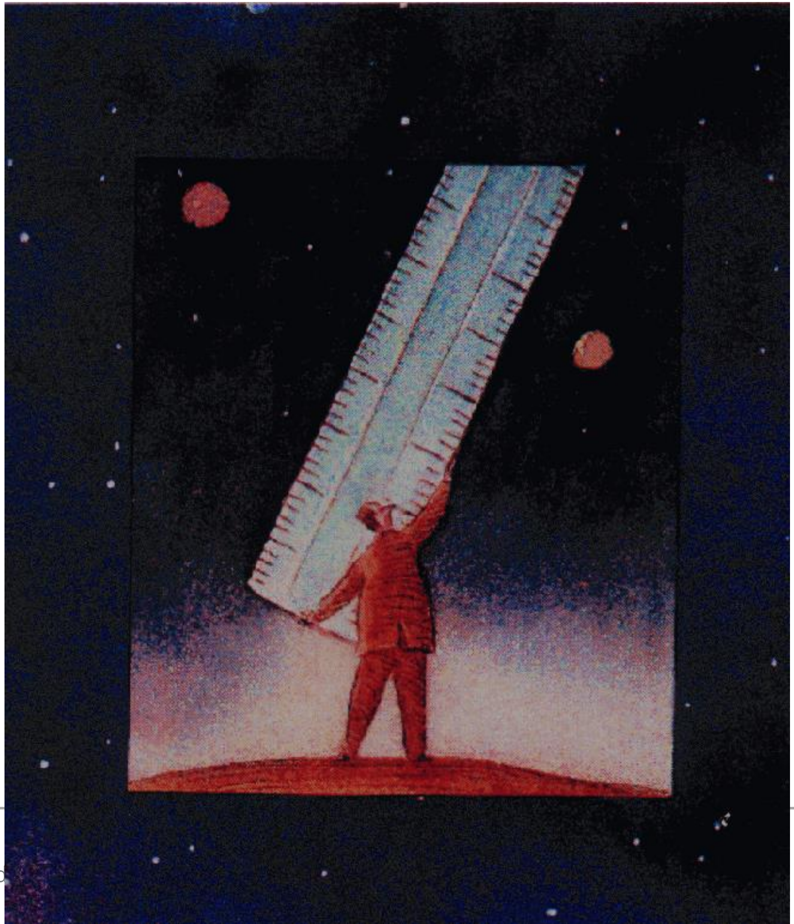
Astronomy cannot touch its objects!

Astronomy cannot make experiments!

Astronomy analyses stellar light:

Astrometry	– direction
Photometry	– quantity
Spectroscopy	– colour and more
Polarimetry	– polarization

+ cosmic particles
++ gravitational waves



Why to bother?

- We need to understand stars.

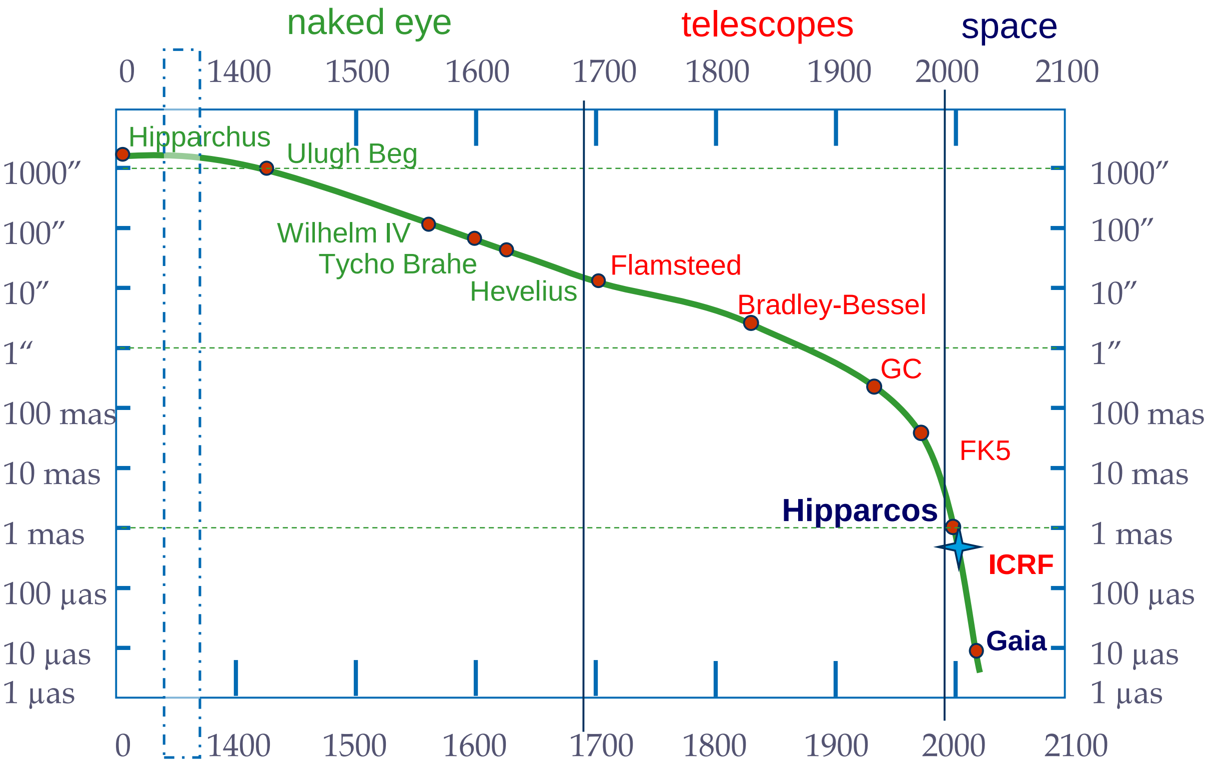
  (our Sun is a star!)

  Without knowing the distance it is not possible to judge if a star is big or small, etc.

- We live in a galaxy.

  We need to understand how our Galaxy was formed.

Accuracy of astrometric observations



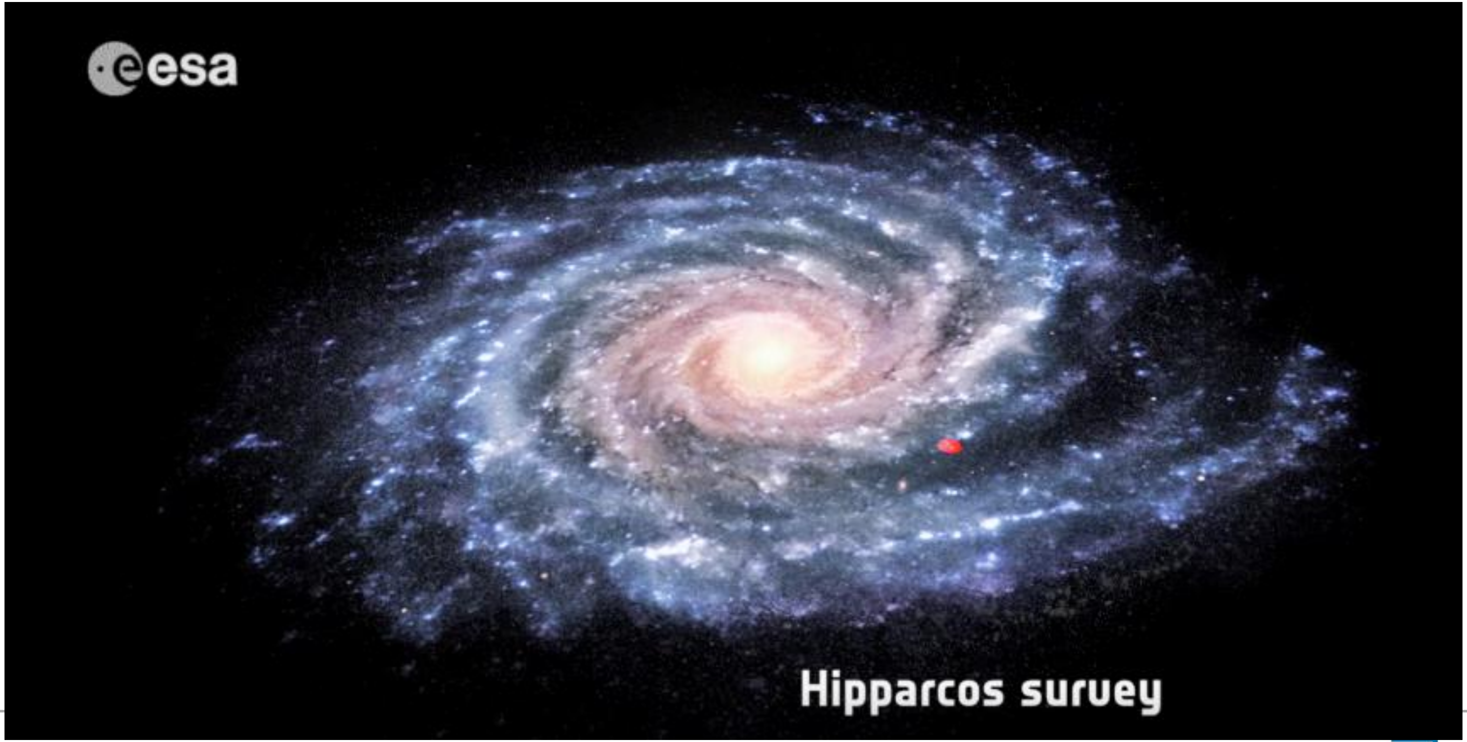
1 µas is the thickness of a sheet of paper seen from the other side of the Earth

What do we know about our Galaxy?

The Sun should be here:

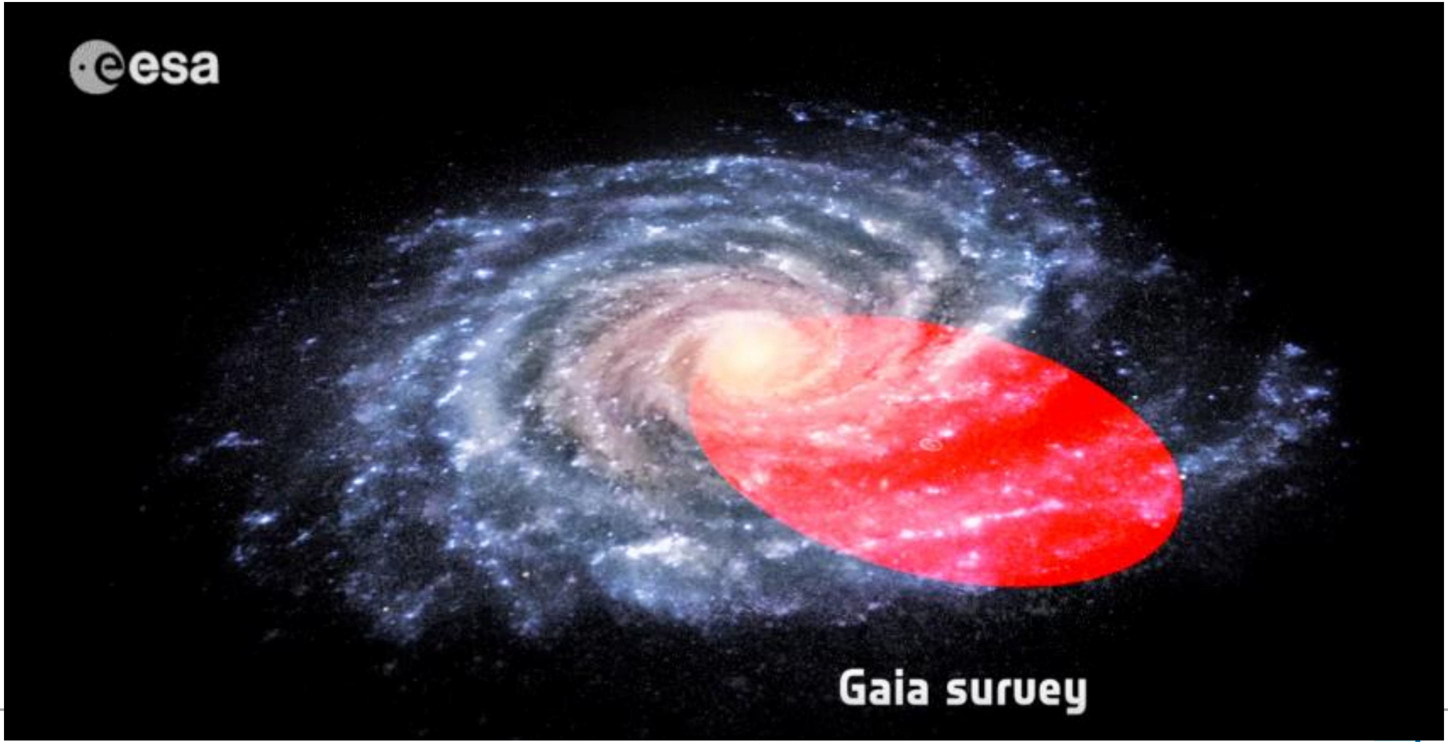
What do we know about our Galaxy?

The stars with distances known till 2016 are all in the small red spot:



What do we know about our Galaxy?

With Gaia we can explore a significant part of our Galaxy:



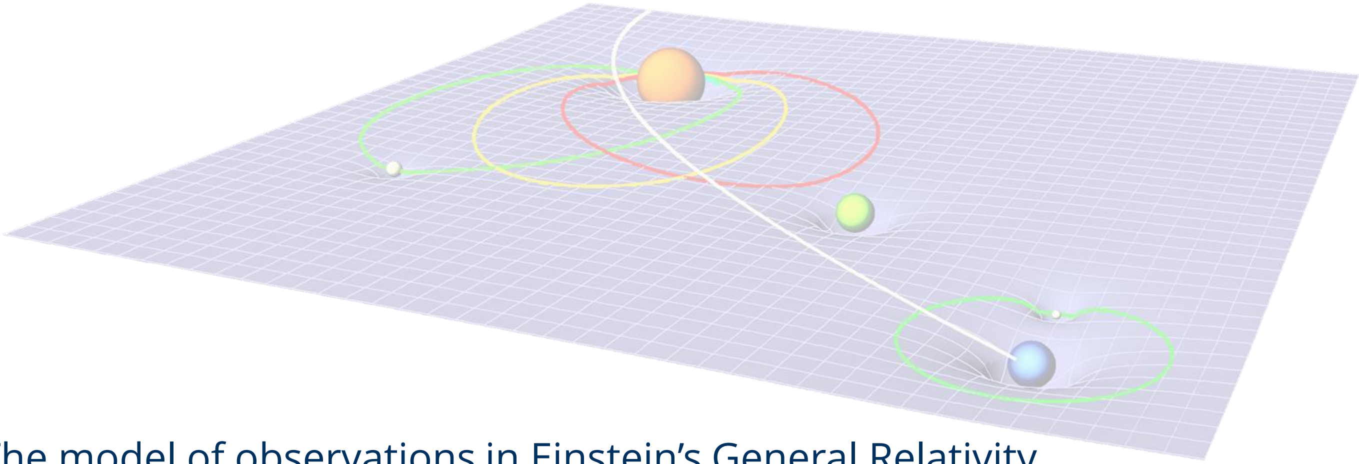
The challenge of data processing

- Parameters
  - At least 5 parameters for each star: $5 \times 1.7\ 10^{9}$
  - 4 parameters of orientation each 15 seconds: 10^{8}
  - 2000 calibration parameters per day: $4\ 10^{6}$
  - global parameters: $<10^{4}$
- Observations

  about 1000 raw images for each star: 10^{12}

- Data volume: 1 PB (iteratively!)
- Computational efforts: $\sim 10^{22}$ flops
- Direct least squares solution is impossible

Gaia in Dresden



1. The model of observations in Einstein's General Relativity
2. Tests of fundamental physical laws with Gaia data
3. Analysis of the Gaia reference frame: quasars
4. Synchronization and monitoring of Gaia's atomic clock
5. Special astrometric solutions: stability and quality verification, special calibration of the instrument, relativistic tests

Special thanks to ZIH for about 3 Million CPU-hours by now!

[Video]

Summary

Computer Science and Computational Science at ZIH

- Exciting computer science research
- Broad spectrum of computational science topics together with application science projects

Contact:

Dr. Andreas Knüpfer
Deputy Director / CTO of ZIH
andreas.knuepfer@tu-dresden.de
Tel. +49 351 463-38323
Willersbau A113