# Hochleistungsrechnen und Big Data - Technologien und Trends

- **Entwicklung**
- **Technologien**
- **Anwendung im wissenschaftlichen Umfeld**

Dr. Uwe Konrad, Leiter

Zentralabteilung Informationsdienste und Computing

Februar 2017

HZDR

HELMHOLTZ ZENTRUM DRESDEN ROSSENDORF

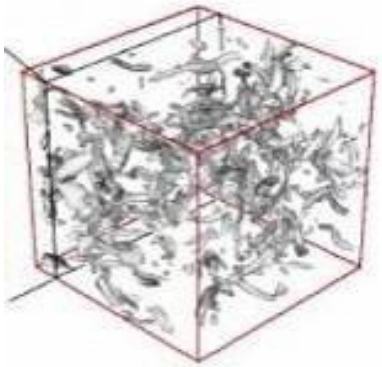
# Entwicklung – Paradigmas der Wissenschaft

- Vor tausend Jahren:
  Wissenschaft war empirisch und beschrieb i.W.
  Naturphänomene

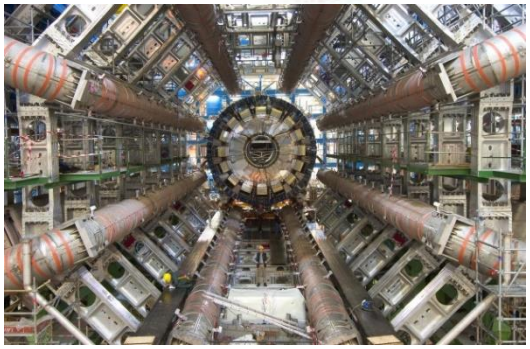
- In den letzten paar hundert Jahren:
  Es entstand ein theoretischer Zweig mit Modellen und Verallgemeinerungen

$$\oint_{\partial A} \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{s} + \frac{\mathrm{d}}{\mathrm{d}t}\left(\int_{A} \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{A}\right) = 0$$

- In den letzten Dekaden:
  Es entstand die numerische, computergestützte Simulation komplexer Phänomene

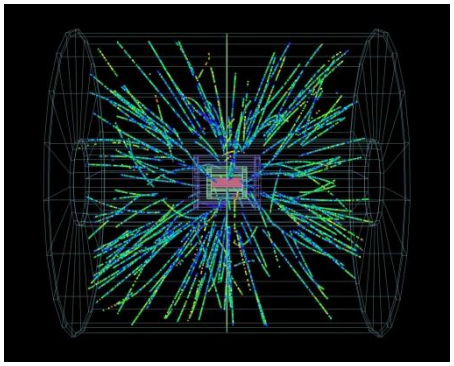


- Heute:
  Wissenschaft verwendet Daten Erkundung (eScience) zur Vereinheitlichung von Theorie, Experiment und Simulation,
  Daten werden durch Instrumente oder Simulation erfasst und durch Software weiterverarbeitet,
  Wissenschaftler analysieren riesige Datenmengen mit Hilfe von Statistik, Korrelation, Heuristik und maschinellem Lernen.



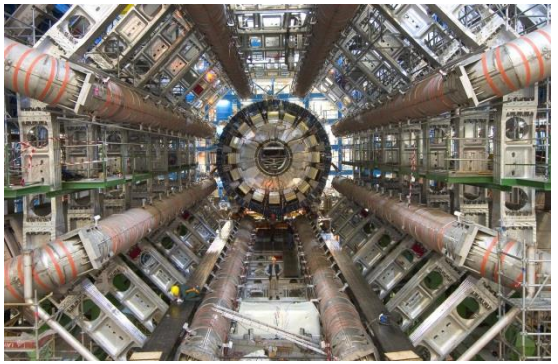
# Entwicklung: Meilensteine von „Big Data“

- **1995**: Die Mission LHC Projektes am CERN ist die Speicherung und Analyse von ca. 25 Petabyte an Daten pro Jahr, zu dem Zeitpunkt mehr als irgendwo sonst.
- **1997:** D. Ellsworth und M. Cox verwenden den Begriff “Big Data” erstmals in einem Paper über Visualisierung.
- **2008:** Google verarbeitet 20 Petabyte Daten an einem Tag.
- **2011**: IBM’s Supercomputer Watson analysiert 200 Millionen Seiten mit etwa 4TB an Daten in Sekunden.
- **2012**: Die Obama Regierung gibt eine „Big Data Research and Development Initiative“ bekannt (Budget 200 M$).
- **2014**: Rechenzentren verbrauchen ca. 2% des Stroms in USA
- **2015-** Google speichert 10.000 Petabyte Daten und verarbeitet ca. 3.5 Milliarden Anfragen täglich.
- **2015-** Amazon hat die größte Rechnerkapazität mit mehr als 1,4 Millionen Servern.



# Entwicklung: Das Large Hadron Collider (LHC) Grid Projekt

- Die Aufgabe des Computing Grid Projektes war die Bereitstellung Speicherung, Verteilung und Analyse der ca. 25 PB an jährlichen Daten zu, die durch den Large Hadron Collider (LHC) am CERN [1] generiert werden.
- Es entstand das „European Grid Project“, bestehend aus 12 Tier-1 Rechenzentren (z.B. KIT), verbunden durch ein dediz. 10 Gbit Netz und >50 Tier-2 RZs.



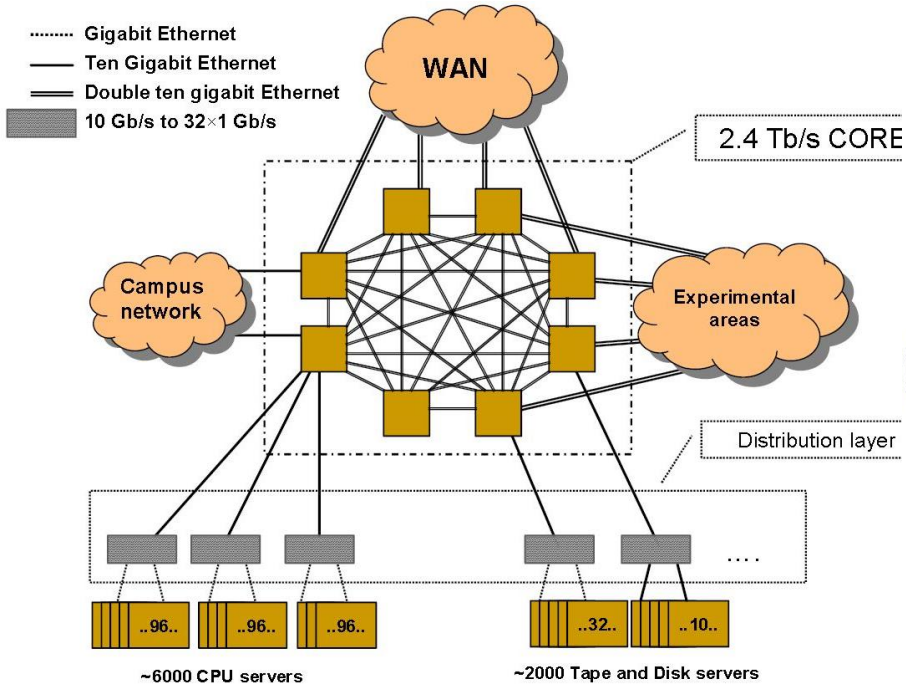



Figure 3.3: Layout of the Tier-0 network

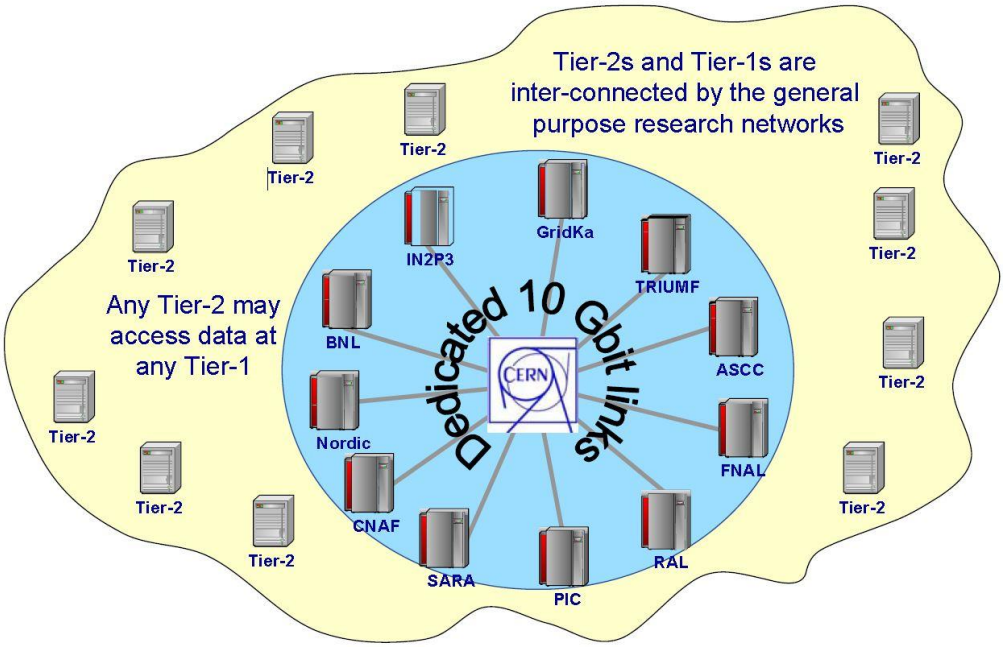



Figure 4.8: Tier-0–Tier-1–Tier-2 interconnectivity

[1] J. Knobloch, et al.: LHC Computing Grid, Technical Design Report, Cern 2005

# Technologien: High Performance Computing für die Datenanalyse

- Zur Verarbeitung großer Datenmengen verwendet man High Performance Computing (HPC) Systeme. Dies sind heute im allgemeinen Cluster, d.h. Netzwerke aus unabhängigen Rechnern.
- Im Juni 1997 erreichte das erste Cluster eine Platzierung in der Top500 Supercomputern. Heute sind 85% der Top500 Cluster. Das stärkste HPC-System war 2016 der chinesische Taihu Light mit 10,6 Mio. CPU Cores und 93 PFlop/s Leistung (15 MW elektrisch).
- Seit einigen Jahren werden Grafikkarten (GPU) zur Beschleunigung der Berechnungen eingesetzt. Der drittstärkste Rechner der Welt (Titan in Oak Ridge) hat 0,5 Mio. CPU Cores und 18.600 GPU Tesla K20 Grafikkarten zur Beschleunigung.
- Der stärkste deutsche Rechner steht derzeit in Stuttgart am HLRS und hat 0,2 Mio. CPU cores und 5,6 PFlop/s (#14).

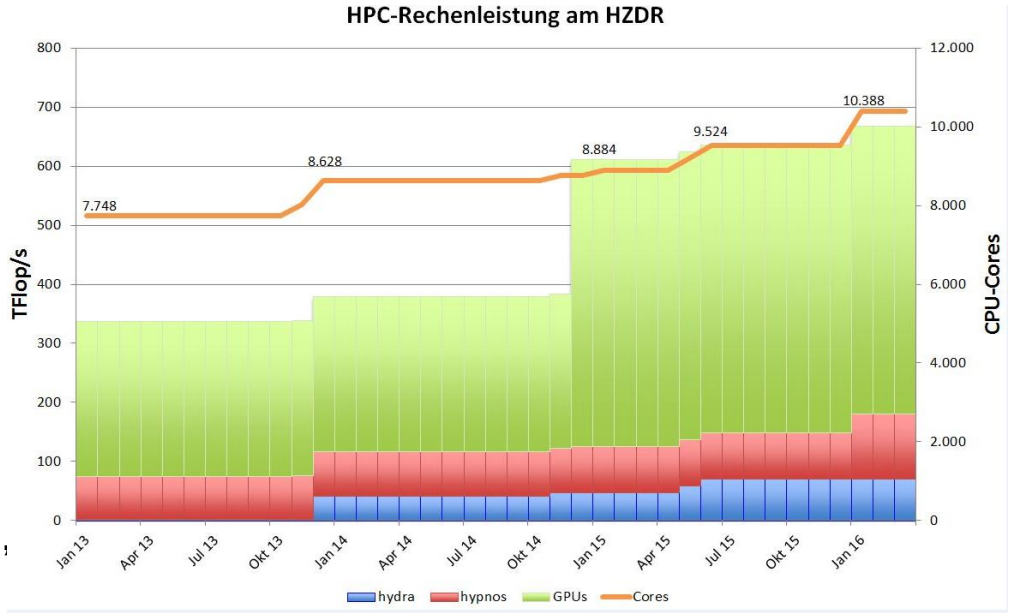
#1: Taihu Light: 10,6 Mio. CPU cores, 93 PFlop/s

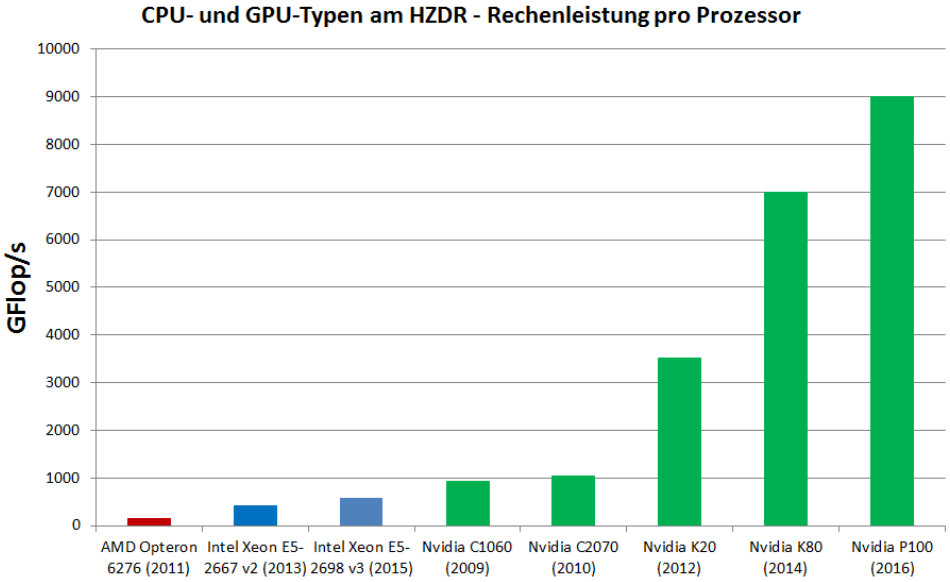
#3: Titan (USA): 0,6 Mio CPU cores, 18.600 GPU, 18 PFlop/s


HZDR: 0,01 Mio. CPU cores, 164 GPU, 0,6 PFlop/s

# Technologien: HPC am HZDR

- Das HZDR besitzt ein tier-3 Rechenzentrum (No.2 in Sachsen!)
- Innerhalb von knapp 5 Jahren hat sich die Rechenleistung um den Faktor 70 erhöht!
- Die Anzahl der CPU-cores erhöht sich auf 10.000 (2016).
- Neben klassischen CPUs werden seit 2010 GPUs zum Rechnen genutzt (grün, 2016: 164 GPU mit 360.000 cores).




- Die Anzahl der Nutzer aus den Instituten des HZDR (FWK, FWI, FWD, FWO, FWH) hat sich auf ca. 100 verdreifacht.
- Rechenaufgaben mit Spitzenanforderungen werden in Zusammenarbeit mit Höchstleistungs-RZs (u.a. Jülich) gelöst
- Das erhebliche Leistungswachstum führte dazu, dass die Kapazität des RZ 2012 erheblich erweitert werden musste!

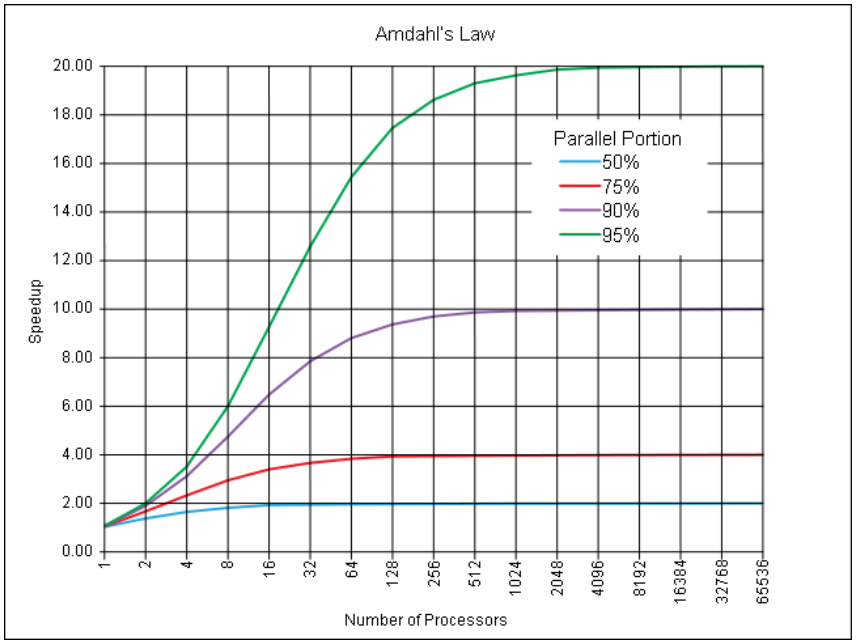
# Technologien: Grenzen der Parallelisierung der Datenverarbeitung

- Notwendige Berechnungen bei der Datenverarbeitung sind heute üblicherweise parallel, d.h. werden von vielen Prozessoren gleichzeitig ausgeführt.
- Idealerweise ist die Berechnungszeit umgekehrt proportional zur Anzahl der verwendeten Prozessoren (Speed-Up). Es gilt jedoch das Gesetz von Amdahl

$$S = \frac{T_1}{T_N}$$

$$S = \frac{1}{(1-P) + \frac{P}{N}} \leq \frac{1}{1-P}$$

S= Speedup, P= Parallel Portion N= Number of Processors




- Schlussfolgerungen aus dem Gesetz von Amdahl:
  - Effizienz der eingesetzten Algorithmen ist entscheidend
  - nicht-parallelisierbare (sequentielle) Programmteile begrenzen die Skalierung
  - Hardware muss intelligent und effizient genutzt werden, guter Code für die Parallelisierung auf GPUs kann nahezu linear skalieren

# Technologien: Grenzen der Parallelisierung der Datenverarbeitung (2)

- Eine andere Sicht: Mehr Prozessoren bearbeiten auch mehr Daten!
- Wenn die zu verarbeitende Datenmenge im gleichen Maße wie die Anzahl der verwendeten Prozessoren wächst, bleibt der Speedup konstant. (Gustafsonsches Gesetz)

$$S = (1 - P) + N \cdot P$$

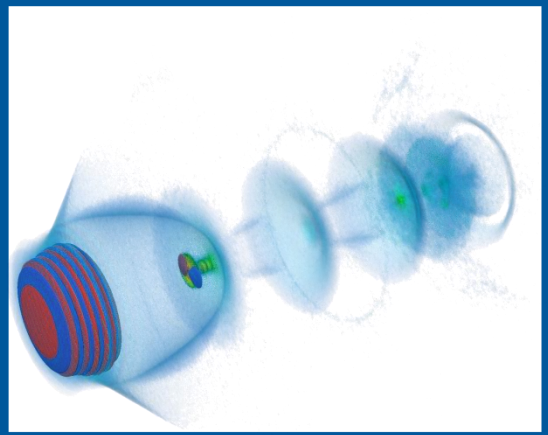
- Die Wahrheit liegt dazwischen
  - Amdahl: starke Skalierung (Gesamtproblem konstant; kleineres Problem pro Prozessor bei mehr Prozessoren)
  - Gustafson: schwache Skalierung (Problemgröße pro Prozessor konstant; Gesamtproblem wächst mit Prozessoranzahl)

# Technologien: Parallelisierung in der Praxis - PIConGPU

PIConGPU

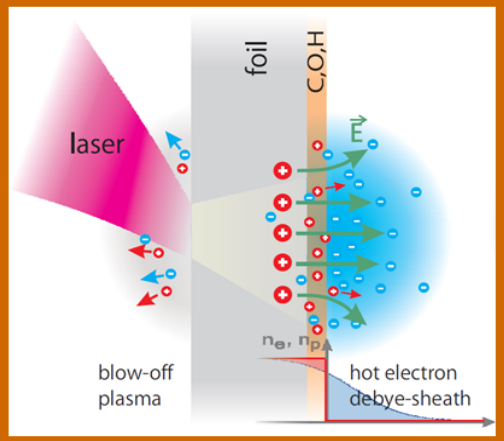
## Electron Acceleration with Lasers

- Compact X-Ray sources



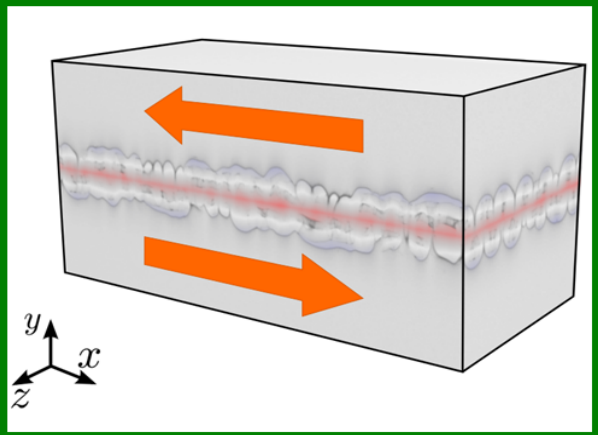
## Ion Acceleration with Lasers

- Tumor Therapy




## Plasma Instabilities

- Astrophysics



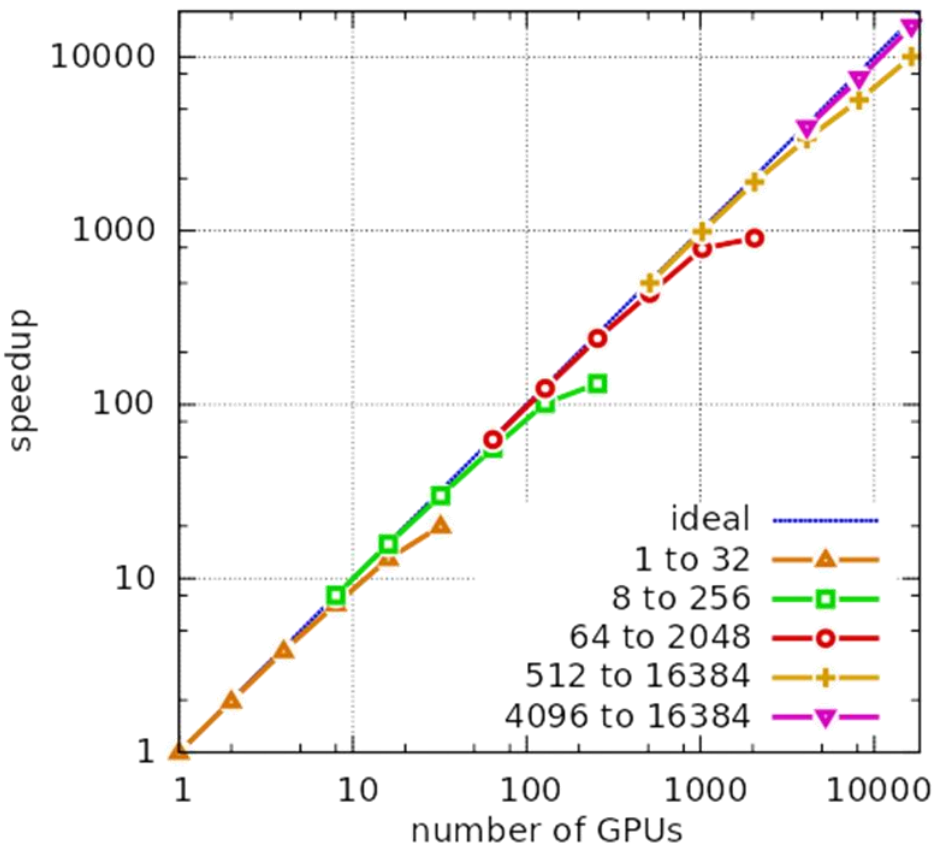

PIConGPU ist ein HZDR code für „particle in cell“ simulationen, der für GPU optimiert ist.

# Technologien: Parallelisierung in der Praxis – PIConGPU Speedups

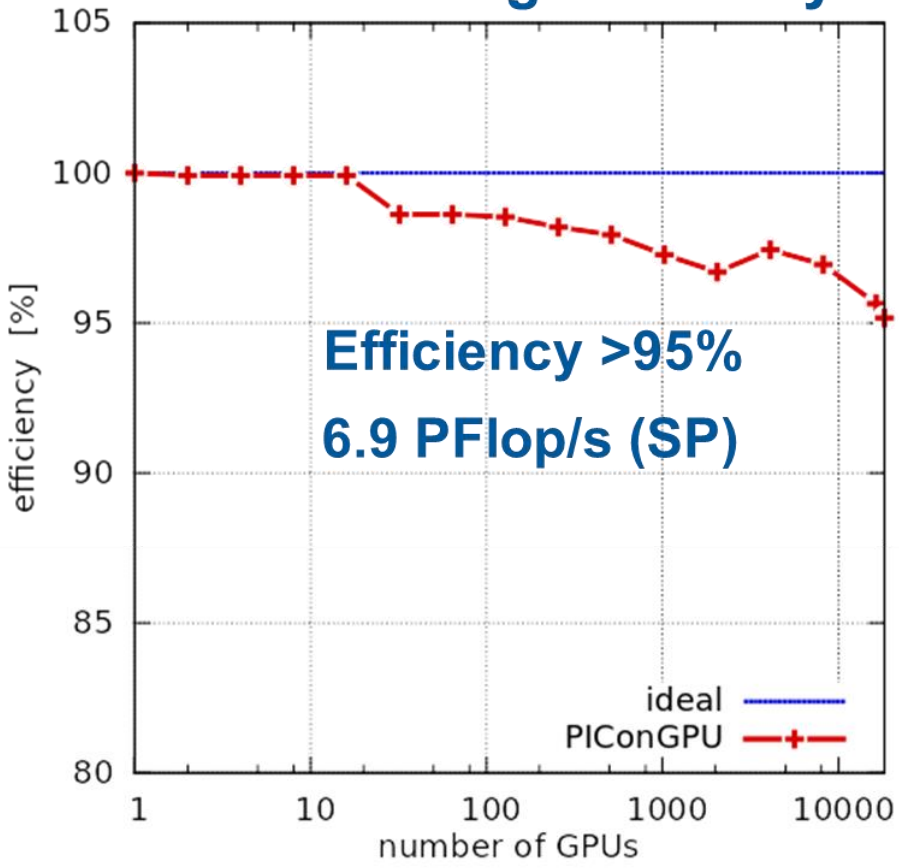
PIConGPU — Scales up to **18,432** GPUs

OAK RIDGE National Laboratory

## strong scaling




## weak scaling efficiency



# Technologien: Schnelle Datenübertragung

Netzwerke für die Datenübertragung müssen hohe Bandbreiten und niedrige Latenzzeiten bieten, dafür gibt es verschiedene Technologien:

- Ethernet Netzwerke (für Intranetze)
  - Bandbreite: Heute üblich 10Gb (Gbit/s)
  - Verfügbar sind auch 40 Gb- und 100 Gb-Ethernet Netzwerke
  - Latenz (gemessen in s): bei 10GbE im Bereich 10 µs
  - Hersteller: Cisco, HP, Extreme, Juniper, Brocade u.v.a.
- InfiniBand Netzwerke (für HPC Systeme)
  - Bandbreite: 56 Gbit/s am HZDR, 100 Gbit/s verfügbar, demnächst 200 Gbit/s
  - Latenz: <1 µs
  - Hersteller: Mellanox, QLogic
- Fibre-Channel Netzwerke (für Speichernetzwerke)
  - Bandbreite: 8/16 Gbit/s am HZDR, verfügbar 32 und 128 Gbit/s
  - Latenz: <1µs
  - Hersteller: Brocade, Cisco, QLogic
- Dateitransfer
  - Dateiformate und - Bibliotheken wie HDF5 unterstützen das schnelle, parallele Übertragen, Verwalten, Visualisieren and Analysieren von Daten

# Technologien: Storage Hardware

Datenspeichersysteme müssen hohe Kapazitäten und hohe Zugriffsgeschwindigkeiten bieten:

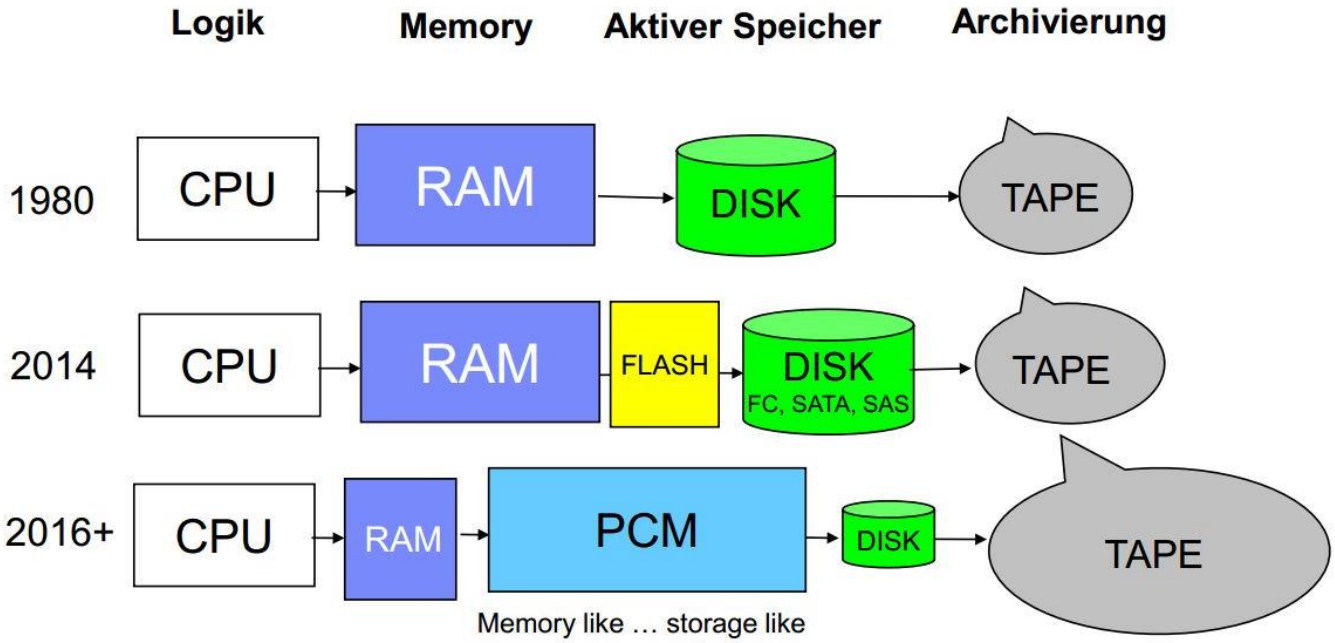
- Kapazität (gemessen in Bytes):
  - liegt heute im Bereich von einigen PetaBytes mit Trends in den ExaByte-Bereich
- Anforderungen wachsen exponentiell
  Zugriffsgeschwindigkeiten (gemessen in Bytes/s, IOPS):

| | |
|---|---|
| Einzelne Magnetplatte: | 120 MB/s, wenige 100 IOPS |
| Bandlaufwerk: | 300 MB/s, wenige Dateien schreibend |
| SSD-Platte: | 500 MB/s, bis zu 10.000 IOPS |

- am HZDR installiertes GPFS Speichersystem (basierend auf 720 handelsüblichen Festplatten): 2 PB, 25 GB/s, über 200.000 IOPS
- Technisch noch möglich: TB/s durch Parallelisierung von Komponenten

# Technologien: Storage Hardware

Neuordnung der Speicherhierarchie (Quelle IBM):
-> Hat die Festplatte hat ausgedient? Hat das Magnetband ausgedient?



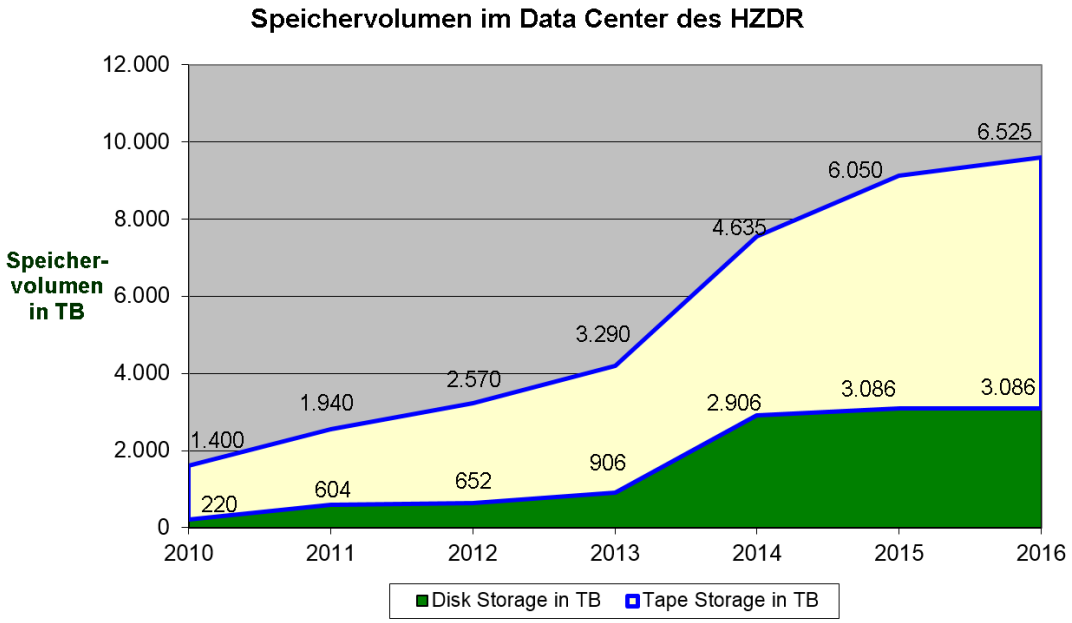

<u>Top Technologie derzeit:</u>

- Hard Disk Seagate: 10 TB ab 2016 (550$)
- SSD Disk Seagate: 60 TB (NAND Flash) ab 2017 (10.000 $?)
- Tape IBM: 32 TB mit LTO8 geplant 2018 (150 $?)

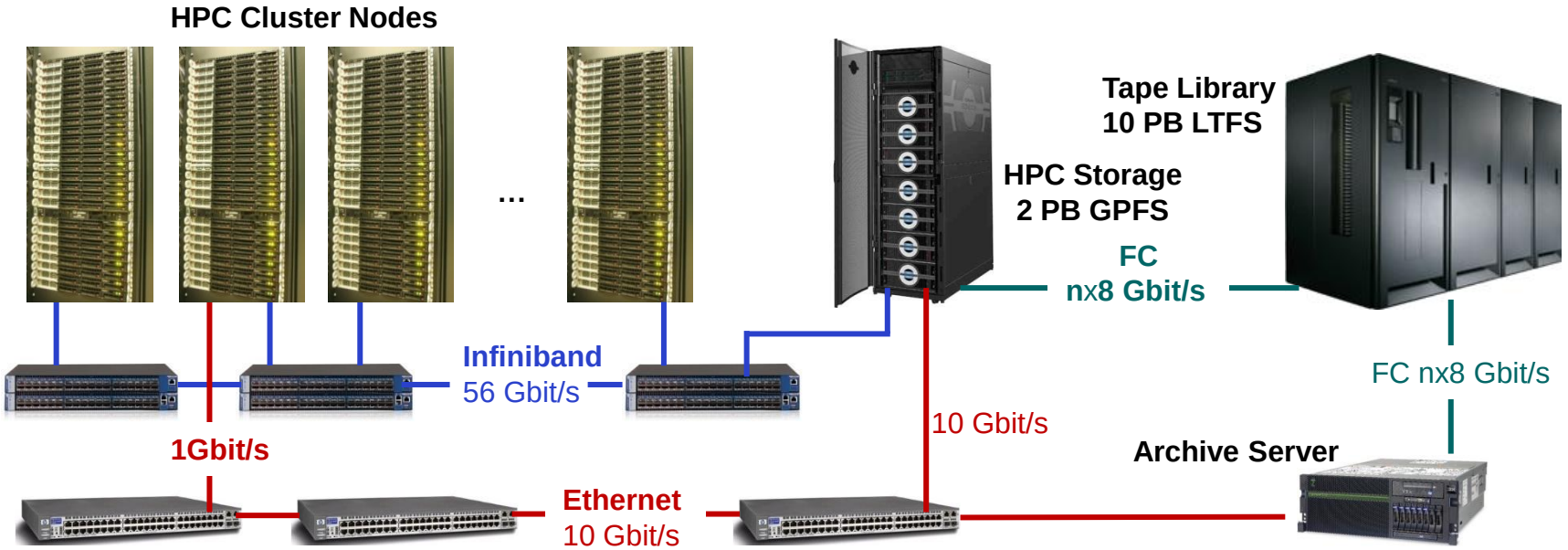
- Und danach? HP Memristor Memory Chip (nach 2018)?

# Technologie: Datenspeicherung am HZDR

- Die Kapazität an schnellem Speicher begrenzt die Simulationsperformance
- Die Kapazität für die Archivierung der Daten auf Bänder ist da aber die Anzahl Files (~500 Mio.) ist ein wesentliches Problem.
- Das Datenaufkommen reicht von 500 TB/w bis zu mehreren PB/a




- HPC Architektur am HZDR:



# Technologie: Verteilte Parallele Filesysteme

- 1998: IBM General Parallel File System (GPFS), es wird z.B. am HZDR und am JUGENE Supercomputer in Jülich wird verwendet, um den hohen Speicher- und Bandbreitenbedarf zu bewältigen.
  -> 2015: 120 PB für ein mp3-Archiv, HZDR: 2 PB


- 2003: Google File System (GFS) ist ein proprietäres, verteiltes Filesystem, das durch Google für eigene Zwecke entwickelt wurde.
  -> 2008: geschätztes Volumen von 400 PB/a


- 2003: CERN (CASTORFS) Filesystem zum Zugriff auf die Experimentdaten des LHC,
  -> Bis 2015 wurden ca. 500 PB gespeichert


- 2003: Lustre Filesystem, zuerst verwendet am Lawrence Livermore National Laboratory (USA), of verwendet auf Supercomputern
  -> Oak Rigde National Lab: 40 PB mit 1.4 TB/s


- 2005: Hadoop™ Distributed File System (HDFS) ist ein hochverfügbares Filesystem um riesige Datenmengen zu speichern. Hadoop wird u.a. bei Facebook (2012: 100 PB), Yahoo! und Twitter verwendet.

# Technologie: Datenanalyse und Visualisierung

- **Schnelle Datenanalyse**:
  Wird u.a. durch Methoden von modernen Filesystemen unterstützt, die ständig weiterentwickelt werden:
  Map-Reduce (Hadoop), Transparente Kompression, Directed-Acyclic-Graph (Flink),

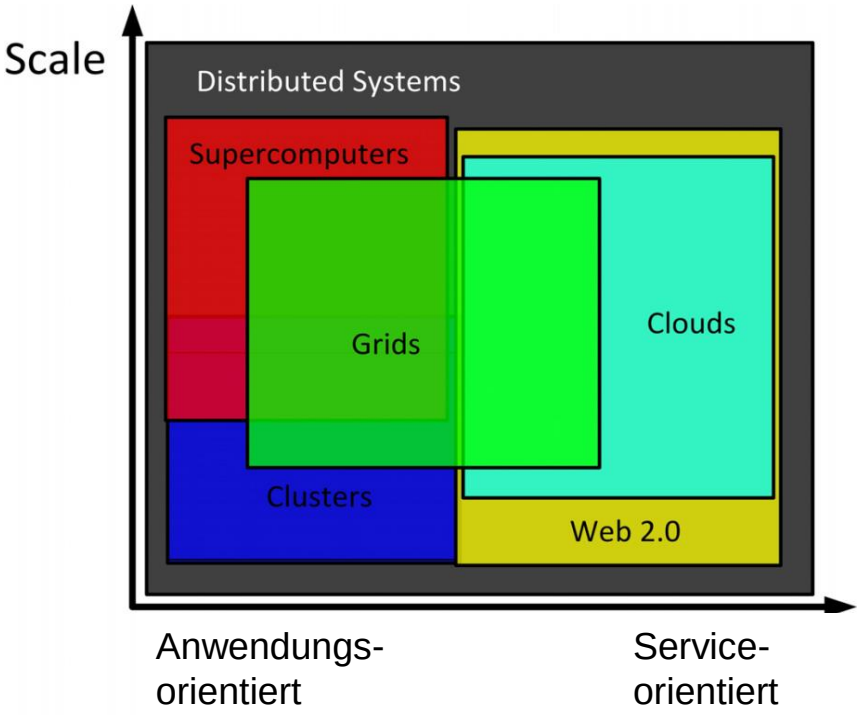
- **Intelligente Methoden**:
  Machine-Learning, Computer Vision (z.B. Objekterkennung), Real-time Processing, Model Reduction (multi-dimensional data sets)

- **Visualisierung und Feedback**
  4D-Visualisierung (auch in-situ), Interaktives Feedback (siehe Bild unten), Explorative System Visualisierung


# Technologien: Cloud Computing

- Cloud Computing ist ein spezielles Modell des verteilten Computings, das von anderen abweicht:

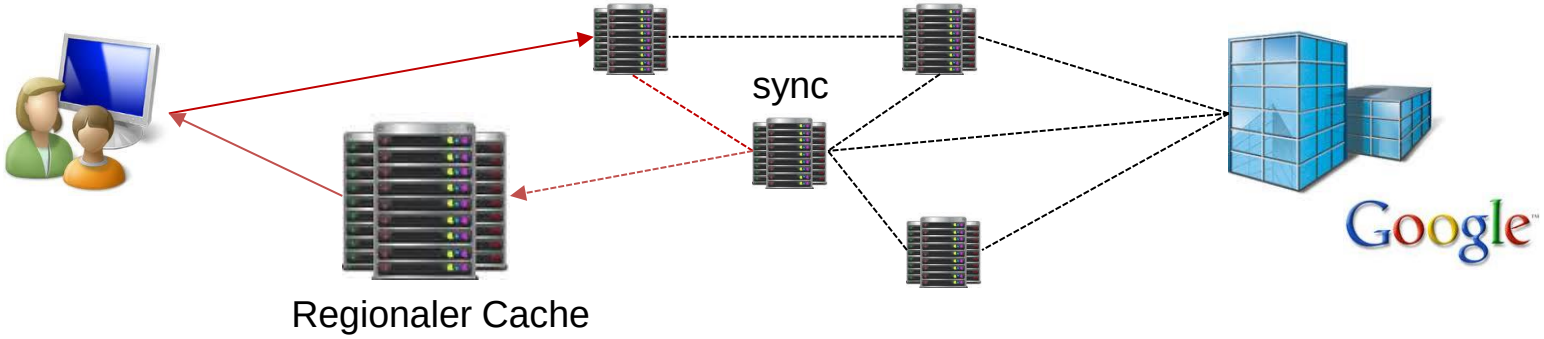
  1) es ist massiv skalierbar,
  2) es wird als abstrakte Einheit gekapselt und bietet verschiedene Services-Levels für Kunden außerhalb der Cloud,
  3) es wird getrieben durch Größenvorteile (economy of scale),
  4) die Services können dynamisch konfiguriert werden (durch Virtualisierung oder andere Lösungen) und werden nach Bedarf geliefert.



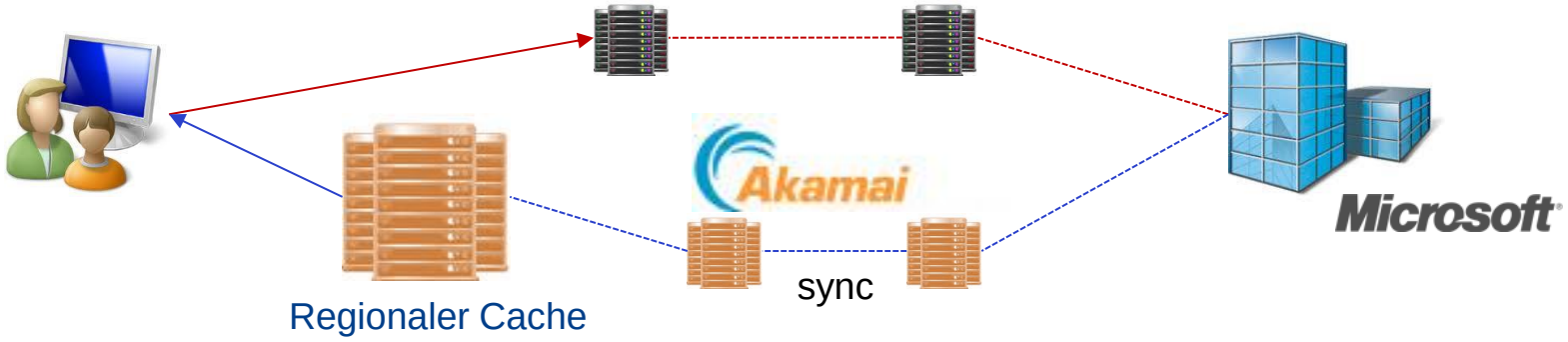

- Cloud Computing wird durch verschiedene Servicemodelle (Infrastruktur, Plattform oder Software) und Einsatzmodelle (Private Cloud, Public Cloud, Private-Public) charakterisiert.
- Systeme wie OpenStack (open source) implementieren Cloud Computing Modelle.

# Technologien: Performance für weltweite Cloud-Services

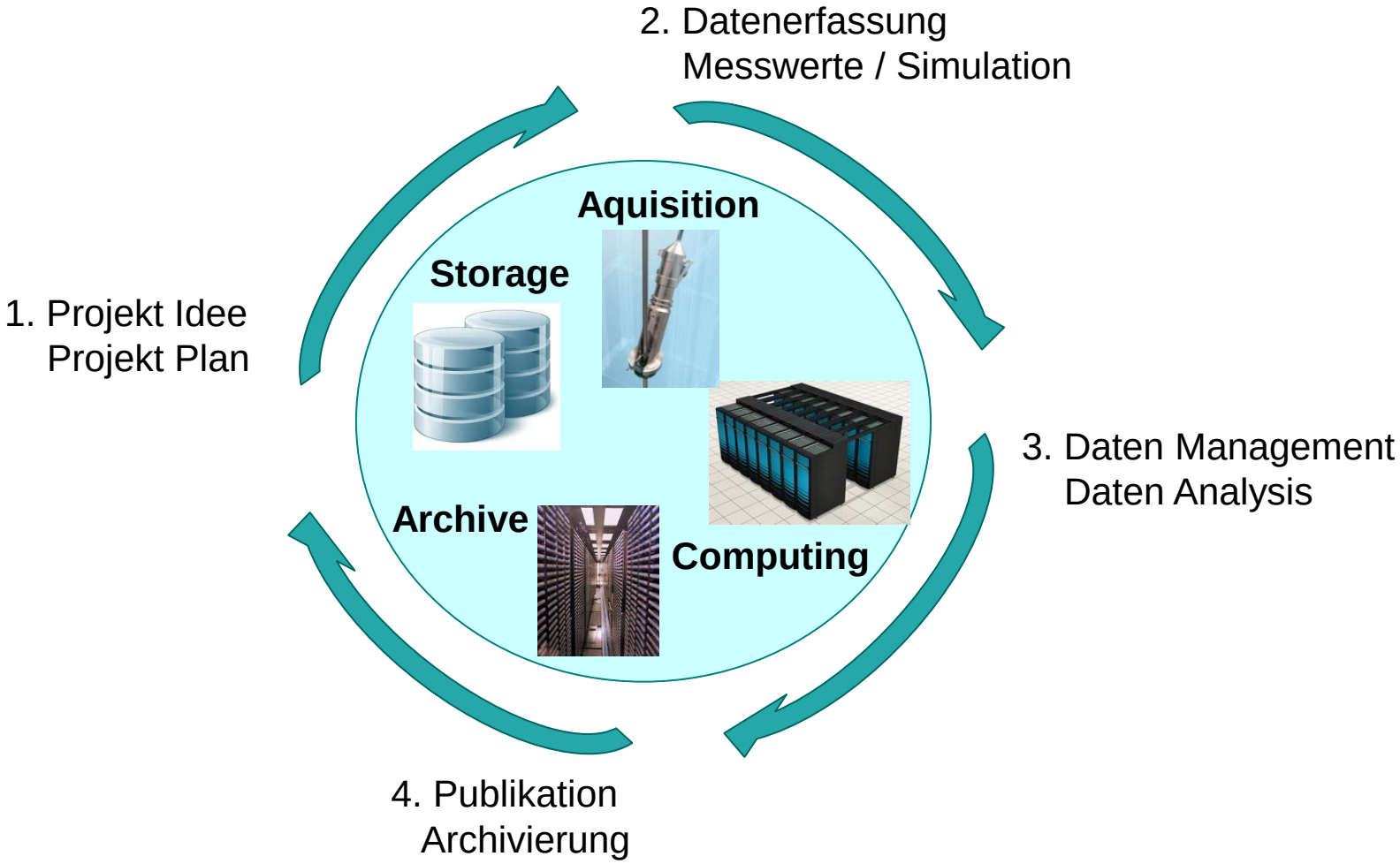
- Theoretisch benötigt eine Nachricht ~200 ms für eine Kabelstrecke von 20.000 km, aber in der Realität ist es viel mehr. Für geeignete Reaktionszeiten müssen die Daten in der Nähe der Nutzer gespeichert sein, dazu sind Cache-Datenzentren erforderlich. Es gibt verschiedene Modelle, um solche cache-Netzwerke zu schaffen:
- Google hat ein gigantisches Netzwerk von tausenden Datenzentren aufgebaut, um Antwortzeiten von <200 ms zu erreichen (Google Instant):




- u.a. Microsoft greift dazu auf Internet-Dienstleister wie z.B. Akamai Technologies zurück



# Wissenschaftliches Umfeld: Lebenszyklus der Daten




- Über den Lebenszyklus müssen ein transparenter Zugriff sowie die Reproduzierbarkeit und Kontrolle der Ergebnisse gewährleistet sein. Dabei gibt es am HZDR jeweils konkrete Herausforderungen, die eine enger Zusammenarbeit der Projektpartner und des Rechenzentrums erforderlich machen!

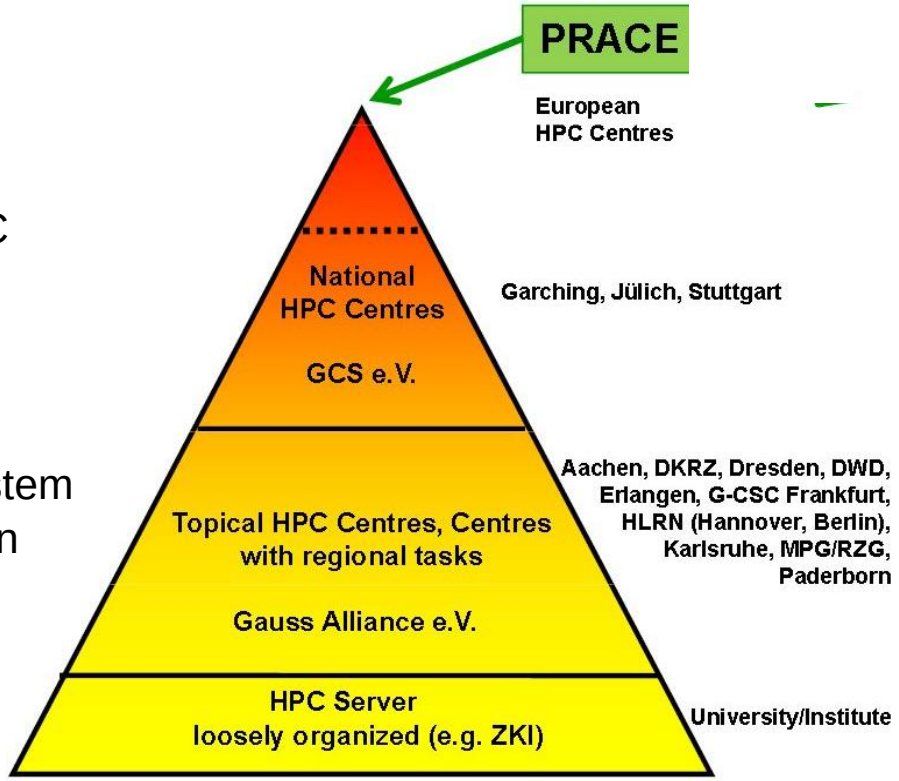
# Wissenschaftliches Umfeld: Datenmanagement und Open Access

- Im Umfeld der „Offenen Wissenschaft“ (Open Science) fordern die nation. und internat. Geldgeber ein Datenmanagement nach den FAIR-Prinzipien: -> Findable, Accessible, Interoperable und Reproducible [2].
- Für die Nachnutzung von Forschungsdaten ist es notwendig, Metadaten zu erfassen sowie den Entstehungskontext und die benutzten Werkzeuge bzw. Software zu dokumentieren. Die Definition erfolgt in internationalen Organisationen wie der Research Data Alliance (RDA) .
- Die nachhaltige Nutzbarmachung von Forschungsdaten bedarf eines sichergestellten Qualitätsmanagements, das den gesamten Lebenszyklus umfasst.
- Die zitierbare Datenpublikation ermöglicht eine nachvollziehbare wissenschaftliche Anerkennung und die Reproduzierbarkeit von darauf aufbauenden Untersuchungen.
- Über internationale Organisationen wie datacite.org werden weltweit eindeutige Objekt-Identifikatoren vergeben. Verlage fordern und verwenden die Nutzung dieser Identifikatoren.
- Über internationale Organisationen wie orcid.org werden eindeutige Personen-Identifikatoren für die Wissenschaftlern vergeben.

[2] European Union: H2020 Programme, Guidelines on FAIR Data Management in Horizon 2020

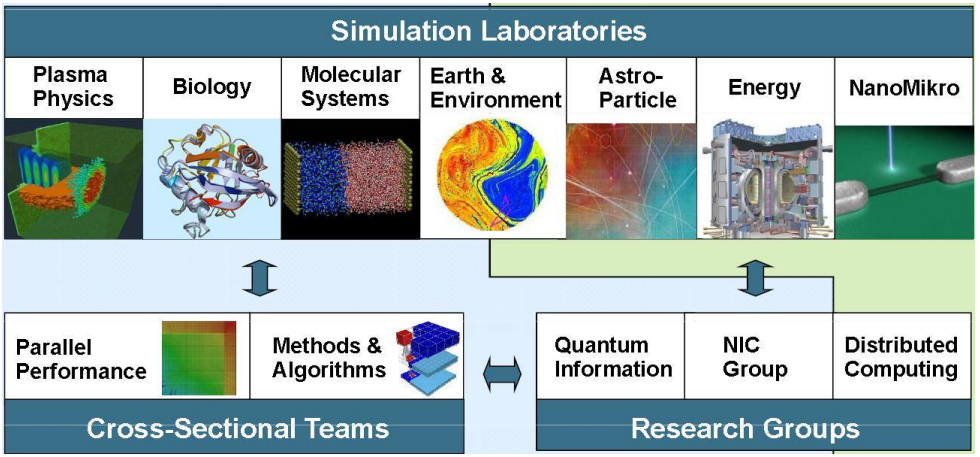
# Wissenschaftliches Umfeld: Pyramide des Supercomputings

- Die Pyramide der High Performance Computing Zentren in Deutschland besteht aus 3 nationalen HPC Zentren, einem Satz von regionalen HPC Zentren und einer großen Anzahl an lokalen HPC Einrichtungen.




- Der Supercomputer am HZDR ist ein lokales System mit ~10.000 CPU cores. Für sehr große Aufgaben nutzen die HZDR Wissenschaftler die nationalen HPC Zentren in Jülich und Garching.

- Die HGF Simulation Labs wurden geschaffen, um die Wissenschaftler in ihren Simulation-Aufgaben zu unterstützen. Dieser Support wird durch HPC-Experten gegeben.

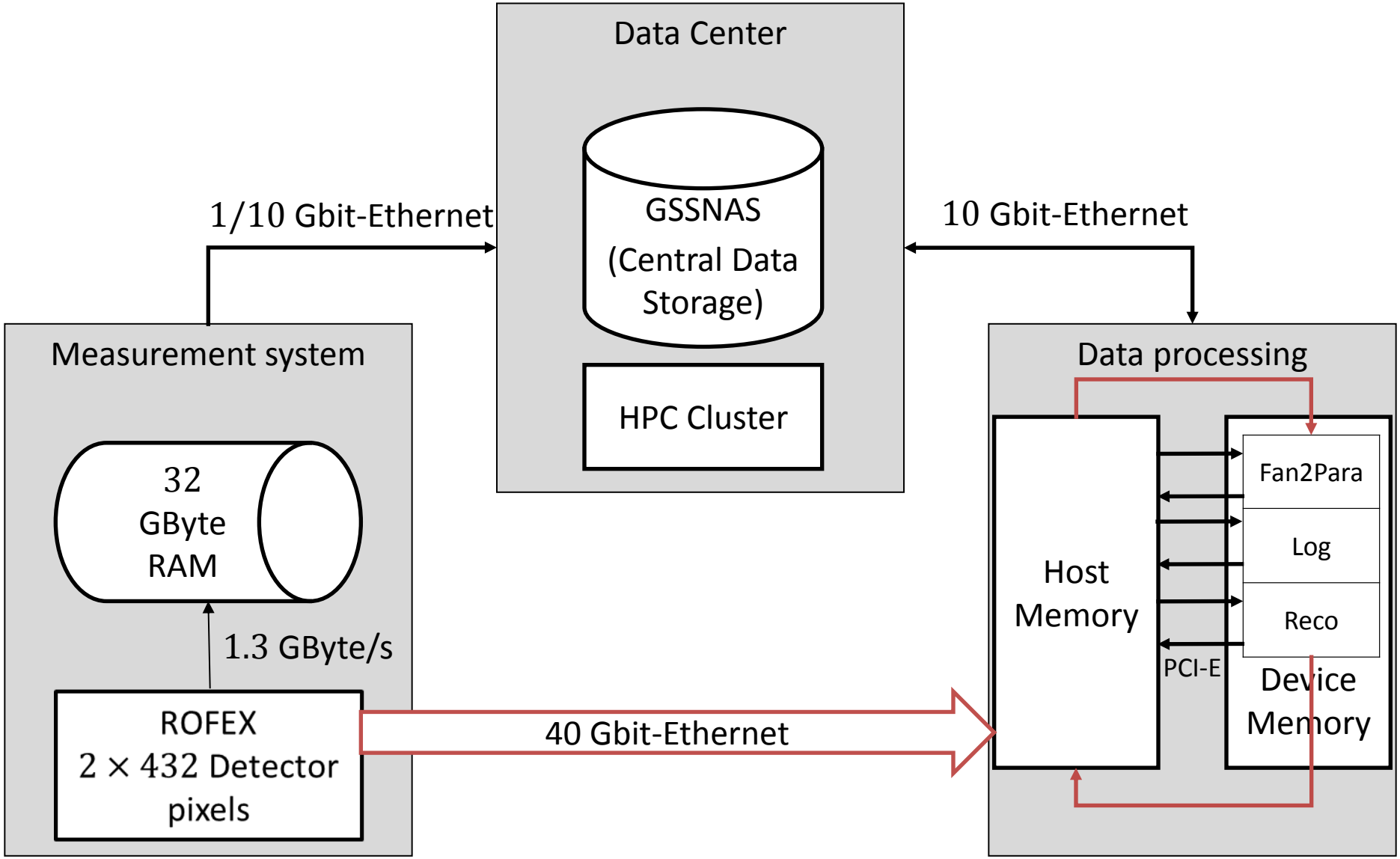


# HZDR: Datenanforderungen aus wissenschaftlichen Projekten

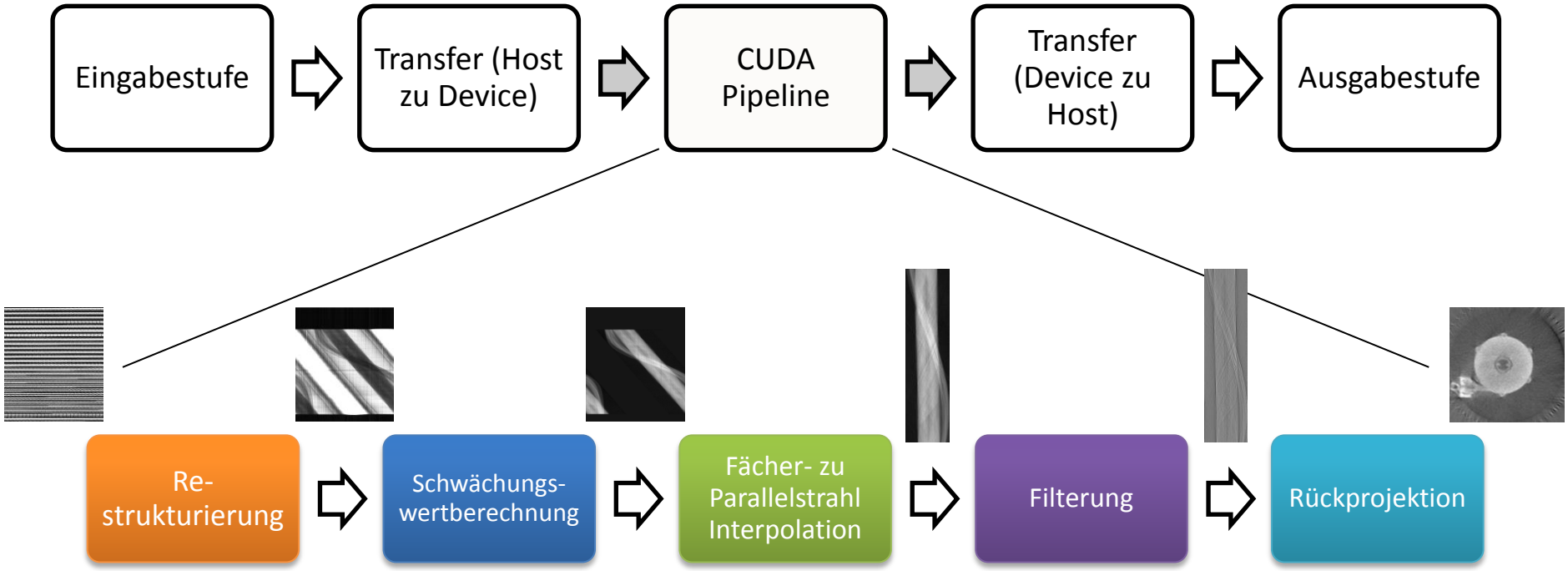
- Die Menge der Experimentdaten wird von weniger als 1 PB/a auf mehr als **20 PByte/a** innerhalb der nächsten Jahre steigen. Dies basiert auf neuen Generationen von Sensoren und hohen Abtastraten (MHz anstelle von kHz).
- So liefert das ROFEX III Experiment jetzt schon 0,5 PB/a mit einer Datenrate von 2,5 GB/s. Ab 2019 wird das THz-Experiment TELBE ca. 13 PB/a mit einer Datenrate von ca. **10 GB/s** produzieren.
- Durch den Einsatz von GPU Knoten an ROFEXIII zur Bildverarbeitung am Experiment (dieses läuft ca. 30s), die Optimierung der Speicher- und Netzwerk-komponenten und die direkte Kopplung zum HPC-System des HZDR im Rechenzentrum konnte die Zeit für die Bereitstellung der Ergebnisdaten von mehreren Stunden auf unter 1 min gesenkt werden.

  ->Nun sind Beobachtungen quasi in Echtzeit möglich!

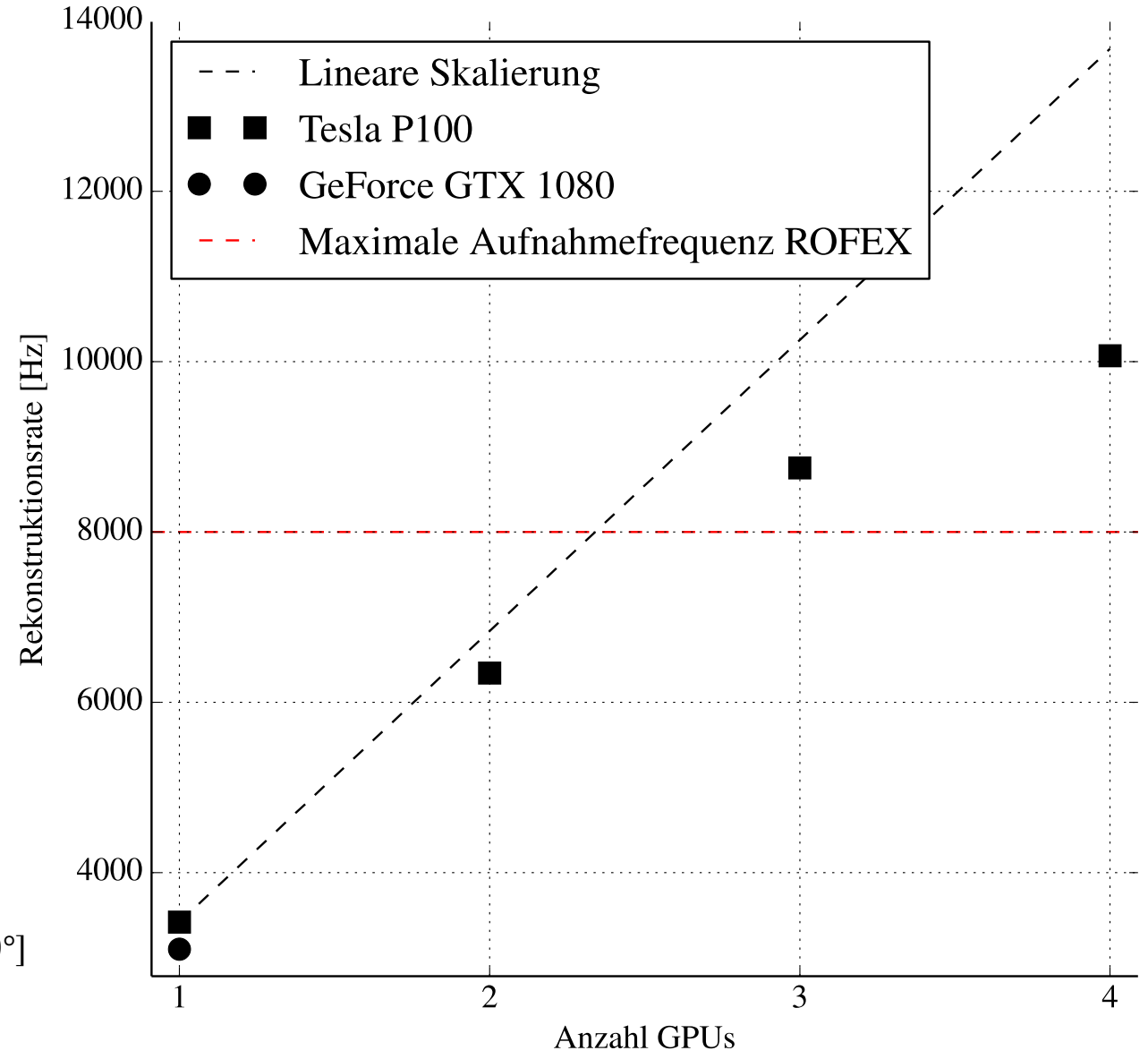
# HZDR: Datenfluss am ROFEX Experiment



# HZDR: Datenverarbeitungs-Pipeline der ROFEX Daten



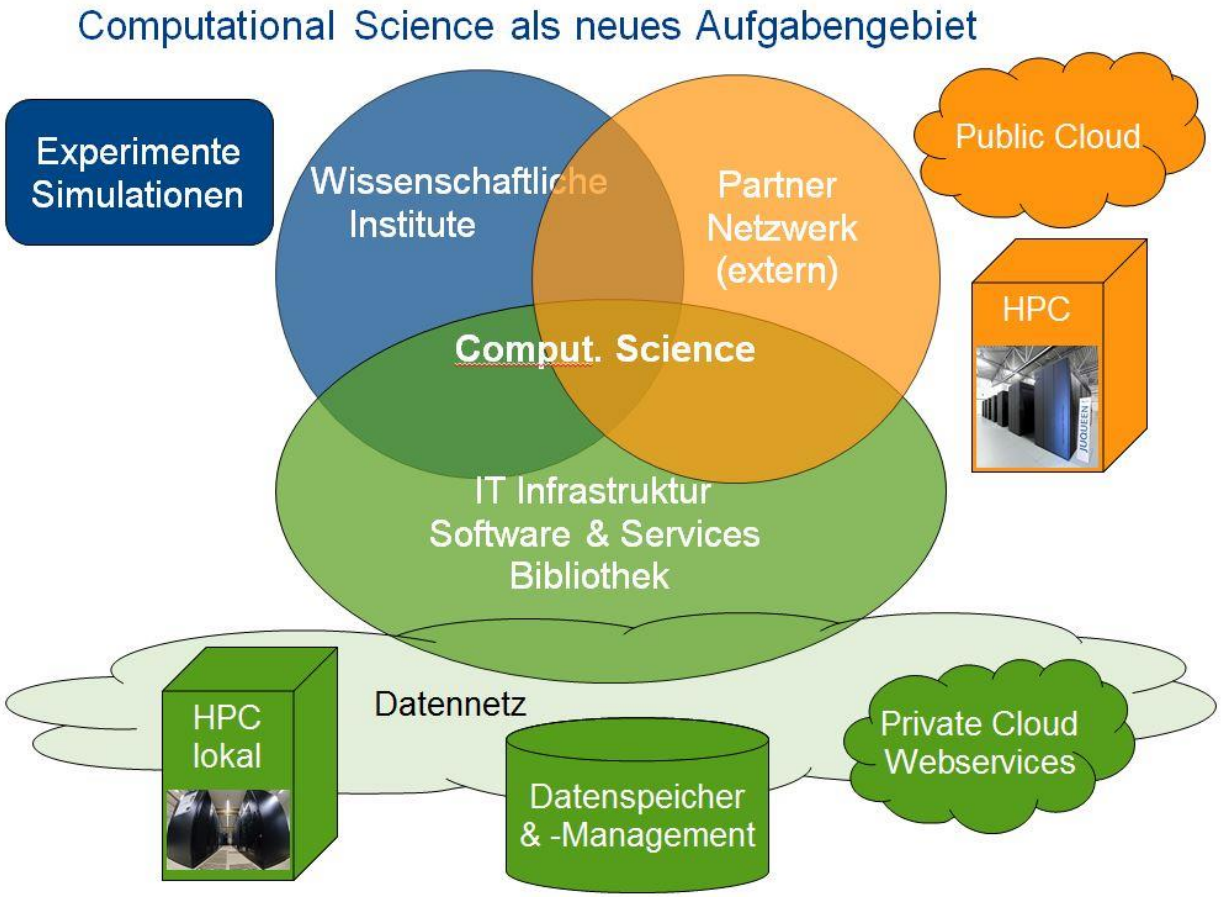
# HZDR: ROFEX Datenverarbeitungsgeschwindigkeit




$N_{\mathrm{parDet}} = 256$
$N_{\mathrm{parProj}} = 512$ [0° bis 180°]
$N_{\mathrm{pixel}} = 256$

# HZDR: Aufgaben und Organisation

- Die Rolle der Informationstechnologie ändert sich, neue Aufgaben an der Schnittstelle zur Wissenschaft stehen an. Dabei spielt auch die Bibliothek als zukünftiger Verwalter von Daten eine wichtige Rolle!
- Dafür wurde 2015 eine neue Organisationsstruktur geschaffen: „Informationsdienste und Computing“ mit vier Abteilungen



# Zusammenfassung

- Wissenschaftlichen Daten gewinnen in der Forschung zunehmend an Bedeutung, die Politik hat dies erkannt. Die Geldgeber fordern und fördern ein systematisches und offenes Datenmanagement. In den USA und GB wird dies schon seit mehr als 5 Jahren praktiziert.
- Die Verarbeitung von großen wissenschaftlichen Datenmengen erfolgt auf High Performance Computing Systemen, die eng an die Datenspeicher gekoppelt sind (Data Intensive Computing). Diese müssen dafür optimiert sein.
- Kompetenz im Management und in der Analyse von Daten sichert eine starke Position in der wissenschaftlichen Community, Datenveröffentlichungen und Datenjournale können zitiert werden und sind international anerkannt (Bsp. earth-system-science-data.net).
- Die Rolle der Rechenzentren und Bibliotheken ändert sich, sie werden stärker in den gesamten Lebenszyklus des Datenmanagements (von der Planung bis zur Veröffentlichung) integriert
- Voraussetzung sind personelle Ressourcen sowohl für das Data Management, als auch für Data Analytics. Neue Berufsprofile wie der „Data Scientist“ oder der „Data Librabrian“ entstehen, Experten werden überall händeringend gesucht.
- Was wir bisher sehen ist nur die Spitze des Eisbergs (das Internet der Dinge und der Menschen)!