# DKRZ
# Data Management Services

## MPG Workshop
## Forschungsdatenmanagement
## 19-April-2018

Hannes Thiemann
Deutsches Klimarechenzentrum GmbH (DKRZ)

# DKRZ

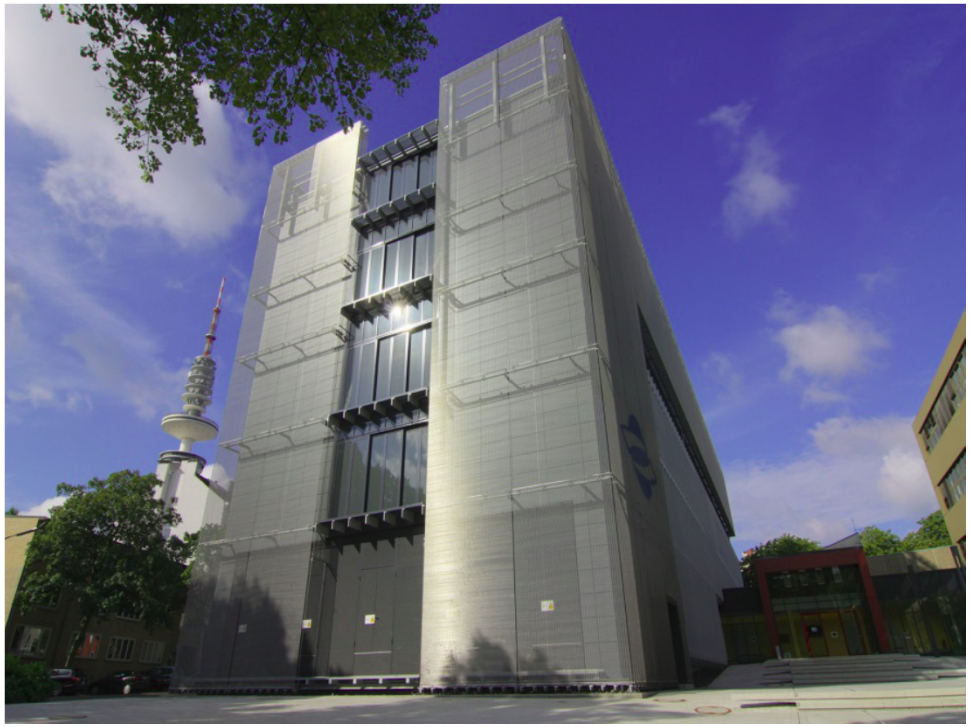
Das DKRZ ist eine gemeinnützige, nicht kommerzielle GmbH mit vier Gesellschaftern:

- der Max-Planck-Gesellschaft
- der Freien und Hansestadt Hamburg (Universität Hamburg)
- dem Alfred-Wegener-Institut für Polar- und Meeresforschung
- dem Helmholtz-Zentrum Geesthacht

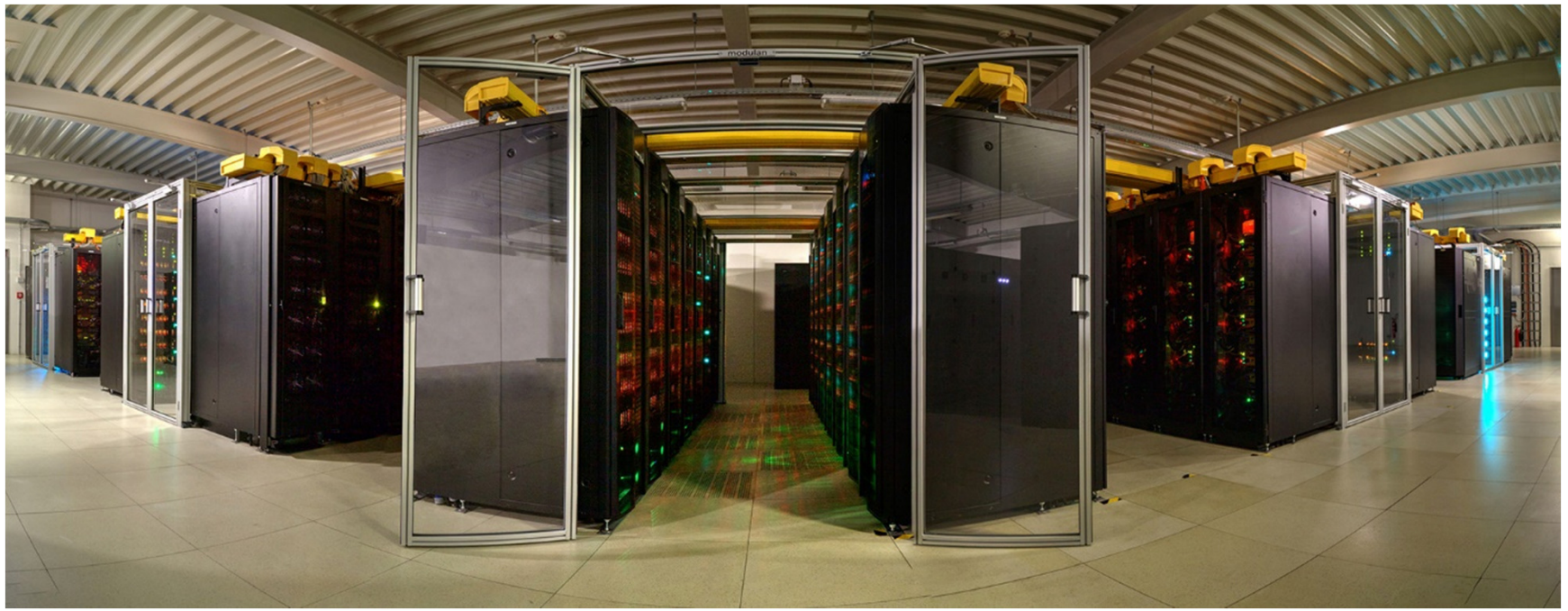
Das DKRZ ist eine zentrale nationale Service-Einrichtung für die Klima- und Erdsystemforschung. Seine Hochleistungsrechner, Datenspeicher und Dienste bilden die zentrale Forschungsinfrastruktur für die simulationsbasierte Klimawissenschaft in Deutschland.

Personal

- 70+ DKRZ
- 10+ Uni, Forschungsgruppe



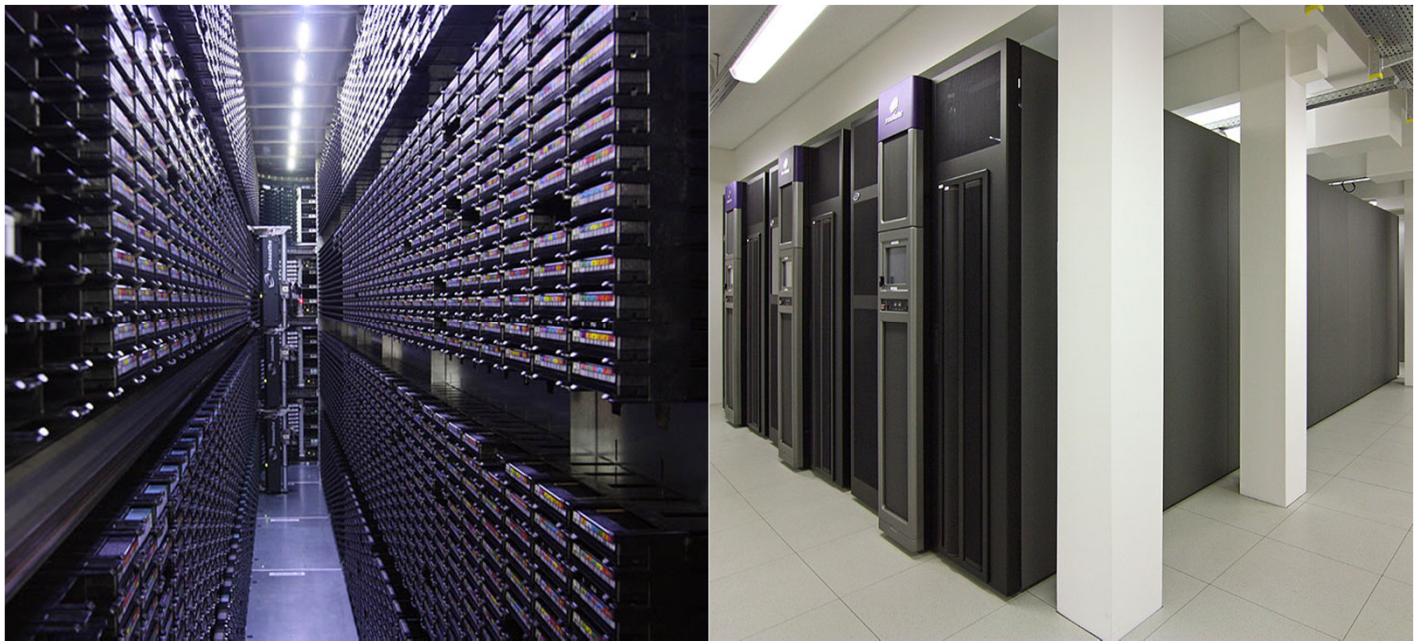
# HLRE-3 – Mistral (2015-2020)



bullx DLC 720, 3,300+ nodes, 100,000+ cores, Haswell/Broadwell, 3.6 PFLOPS
240 TB main memory, **54 PB disk storage**, 450 GB/s mem-disk rate, FDR network
21 nodes for visualization
hot liquid cooling with high efficiency

# High Volume Data Archive

- 65,000 slots for tapes in Hamburg (10,000 remote)
- 60+ PB of climate simulation data
- increase 8 PB/y before 2015, now up to 9x more
- **400 PB capacity until 2020, used 100-300 PB**



# Anforderungen I

- Daten für und von der Erdsystemforschung.
- Datenmengen im PB Bereich.
- Sowohl inhouse (DKRZ) als auch weltweit erzeugt.

=> Technische Randbedingungen (Platten/Bandplatz, Netzwerkbandbreite)

- Daten an (möglichst) einer Stelle zusammenführen um eine übergreifende Analyse zu ermöglichen.
  - Datenstandardisierung
  - „eine Stelle“ wird teilweise schon schwierig.
    - Datenauswahl (wer bestimmt?)
  - Versionierung
  - Identifikation

# Anforderungen II

- Datennutzung außerhalb eines Projektes
  - Dokumentation
  - Auffindbarkeit
  - Verwertbarkeit
  - Nachnutzung dokumentieren
  - Anerkennung / Reputation
  - Lizenz

# Anforderungen III

- Policies& Rules + Transparenz + Impact
  - Nachvollziehbarkeit
  - Reproduzierbarkeit
  - Open Science, Open Access, Open Methods
  - FAIR

# Lösungen I

- Assistenz bei der Planung des Datenmanagements
  - Workflows
  - DMPs
  - Standardisierung
  - Versionierung
- Datensammlungen
  - Unterstützung
  - Realisierung (wenn „übergreifendes“ Interesse)

# Lösungen II

- Datenverteilung
  - Inhomogen (Cloud)
  - Homogen (ESGF) (-> FAIR)
    - Vielzahl von Diensten on top möglich
- Datenarchivierung (10 Jahre+)
  - DOKU (einfach)
  - WDCC (umfaßend) (-> FAIR), Zertifizierung
- Datenpublikation (DataCite DOI + PID)
  - Anerkennung + eindeutige Identifikation

# Aktuelle Themen

- Transparenz der Methoden (Software !)
- Machine Learning  (Unterstützung bei der Auswertung mittels ML)
- Eindeutige Identifikation
- Provenance
- Suche in/nach unstrukturierten/undokumentierten  Daten
- Datenanalyse über verteilten Datensammlungen

thiemann@dkrz.de