HPC FOR TURBULENCE RESEARCH

Workshop zum Forschungsdatenmanagement in den Ingenieurwissenschaften

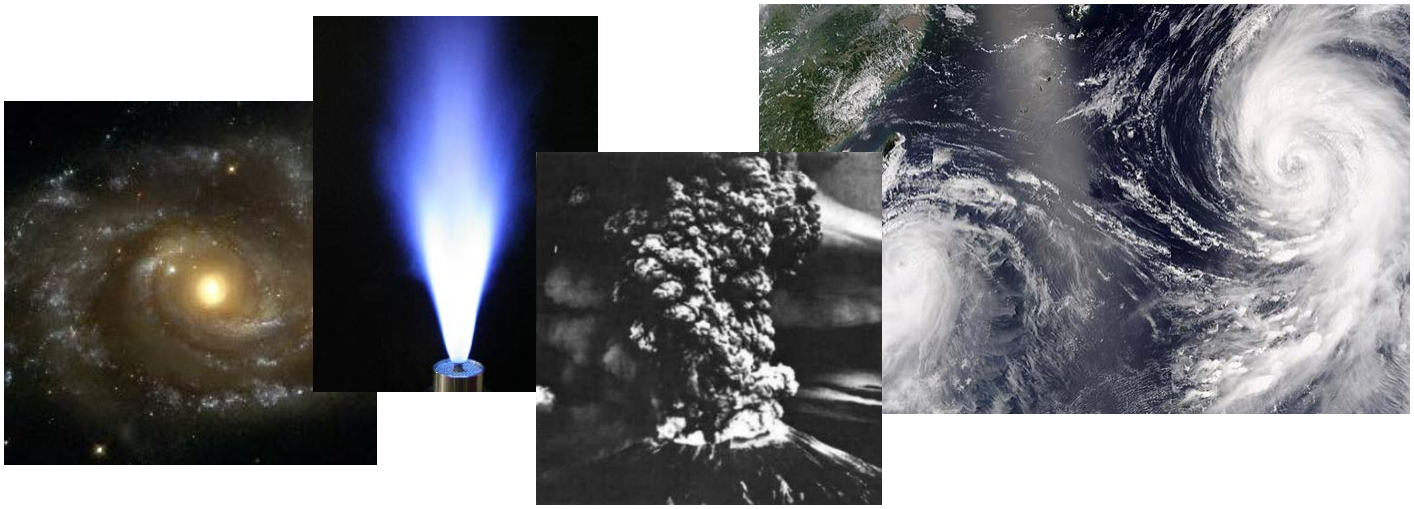
08.03.2017 I JENS HENRIK GÖBBERT - J.GOEBBERT@FZ-JUELICH.DE

TURBULENZ

Welchen Bereich der Ingenieurwissenschaften deckt Ihre Forschung ab?

Turbulenz …

… ist phenomeologisch ein Strömungszustand, charakterisiert durch **chaotische** und **statistische** Eigenschaftsänderungen.



Die Beschreibung des Verhaltens von kleinskaliger Turbulenz gilt als eins der ungelösten Probleme der klassischen Physik …

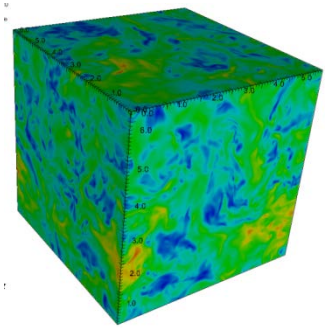
… und ist für große Bereiche der Ingenieurswissenschaften von entscheidender Bedeutung.

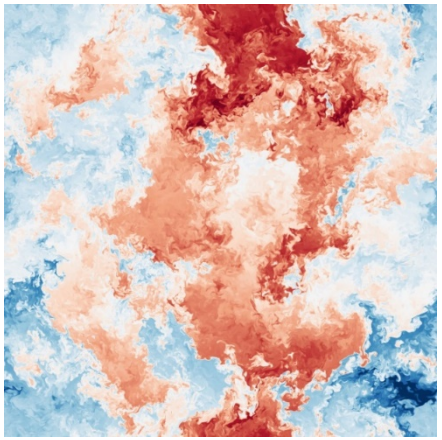
FORSCHUNGSDATEN

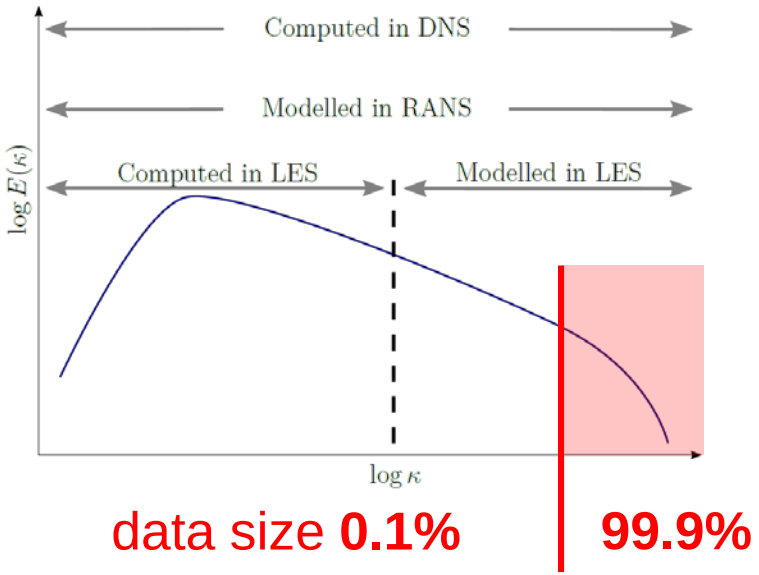
Welche Forschungsdaten sind characteristisch für Íhr Gebiet?

Direkte Numerische Simulation (DNS) …

… löst die Navier-Stokes Gleichung – ohne ein weiteres Modell (also für alle Skalen)







FORSCHUNGSDATEN

Welche Forschungsdaten sind characteristisch für Íhr Gebiet?

	A0	A1	A2	A3	A4	A5	A6
N^3	512^3	1024^3	1024^3	2048^3	2048^3	4096^3	4096^3
Re_λ	88	119	184	215	331	529	754
file size (GB)	8	64	64	512	512	4096	4096
M	180	60	60	10	10	10	10
data size (TB)	1.44	3.81	3.81	5	5	22	22

	B0	B1	B2	B3	B4	B5	B6
N^3	720^3	1440^3	1440^3	2816^3	2816^3	5632^3	6144^3
Re_λ	84	115	173	207	297	529	770
file size (GB)	22	177	177	1331	1331	5324	6912
file size compressed (GB)	6.6	52.6	52.6	393.2	393.2	1572.8	2041.9
M	40	20	20	10	10	5	5
data size (TB)	0.88	3.54	3.54	13.3	13.3	22.4	34.5
data size compressed (TB)	0.26	1.05	1.05	3.9	3.9	6.6	10.3

FORSCHUNGSDATEN

Sind die Daten nach vereinbarten Standards strukturiert oder mit Metadaten versehen?

Community

- Definiert über gemeinsames Forschungsinteresse, Konferenzen, Paper
- Nicht einheitlich organisiert –> loser “Interessensverbund”
- Kontakt zwischen den Wissenschaftlern auf Konferenzen, über Paper, persönl. Kontakt
- Große Variation an Simulationscodes und numerischen Methoden

Daten

- Austausch von Daten i.d.R. über persönlichen Kontakt
- Keine festen Datenformate - Standards nicht üblich bzw. nicht möglich
- Eingesetzt werden z.B. HDF5, netCDF - aber auch pure MPI-IO
- Auch bei gleichem Dateiformat ist die innere Struktur i.d.R. unterschiedlich!
- Ohne Metainformationen bzw. persönl. Kontakt kann die Interpretation schwierig sein

VERÖFFENTLICHEN & TEILEN

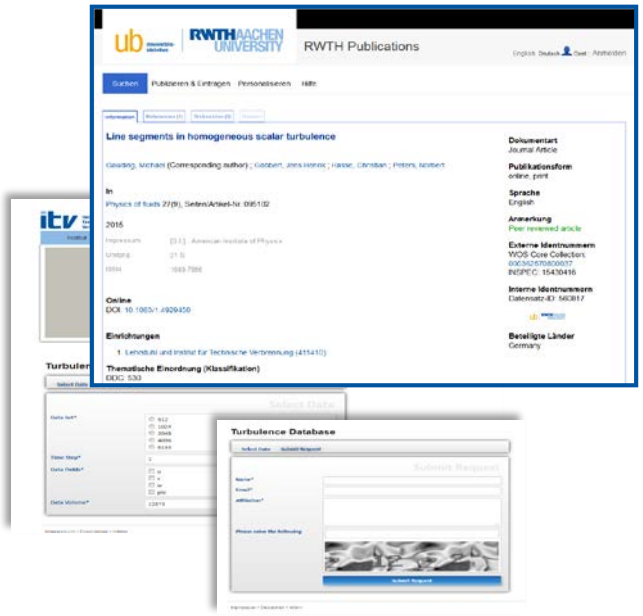
Veröffentlichen oder Teilen Sie Ihre Daten
und aus welchem Lebenszyklus stammen die Daten?
Auf welcher Platform werden die Daten veröffentlicht …

Workflow

- Simulation -> Daten -> Auswertung, Auswertung, Auswertung -> Ergebnisse
  - Simulation (rechenzeitaufwendig)
  - Analyse (“menschen”zeitaufwendig)

Veröffentlichen / Teilen von Daten

- scp, wget, etc.
- (John-Hopkins Turbulence Database)
- Testweise über
  - Rosetta
  - B2Drop + B2Share (EUDat)
  - Kombination etablierter Techniken



WÜNSCHE

… und was könnte zu mehr Datenaustausch führen?
Was würden Sie sich von einer Infrastruktur zum Forschungsdatenmanagement für Ihr Fachgebiet wünschen?

Was motiviert den “kleinen” Forscher eine Infrastruktur zum FDM zu nutzen?

- Weil er/sie es muss (schlechter Grund)
- Weil sein/ihr tägliches Arbeiten dadurch einfacher wird (was genau wird einfacher?)
- Weil er/sie ihre Daten einfacher teilen kann (guter Grund)
  - Einfachere Zusammenarbeit mit anderen Forschern/Gruppen
    - Seine/ihre Daten von einer größeren Gruppe genutzt warden.
      - Er/Sie so öfter zitiert wird -> eigene Ideen warden verbreitet.

Das Teilen von großen Daten auf dem HPC-System muss einfacher werden!

WÜNSCHE

… und was könnte zu mehr Datenaustausch führen?
Was würden Sie sich von einer Infrastruktur zum Forschungsdatenmanagement für Ihr Fachgebiet wünschen?

Das Teilen von großen Daten auf dem HPC-System muss einfacher werden!

- Integration vom FDM in die HPC-Infrastruktur.
  - Aus täglicher Arbeitsumgebung ($HOME, $WORK, $ARCH) Dateien teilen
    - HTTP-Links direkt zu den Dateien auf dem HPC-System
    - Keine Kopie auf ein anderes System (nicht möglich/zeitaufwendig)
  - Persistent Identifier für Daten in $ARCH generieren (referenzierbar)

Das Reproduzieren von Ergebnissen anderer muss einfacher werden!

- Essentiel wichtig für den Start in eine erfolgreiche Kooperation
- Engere Verzahnung von Daten mit Analyse und Ergebnissen (z.B. Jupyter Notebooks)
  - Einheitliches Dateiformat / Metadaten ist nicht die universelle Lösung