Anmerkungen zu Big Data für HEP Computing

Stefan Kluth
MPI für Physik

Big Data

- HEP
  - Verarbeiten sehr grosser Datenmengen
  - Komplexe Analysen
  - “Echtzeit”, d.h. schnelle Zyklen
  - Neue Ideen auf vorhandenen Daten
- Andere
  - Unstrukturierte Daten (z.B. Textanalyse)
  - Zusammenführen getrennter Datensätze
  - Echte Echtzeit

Was will BMBF?

- Nach Besuch der Big Data Konferenz
  - Entwickeln und Verbreiten von Kompetenz für Big Data Probleme
  - M.E. relativ offen für Vorschläge
  - Fachübergreifende Lösungen gesucht
  - Kommerzialisierung?

Was könnten wir?

- Antrag als Verbund
  - HEP Helmholtz (DESY/KIT/GSI?), Universitäten, MPP
- HEP Grid Computing
  - Z.T. Prototypische Big Data Anwendung
  - Aber durch Grid Infrastruktur nicht übertragbar auf andere Bereiche
  - Übertragen des HEP Grid computing auf moderne Big Data fähige Infrastruktur
  - Dadurch Perspektiven und Lösungen für andere

Ausbildung

- Data Scientist und Data Librarian
  - Neues Berufsbild in Wissenschaft/IT?
  - Informell bei uns vorhanden
  - Viele unserer Dr.anden und PostDocs gerade deshalb erfolgreich auf dem Arbeitsmarkt
  - Verstärkung dieser Ausbildung im Antrag

Analogien zu HPC Zentren

- Grosse Resourcen für heterogene Benutzer
  - Bearbeitung von Problemen, die für jede einzelne Gruppe zu gross / zu teuer sind
  - Applikationsgruppen an HPC Zentren
    - Zentrale technische Unterstützung der Wissenschaftler
- High Throughput Computing (HTC) Zentrum
  - Sehr ähnliche Hardware zu HPC
  - Plus grosse und performante Storage
  - Applikationsgruppe
    - Big Data Experten für Unterstützung der Nutzer

Vorschlag für Antrag

- HEP Computing als prototypische Big Data Anwendung auf moderner Infrastruktur
- Hardware für Tier1/Tier2/Tier3 z.B. am KIT
  - Nur z.T. im Big Data Antrag (Hardware kann gefördert werden, z.B. Als
- HTC optimiert
  - GPFS oder CEPH cluster file system, dCache Anbindung
  - Infiniband LAN
- OpenStack Virtualisierung
  - Heterogene Nutzergruppen
  - Tier1/Tier2 gemeinsam
  - Tier3 als eigene virtuelle cluster der Gruppen
- Aufbau einer Big Data Applikationsgruppe
- I.W. FTEs für Aufbau und Betrieb des HTC Zentrums und Big Data Applikationsgruppe

Virtuelle Cluster

- Virtuelle Maschinen auf OpenStack
  - Volle Kontrolle der Nutzergruppe
  - I.d.R. flexibler als Rechenzentren
- Computing Tätigkeiten der Unigruppen
  - ATLAS spezifische Arbeiten gehen weiter wie bisher
  - Administration liegt weiter bei den Gruppen