45. Sitzung des Wissenschaftlichen Lenkungsausschusses der Deutsche Klimarechenzentrum GmbH

Beginn der Sitzung: 26. Mai 2023 um 10:05 Uhr

Teilnehmende

Prof. Dr. Arne Biastoch, GEOMAR Helmholtz-Zentrum für Ozeanforschung Kiel
Dr. Hendryk Bockelmann, DKRZ
Dr. Frauke Feser, Helmholtz-Zentrum Hereon (Vorsitz)
Dr. Bernadette Fritzsch, AWI Bremerhaven (Vorsitzende des DKRZ-Usergroup-Komitees)
Dr. Helge Goessling, AWI Bremerhaven
Dr. Patrick Ludwig, Institut für Meteorologie und Klimaforschung, Karlsruher Institut für Technologie
Prof. Dr. Thomas Ludwig, DKRZ
Dr. Armin Mathes, DLR PT (BMBF)
Prof. Dr. Juan Pedro Mellado González, Meteorologisches Institut, Universität Hamburg
Prof. Dr. Johannes Quaas, Institut für Meterorologie, Universität Leipzig
Dr. Mathis Rosenhauer, DKRZ (Protokoll)
Hannes Thiemann, DKRZ
Prof. Dr. Uwe Ulbrich, Institut für Meteorologie, Freie Universität Berlin
Dr. Sebastian Wagner, Helmholtz-Zentrum Hereon
Prof. Dr. Sönke Zaehle, MPI für Biogeochemie

1. Annahme der Tagesordnung

Die Tagesordnung wird angenommen.

2. Begrüßung der neuen WLA-Mitglieder

Die Vorsitzende Frauke Feser begrüßt Arne Biastoch, Helge Goessling und Sebastian Wagner als neue Mitglieder im WLA.

3. Organisatorisches

a) Annahme des Protokolls der 44. Sitzung

Das Protokoll wird angenommen.

b) Ort und Termin der nächsten Sitzung

Die nächste Sitzung wird am 01.12.2023 in Hamburg stattfinden.

4. Bericht DKRZ

a) Nutzung HLRE-4 (H. Bockelmann)

HLRE-4 ist seit Beginn des Jahres voll betriebsbereit. Als letzter Abschnitt sind 60 GPU-Knoten für die allgemeine Nutzung freigegeben worden. Die CPU-Partitionen werden noch etwas aufwachsen, da eine Kompensation für die Minderleistung des bisherigen Systems erwartet wird.

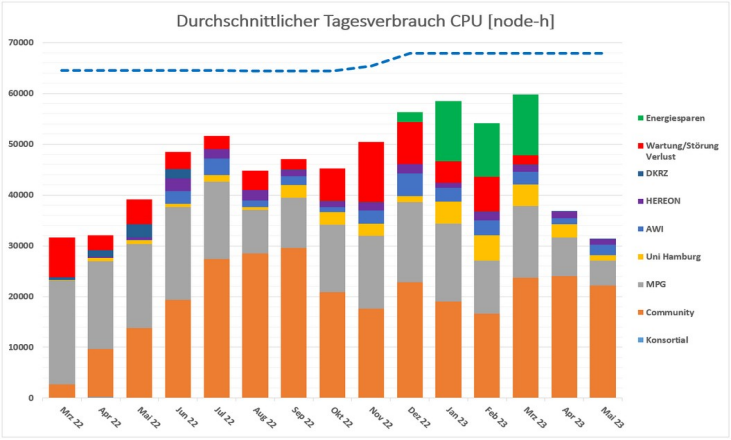

Abbildung 1: Durchschnittliche Auslastung der CPU-Rechenknoten von Levante

Die Auslastung des Systems war in den vergangenen Monaten eher gering. Da der Rechner noch relativ neu ist, wird eine bessere Auslastung mit zunehmender Anpassung aller Nutzergruppen an das System erwartet.

Seit Jahresende 2022 werden unbenutzte Knoten ausgeschaltet. Das Hochfahren abgeschalteter Knoten erfolgt derzeit noch manuell. Bei einem automatisierten Hochfahren kann noch nicht sichergestellt werden, dass Jobs nicht auf einzelne Knoten mit Startproblemen warten müssen. Daher wird eine stets verfügbare Reserve an laufenden Knoten vorgehalten. Der Energieverbrauch von Racks mit abgeschalteten Knoten ist signifikant geringer gegenüber Knoten im Leerlauf.

Um die Auslastung zu erhöhen, beschließt der WLA nicht vergebene Rechenzeit allen Community-Projekten anteilig ihres Rechenzeitverbrauchs in 2023 zuzuteilen.

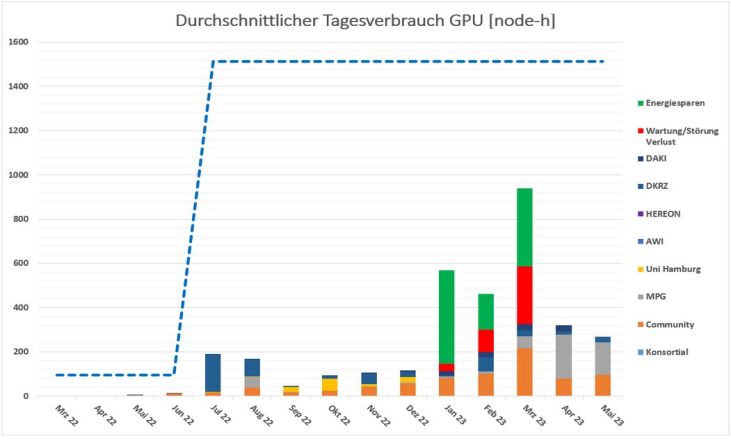

Abbildung 2: Durchschnittliche Auslastung der GPU-Knoten von Levante

Nach der allgemeinen Verfügbarkeit der GPU-Knoten werden diese allmählich von Projekten eingesetzt. Die Auslastung ist insgesamt noch gering. Die anteilige Nutzung durch die MPG ist aufgrund einiger Großprojekte recht hoch.

Die Nutzung der GPUs durch Maschinelles Lernen ist derzeit noch recht gering, da sich ML-Jobs meist auf eine GPU oder höchsten einen Knoten beschränken. Im Gegensatz dazu kann die auf GPUs angepasste Version von ICON bereits eine große Zahl von GPU-Knoten parallel verwenden und ist daher der Hauptnutzer der Partition..

ICON für GPUs wird hauptsächlich vom MPI-M eingesetzt aber auch andere Gruppen sind dabei ihre Nutzung auszubauen. Neben ICON gibt es weitere Portierungsarbeiten auf GPUs für das Modell FESOM. Insgesamt wird jedoch erwartet, dass der Technologiewechsel auf GPUs zu einer Konsolidierung der Modellpalette führt. Der WLA begrüßt ausdrücklich die Arbeiten des DKRZ im Rahmen der vom BMBF finanzierten Nat-ESM Initiative, in der durch das DKRZ Unterstützung bei der GPU-Portierung durch coding-sprints geleistet wird. Dieser Service steht auf Antrag unter Bewertung durch das Nat-ESM Steering Committee allen Gruppen bei, die einen Beitrag zur Nationalen Erdsystemmodellierung leisten möchten. Kontakt: nat-esm.de oder per E-Mail an support-request@nat-esm.de.

Im Verlauf des Jahres wird der Fair-Share-Mechanismus des Batch-Systems für die GPU-Partition aktiviert werden, was bisher wegen der vergleichsweise geringen Nutzung noch nicht notwendig war.

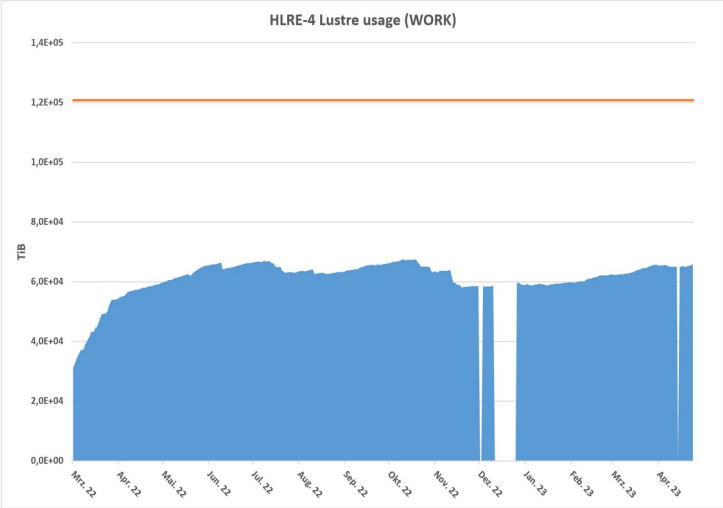

Abbildung 3: Belegung des Lustre Dateisystems auf Levante [TiB]

Die Auslastung des parallelen Dateisystems liegt bislang noch bei ca. 50%. Der Quota-Mechanismus verhindert zuverlässig, dass einzelne Projekte oder User das System über das zugewiesene Maß hinaus belasten.

Neu eingeführt wird „Fastdata“. Das System besteht aus SSDs und HDDs und erlaubt den besonders schnellen Zugriff auf häufig genutzte Daten, die bevorzugt auf SSDs gespeichert werden.

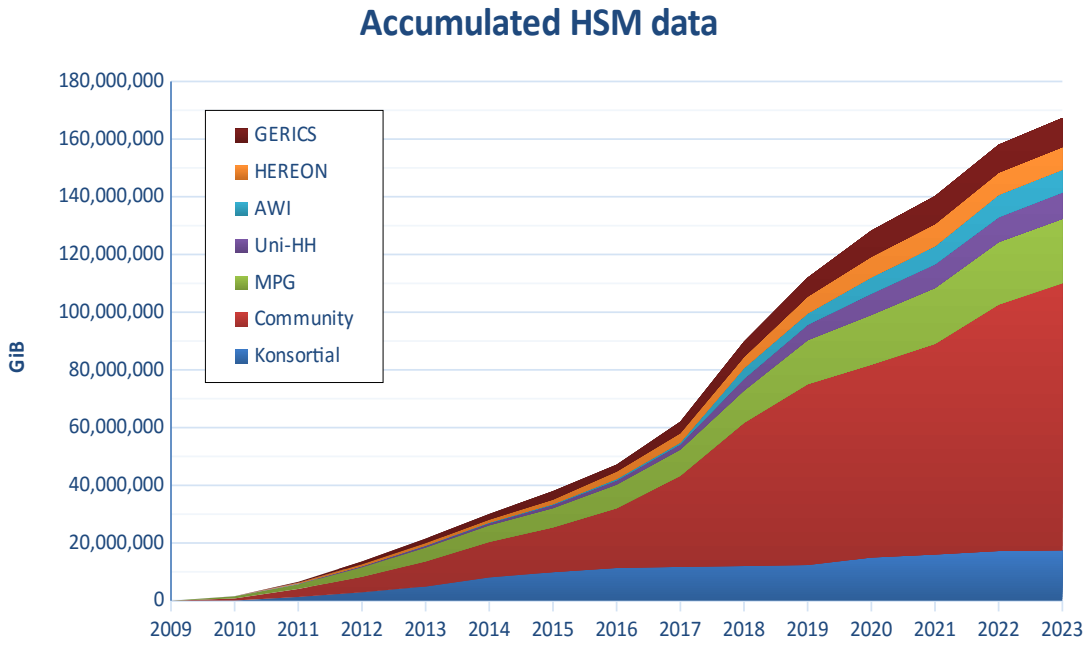

Abbildung 4: Belegung des Bandarchivs [GiB]

Das neue Archiv wird insbesondere von Community-Projekten intensiv genutzt. Viele der anfänglichen Probleme sind inzwischen behoben oder können umgangen werden. Beim Zurückladen der Daten von Bändern gab es zwar noch Verzögerungen, jedoch hat die Inbetriebnahme einer neuen Bibliothek die Wartezeiten deutlich verkürzt.

b) Perspektiven (T. Ludwig)

Der Antrag für den neuen HLRE-5 wird im Frühsommer bei der HGF eingereicht werden. Ein hoffentlich positiver Bescheid wird dann im kommenden Jahr erwartet. Ab 2025 würde der Ausschreibungsprozess beginnen. Für das Jahr 2026 wäre der erforderliche Umbau des Gebäudes in der Bundesstraße 45a vorgesehen. Im Jahr 2027 würde die Aufstellung und Übergabe an die Benutzer erfolgen. Dieser Zeitplan entspräche einer Standzeit des HLRE-4 von sechs Jahren.

Der Übergang von HLRE-3 auf HLRE-4 lieferte einen Faktor vier in der Rechenleistung. Ein weiterer Faktor vier für HLRE-5 ist wahrscheinlich nur mit einem größeren Ausbau des GPU-Anteils möglich.

Der Umbau des Gebäudes betrifft insbesondere die Kühlanlage, welche auf das Dach des Hauptgebäudes umziehen muss.

Im Leistungsverzeichnis, welches 2025 erstellt wird, müssen der maximale Stromverbrauch, GPU-Anteil des Rechners, sowie der Anteil für den Festplattenspeicher festgelegt werden. Diese Aufteilung muss von einer Kommission aus Gesellschaftern, Usern und dem WLA bestimmt werden.

Der WLA würde es begrüßen, wenn für den HLRE-5 zusätzlich zu den Finanzmitteln für das System auch Mittel für zusätzliches Servicepersonal bereitgestellt werden könnten.

Auch die Zusammensetzung der Benchmarks muss im Vorfeld erörtert werden. Der Prozess begann bei früheren Ausschreibungen etwa ein Jahr vor Veröffentlichung des Leistungsverzeichnisses. Für HLRE-5 wird es wichtig sein eigene Benchmarks für den GPU-Anteil zu definieren.,

5. Neue Geschäftsordnung WLA

Ein Entwurf für eine Geschäftsordnung wurde erstellt und an die Juristen der Gesellschafter weitergeleitet. Für die Satzungsänderung des DKRZ muss die neue Geschäftsordnung des WLA vorliegen. Nach der notariellen Beglaubigung der Satzungsänderung kann die Geschäftsordnung dann unabhängig von der Satzung des DKRZ verändert werden.

6. Bericht aus der DKRZ-User-Group (B. Fritzsch)

Die Sitzungen der Usergroup haben inzwischen einen öffentlichen Teil, in dem alle Nutzer über Erfahrungen und Probleme mit dem Rechner berichten können.

Aufgrund von weiterhin bestehenden Problemen mit Levante, konnten einige Projekte noch nicht wie geplant ihre Rechenzeit abrufen, was bei der Begutachtung berücksichtigt werden sollte. Zwar werden die Bemühungen des DKRZ für Stabilität zu sorgen anerkannt, jedoch zwingen die Schwierigkeiten einige Nutzer nach alternativen Möglichkeiten zu suchen.

Zum Anteil von GPUs am nächsten Rechner soll es seine Umfrage geben. Dabei werden Status und geplante Portierungen von Modellen abgefragt. Auch zu ML/AI sollen die Vorhaben erfasst werden. Vom WLA wird Unterstützung der Umfrage gewünscht, um deren Rücklaufquote zu erhöhen.

7. Sonstiges

Keine Diskussionspunkte.

1. Rechenzeitanträge (intern)

Im nichtöffentlichen Teil der Sitzung wurde unter anderem über die Rechenzeitanträge für Community- und Konsortial-Projekte beraten.

Es wurden Ressourcen für Neu- und Folgeprojekte über den Zeitraum vom 01.07.2023 bis 30.06.2024, sowie zusätzliche Ressourcen über den Zeitraum vom 01.07.2023 bis 31.12.2023 bewilligt. Im einzelnen sind dies:

	Beantragt	Bewilligt
Levante CPU [Node hours]	4.740.109	3.664.203
Levante GPU [Node hours]	94.000	93.700
Levante storage [TiB]	9.598	7.801
Archive project [TiB]	10.836	4.556
Archive long term [TiB]	3.018	1.448

Ende der Sitzung: 16:03