**Sachbericht zum Verwendungsnachweis**

---

Vorhabenbezeichnung: High-Performance Computing for Applied Artificial Intelligence (HPC4AAI)
FKZ 13FH050KI1

---

Laufzeit des Vorhabens: 01.08.2021 – 29.02.2024

---

Ziel des Vorhabens ist die Etablierung von HPC-Ressourcen zur Unterstützung von Studierenden, Nachwuchswissenschaftlerinnen und -wissenschaftlern und Forschungsgruppen mit KI-Bezug an der Hochschule für Angewandte Wissenschaften in Coburg (HSCo). Dadurch wird ein Beitrag für die Entwicklung der KI-Kompetenzen an der Hochschule geleistet. Zur Zielerreichung wurde ein initialer Arbeitsplan eingereicht, der sich aus den Phasen Anforderungserfassung (AP1), Entwicklung eines Nutzungskonzepts (AP2), Anschaffung und Inbetriebnahme der Hardware (AP3), Umsetzung der Softwareinfrastruktur und der Benutzungsverwaltung (AP4), sowie Schulungen / Rollout (AP5) zusammensetzt.

Im ersten Projektjahr verzögerten sich Arbeitspakete initial durch die angespannte Situation am Arbeitsmarkt bei der Besetzung einer TV-L E13 Stelle im Bereich Künstliche Intelligenz und High-Performance Computing. Die im Antrag initial avisierten Personen standen für diese Stelle aufgrund persönlicher Gründe nicht mehr zur Verfügung.

Aufgrund einer aktualisierten Bedarfsanalyse (AP1) wurde im ersten Projektjahr zudem eine Aufteilung der Rechenkomponenten als sinnvoll erachtet. Dazu wurde die Verortung eines Teils der Rechenknoten in das Regionale Rechenzentrum Erlangen (RRZE) sowie ein Teil für lokales Hosting (On Premise an der Hochschule Coburg) vorgesehen. Dazu wurde Ende Oktober 2021 ein entsprechender Änderungsantrag eingereicht und positiv beschieden. Die Rahmenbedingungen zum Hosting des RRZE wurden erfolgreich abgestimmt und die Beschaffung eines Teils der Komponenten („GPU-Knoten zum Hosting in Erlangen“ laut Änderungsantrag) eingeleitet. Durch die Integration dieser Komponenten in das RRZE konnte für diese Komponenten das Nutzungskonzept (AP2) und die Benutzerverwaltung (AP4) teilweise übernommen werden.

Die Beschaffung der CPU-Knoten zum Hosting in Erlangen sowie der lokalen Knoten in Coburg wurden im ersten Projektjahr vorbereitet, aber aufgrund substantieller Lieferschwierigkeiten seitens potentieller Anbieter und massiver Preissteigerungen (ca. 25%) erst später abgeschlossen.

Im zweiten Projektjahr konnte die Anschaffung und Inbetriebnahme der Hardware (AP3) substantiell vorangetrieben werden. Aufgrund vorhandener Preisdynamiken wurde auf die

Beschaffung lokaler Clusterkomponenten verzichtet und aufgrund des angespannten Arbeitsmarktes eine Laufzeitverlängerung beantragt.

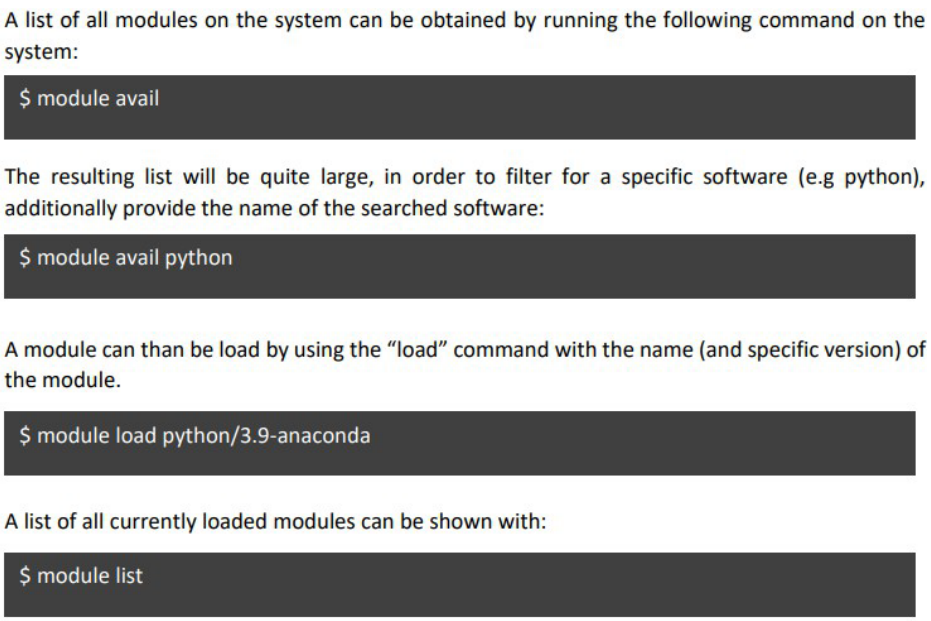
Die GPU-Cluster Komponenten konnten im zweiten Projektjahr erfolgreich beschafft und in Betrieb genommen werden. Dies geschah in enger Abstimmung mit dem Technologieintegrator sowie dem Regionalen Rechenzentrum Erlangen (RRZE). Die Mitarbeiterstelle konnte in Q3/2022 erfolgreich besetzt werden.

Zudem wurde an einem initialem Nutzungskonzept (AP2) sowie der Benutzerverwaltung und Softwareinfrastruktur (AP4) gearbeitet. Nach der Erörterung verschiedener Zugangsmodelle wurde sich dafür entschieden einen möglichst niedrigschwelligen Zugang für einen größtmöglichen Nutzungskreis zu ermöglichen. Konkret wurden zeitlich begrenzte Zugänge (Studierende: Ein Semester, Mitarbeitende: 1 Jahr) geplant, die mittels einem Online-Formular (initial per E-Mail) über das Intranet der Hochschule Coburg beantragt (und verlängert) werden können. Die Nutzenden bekommen dadurch Zugang zu dem jeweiligen Compute-Cluster. Die Benutzerverwaltung wurde (und wird weiterhin) durch eine zentrale Person gewährleistet (während der Projektlaufzeit über die Projektstelle, jetzt über eine dauerfinanziere Stelle) und erfolgt in einem online-basierten Managementsystem, welches zusammen mit dem RRZE genutzt und weiterentwickelt wird. In diesem System wird zwischen der Nutzungsgruppe der Studierenden und der Mitarbeitenden unterschieden (mit den oben angegebenen unterschiedlichen Laufzeiten). Zur Nutzung des GPU-Clusters wurde sich für die Verwendung von Containervirtualisierungen (Docker, Singularity) in Zusammenspiel mit einem Workloadmanager (Slurm) entschieden – auch dies erfolgte in enger Absprache mit dem RRZE.

Im zweiten Projektjahr wurden zudem Testanwenderinnen und -anwendern aus den Bereichen Bioinformatik, Autonomes Fahren, Physik und Informatik Zugang zum System gewährleistet (AP 5). Auf Basis der initialen Rückmeldungen wurde damit begonnen Schulungsmaterialien (Schritt-für-Schritt Anleitungen sowohl als schriftliches Dokument als auch als Video) zu konzipieren, sowie Anpassungen am Antragsvorgang (Umstellung von E-Mail auf Online-Formular, Präzisierung von im Formular zu machenden Angaben) vorgenommen.

Im dritten Projektjahr wurde sich auf das Rollout und die Schulungen von Nutzenden fokussiert. Um die verschiedenen Nutzergruppen in die Lage zu versetzen, die bereitgestellte KI-Computing-Infrastruktur selbständig zu bedienen und ein allgemeines Bewusstsein für die Existenz dieser neuen Option zu schaffen, wurde ein mehrstufiger Rollout durchgeführt. Dabei wurden spezifisch die heterogenen Hintergründe unterschiedlicher Nutzergruppen beachtet. Insbesondere wurde auf der hochschulinternen Lernplattform Moodle ein für alle Interessierten zugänglicher Kurs eingerichtet, der als zentrale Plattform für alle Tutorials und Schulungen rund um den HPC- Cluster dient. Englisch wurde als primäre Sprache für alle Ressourcen gewählt, um die Barriere für

internationale Studierende an der Hochschule Coburg zu senken. Für relevante Informationen zum Cluster wurde ein (bisher) 25-seitiges PDF-Dokument erstellt, das als lebendes Dokument bei Bedarf durch Updates erweitert wird (vgl. Auszug in Abbildung 1).


A list of all modules on the system can be obtained by running the following command on the system:

```
$ module avail
```

The resulting list will be quite large, in order to filter for a specific software (e.g python), additionally provide the name of the searched software:

```
$ module avail python
```

A module can than be load by using the "load" command with the name (and specific version) of the module.

```
$ module load python/3.9-anaconda
```

A list of all currently loaded modules can be shown with:

```
$ module list
```

Abbildung 1: Auszug aus dem schriftlichen Tutorial.

Das Dokument behandelt die folgenden Aspekte des Clusters: 1) Einführung: Grundlegende Informationen zum Tutorial. 2) Arbeitsablauf: Schematische Darstellung der Arbeit mit dem Cluster als Flussdiagramm. 3) Antrag für das HPC-System: Informationen zum Antragsverfahren, zur ersten Anmeldung am Portal und zur Aktivierung des Kontos. 4) Verbindungsaufbau mit dem HPC-System: Verbindungsaufbau zum Cluster mittels SSH, RSA- Schlüsselgenerierung und SSH-Konfigurationsdatei. 5) Dateisystem: Vorstellung und Erläuterung der verschiedenen auf dem Cluster verfügbaren Dateisysteme. 6) Übertragung von Daten: Übertragen von Dateien mit FileZilla und dem Befehl scp. 7) Module: Laden und Verwenden vorinstallierter Softwaremodule. 8) Conda: Ausführen eigener Programme auf dem Cluster mit Conda- Umgebungen. 9) Singularity: Ausführen eigener Programme auf dem Cluster mit Singularity und allgemeines Arbeiten mit Containern. 10) Docker: allgemeine Arbeit mit Docker. 11) Konvertierung: Umstellung von Docker auf Singularity. 12) Batch: Arbeiten mit Jobskripten in Slurm und Reservierung von GPU-Ressourcen im interaktiven Modus. 12) Auftragsüberwachung mit ClusterCockpit: Vorstellung des ClusterCockpit-Systems zur Überwachung und Kontrolle aktueller und vergangener Aufträge.

Zudem wurde auch eine Reihe von Video-Tutorials produziert, welche die wichtigsten Aspekte der Nutzung des Clusters vermitteln (siehe Abbildung 2).

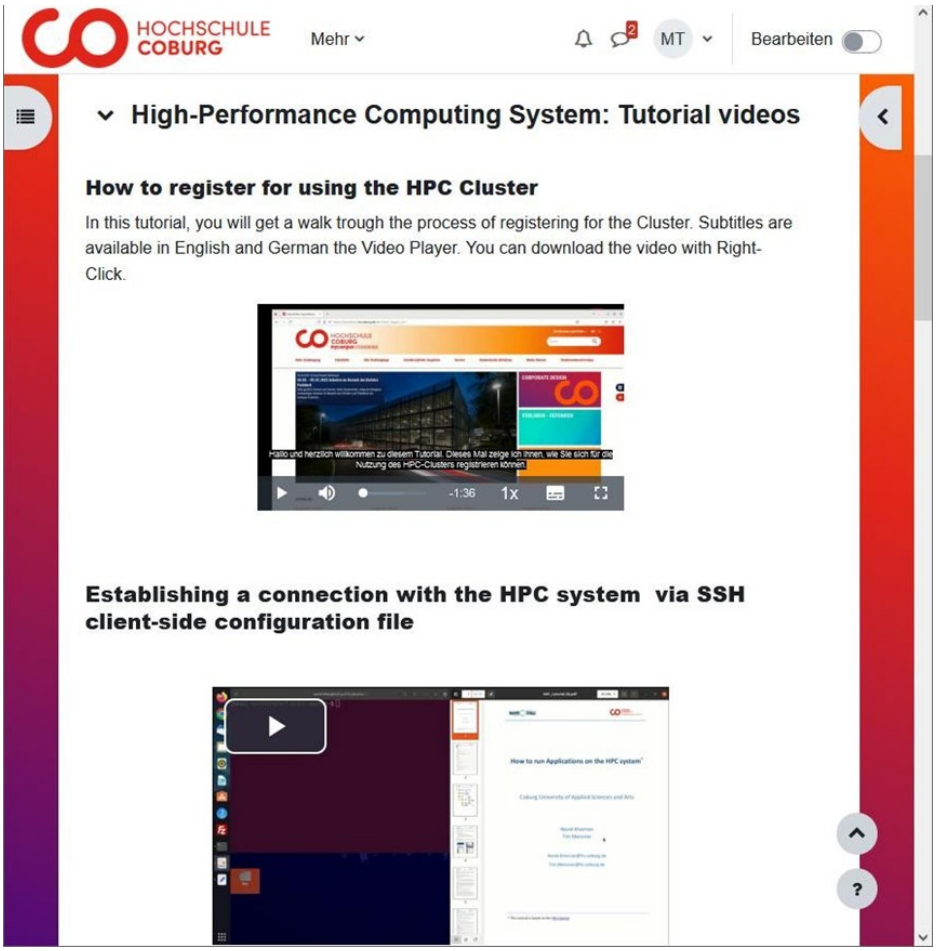



Abbildung 2: Moodle Kurs mit Video-Tutorials.

Analog zu ausgewählten Inhalten der schriftlichen Dokumentation, behandeln die Videos folgende Themen im Detail: 1) Anmeldung für den HPC-Cluster: Dieses Video beschreibt den kompletten Registrierungsprozess für neue Nutzer, von der ersten Anmeldung in mycampus bis zum ersten Login im HPC-Portal. 2) Verwendung von SSH und Herstellung einer Verbindung zum Cluster (Linux): In diesem Video werden die Benutzer durch den gesamten Prozess der Anmeldung beim Cluster und der Herstellung einer Verbindung geführt. Dazu gehören die Erstellung von RSA-Schlüsseln über das Terminal, das Hochladen des öffentlichen Schlüssels auf den Cluster und die Verbindung über SSH. Bei der Verbindung über SSH wird besonders auf die Verwendung einer Konfigurationsdatei geachtet, um den Prozess so effizient wie möglich zu gestalten. 3) Verwendung von SSH und Aufbau einer Verbindung zum Cluster (Windows): Dieses Video beschreibt vergleichbare Abläufe wie sein Gegenstück für Linux, jedoch unter Windows, um allen, die an diesem Betriebssystem interessiert sind, den Zugang zu erleichtern. 4) Dateiübertragung zum und vom Cluster: In diesem Tutorial wird die Übertragung von Dateien zwischen einem lokalen Rechner und dem Cluster mittels scp in beide Richtungen erläutert. Neben dem regulären Konsolenbefehl wird auch das darauf aufbauende Tool FileZilla vorgestellt, das den Vorgang durch

eine grafische Benutzeroberfläche erheblich vereinfacht. 5) Module und Conda auf dem Cluster: Dieses Video erklärt, wie Benutzer die vorinstallierten Softwaremodule auf den Cluster laden können und wie sie mit Conda ihre eigenen Python-Umgebungen erstellen können, um ihre eigenen Programme auszuführen. 6) Arbeiten mit Singularity-Containern: Dieses Tutorial zeigt, wie man einen Singularity-Container erstellt und ihn benutzt, um seine eigene Software auf dem Cluster laufen zu lassen. Es werden drei Möglichkeiten vorgestellt: die Erstellung über eine Definitionsdatei, die Installation von Modulen in einem bereits erstellten Container und die Konvertierung von Docker zu Singularity. 7) Arbeiten mit Docker-Containern: Dieses Video behandelt Docker-Container auf die gleiche Weise wie 6). Nach einer Erklärung der Grundlagen der Containerisierung wird gezeigt, wie Container in Docker erstellt, konvertiert und dann verwendet werden können. 8) Auftragsverwaltung über Batch: Dieses Video zeigt, wie Jobs auf dem Cluster erstellt und geplant werden können. Es geht explizit auf die Verwendung von Jobskripten für Slurm und die Reservierung von Ressourcen im interaktiven Modus ein.

Im Oktober 2023 wurde ein Workshop geplant und durchgeführt, um Interessenten den Einstieg in die Nutzung des Clusters zu erleichtern und für die Möglichkeit der Nutzung in größerem Rahmen zu werben. Alle Professorinnen, Professoren, Mitarbeitenden und Studierende der Hoschule waren eingeladen.

Nach einer kurzen Einführung in die Grundlagen des Cluster-Computings und einer Vorstellung der verfügbaren Cluster wurden verschiedene Prozesse und Technologien anhand eines konkreten Beispiels erläutert. Alle Teilnehmenden hatten ihre eigenen Laptops mitgebracht, mit denen sie sich mit dem Cluster verbinden konnten. Konkret ging es darum, das Text-Bild-Modell StableDiffusion gemeinsam mit allen Teilnehmenden über ein Python-Skript auf dem Cluster laufen zu lassen.

Alle notwendigen Schritte, von der ersten Anmeldung am Cluster über die Übertragung von Dateien bis hin zur Reservierung von GPU-Ressourcen, wurden so durchgeführt, dass alle Teilnehmenden mitmachen konnten. Falls es bei einem der Teilnehmenden Probleme gab, wurden diese zunächst behoben, bevor mit dem Walkthrough fortgefahren wurde. So wurde sichergestellt, dass auch technisch weniger versierte Teilnehmende nicht auf der Strecke blieben. Um die verschiedenen Möglichkeiten der Nutzung eigener Software auf dem Cluster abzudecken, wurde die Ausführung des Python-Skripts und die Installation der notwendigen Abhängigkeiten einmal mit einer Conda-Umgebung und einmal mit einem Container demonstriert.

Neben einem konkreten Beispiel wurden auch einige grundlegende Tipps zur Arbeit mit dem Cluster erläutert, wie z.B. die Eigenschaften der verschiedenen verfügbaren Dateisysteme.
Die Teilnehmenden hatten während des gesamten Workshops die Möglichkeit, Fragen zu stellen, aber am Ende, nach einer kurzen Zusammenfassung, gab es eine weitere Zeitspanne, die

ausdrücklich für diesen Zweck vorgesehen war. Der Workshop wurde zudem aufgezeichnet (vgl. Abbildung 3) um Nutzenden, die nicht am Workshop teilnehmen konnten, die Möglichkeit zu bieten den Ausführungen nachträglich zu folgen.

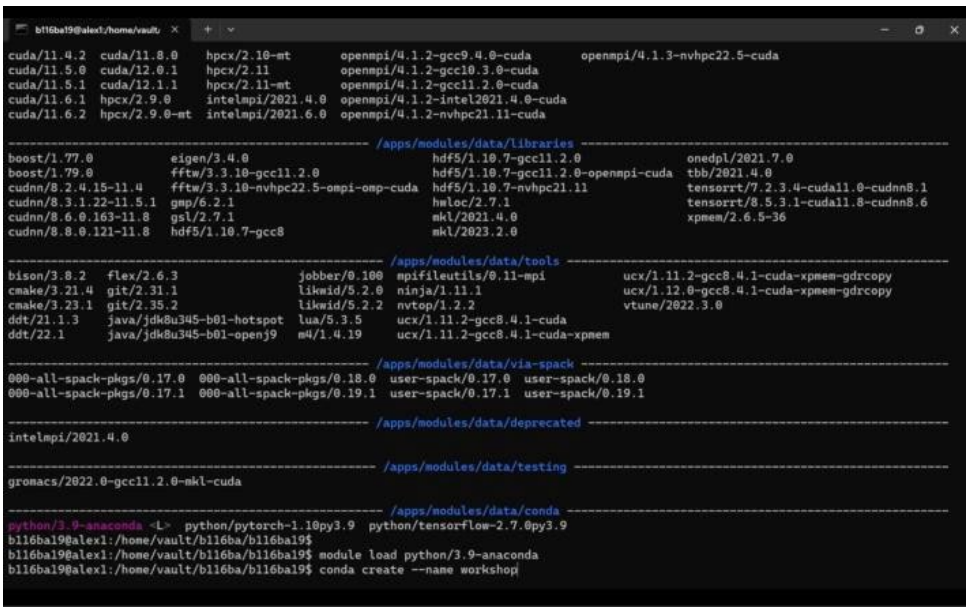



Abbildung 3: Screenshot aus der Bildschirmaufzeichnung des Workshops

Die Aufzeichnung der Veranstaltung wurde wie die zugehörige Präsentation, auf Moodle hochgeladen, damit Interessierte welche nicht teilnehmen konnten, sie Online nachholen können.

Um den Nutzern des Clusters einen angemessenen Support zu bieten, wurde ein ticketbasiertes Supportsystem auf Basis von Zammad eingerichtet (vgl. Abbildung 4). Alle Anfragen zu Neuregistrierungen, Verlängerungen und Support sind mit einer Formular-/Support-E-Mail-Adresse verknüpft. Es ist auch möglich, sich an eine Support-Telefonnummer zu wenden. Seit dem Start des Supportsystems Anfang März 2023 konnten bereits mehrere Dutzend Anfragen erfolgreich gelöst werden.

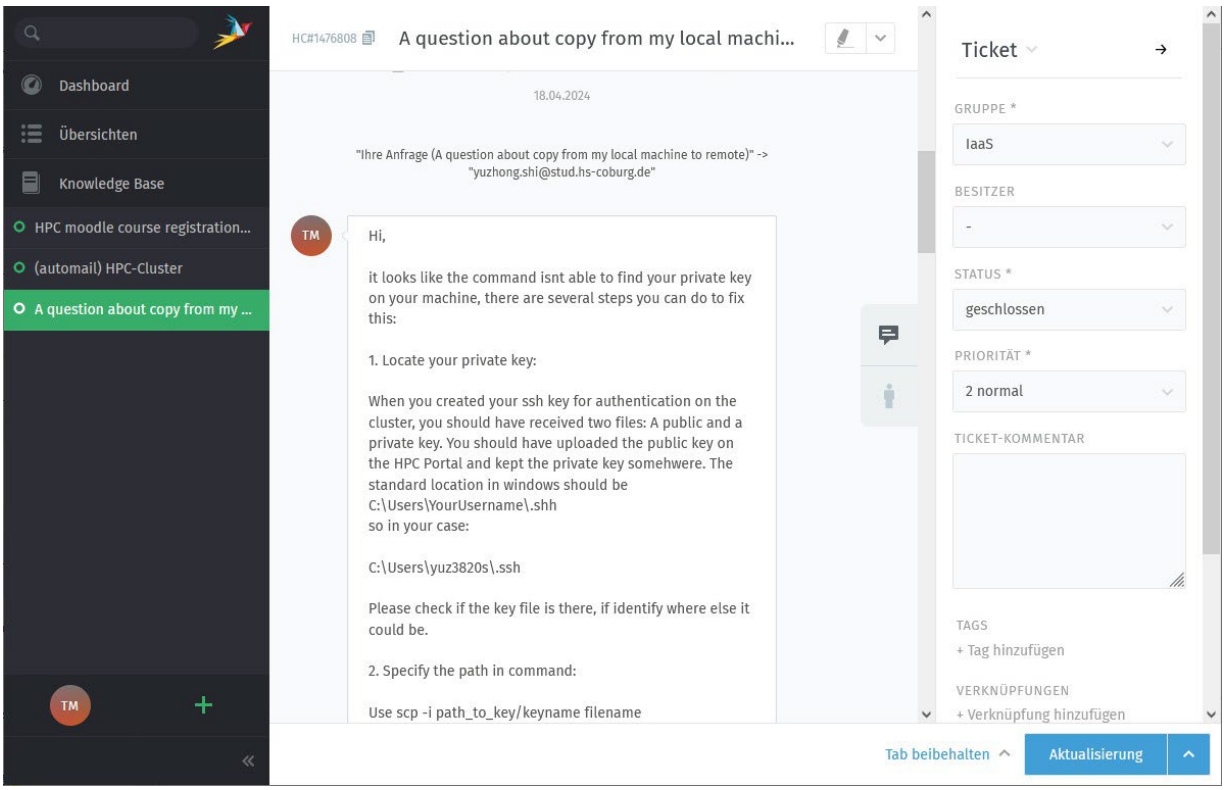



Abbildung 4: Screenshot eines Support-Tickets.

### Wichtigste Positionen des zahlenmäßigen Nachweises

Es fielen zwei hauptsächliche Kostenpositionen an:

1) Beschäftigungsentgelte für einen wissenschaftlichen Mitarbeiter (Ausgabeart 0812) in Höhe von insgesamt 90.332,31 €, davon 22.513,43 € im Jahr 2022, 58.705,10 € im Jahr 2023 und 9.113,88 € im Jahr 2024.

2) Die Kosten der beschafften HPC-Komponenten (Ausgabeart 0850) in Höhe von insgesamt 984.606,00 €, davon 806.106,00 € m Jahr 2022 und 178.500,00 € im Jahr 2023.

### Notwendigkeit und Angemessenheit der geleisteten Projektarbeiten

Das Projekt hat die gesteckten Projektziele trotz Verzögerungen bei der Personaleinstellung, massiver Preissteigerungen in der angeschafften Hardware und der damit verbundenen Anschaffungsverzögerungen erreicht. Zur effizienten Zielerreichung hat die konstruktive Zusammenarbeit mit dem RRZE beigetragen. Die durchgeführten Arbeiten entsprachen, unter Berücksichtigung der Änderungsanträge, den ursprünglich geplanten Arbeiten. Ohne diese notwendigen Arbeiten hätte das HPC entweder nicht beschafft oder nicht in Betrieb genommen werden. Die Arbeiten waren angemessen, da diese ressourceneffizient mittels der bewilligten TV-L E13 Stelle umgesetzt wurden. Die Arbeiten wurden weiterhin durch zusätzliches Personal komplementiert.

**Voraussichtlicher Nutzen, insbesondere die Verwertbarkeit des Ergebnisses - auch konkrete Planungen für die nähere Zukunft - im Sinne des fortgeschriebenen Verwertungsplans**

Der bisher erreichte und nach Projektende (durch die zur Verfügung stehende Dauerstelle sowie Verortung der Hardware am RRZE) weiterbestehende Nutzen liegt in dem realisierten und nachhaltig bestehenden Zugang von Studierenden, Nachwuchswissenschaftlerinnen und -wissenschaftlern und Forschungsgruppen mit KI-Bezug zu KI-HPC-Ressourcen - auch in Forschungskooperationen mit externen Partnern aus Wissenschaft und Wirtschaft. Der Nutzen geht sogar über die initial antizipierte Nutzendengruppe hinaus, der sich primär an Studierende und wissenschaftlichen Nachwuchs aus naturwissenschaftlichen-technischen Bereichen ausrichtete. Insbesondere wurden zusätzlich Studierende mit gestalterischem Fokus an der Fakultät Design im Studiengang Integriertes Produktdesign dazu befähigt, mittels dem HPC-System aktuelle Fragestellungen der generativen KI zu bearbeiten (vgl. https://www.hs-coburg.de/news-detailseite/mode-zeichnungen-produkte-ki-als-designer.html und https://ai-productdesign.de/ ).

**Bekannt gewordener Fortschritt auf dem Gebiet des Vorhabens bei anderen Stellen**

Bei dem Projekt handelt es sich primär um ein Projekt zur Etablierung von Rechenressourcen zur Ermöglichung neuer angewandter Forschung, aber nicht um ein eigenes Forschungsprojekt. Das Angebot wird durch den Ausbau nationaler Hochleistungsrechenzentren komplementiert.

**Erfolgte oder geplante Veröffentlichungen des Ergebnisses**

Da es sich bei dem Projekt primär um ein Projekt zur Etablierung von Rechenressourcen zur Ermöglichung neuer angewandter Forschung handelt, aber nicht um ein eigenes Forschungsprojekt, sind keine spezifischen wissenschaftlichen Veröffentlichungen über das Projekt selbst geplant. Das Projekt wurde jedoch im Rahmen von Vorträgen der Nutzendengemeinschaft von HPC-Systemen in Bayern vorgestellt. Zudem ist nach Projektabschluss eine weitere Vorstellung und Vernetzung im Rahmen des Kompetenznetzwerk für Technisch-Wissenschaftliches Hoch- und Höchstleistungsrechnen in Bayern geplant. Zudem wird durch die HPC-Infrastruktur die Publikation von Projektergebnissen in wissenschaftlichen Veröffentlichungen unterstützt.

**Kurzbericht zum Verwendungsnachweis**

---

Vorhabenbezeichnung: High-Performance Computing for Applied Artificial Intelligence (HPC4AAI)
FKZ 13FH050KI1

---

Laufzeit des Vorhabens: 01.08.2021 – 29.02.2024

---

An der Hochschule Coburg finden KI-spezifische Lehr-, Forschungs- und Transfertätigkeiten an drei Standorten und in fünf Fakultäten in Studiengängen von A wie Autonomes Fahren bis Z wie ZukunftsDesign statt. Nachdrückliches Ziel der Hochschule ist es Synergien zwischen den Akteuren innerhalb der Hochschule, mit regionalen KMUs in der Region und weiteren Hochschulpartnern nachhaltig zu stärken. Die Schaffung von hochschulweit effizient nutzbaren HPC Rechenressourcen für KI ist ein zentraler Grundpfeiler für den Erfolg der regionalen KI-Forschungsstrategie der HAW Coburg.

Zur Zielerreichung wurde ein initialer Arbeitsplan verfolgt, der sich aus den Phasen Anforderungserfassung (AP1), Entwicklung eines Nutzungskonzepts (AP2), Anschaffung und Inbetriebnahme der Hardware (AP3), Umsetzung der Softwareinfrastruktur und der Benutzungsverwaltung (AP4), sowie Schulungen / Rollout (AP5) zusammensetzt.

Nach einer Bedarfsanalyse wurde eine Beschaffung von GPU-, CPU- und Speicherkomponenten angestrebt. Eine Kooperation mit dem Regionalen Rechenzentrum Erlangen (RRZE), insbesondere dem Zentrum für Nationales Hochleistungsrechnen, erlaubte die effiziente Integration und Inbetriebnahme der Hardware. Insgesamt wurden sieben GPU Knoten mit jeweils acht NVIDIA A100 80 GB GPUs und acht Sapphire Rapid Knoten mit jeweils 104 Rechenkernen angeschafft. Ergänzt wurden diese Rechenkapazitäten durch zwei NVMe Koten und einer Speicherkapazität von 300 TB.

Nach Abwägung verschiedener Zugangsmodelle wurde sich dafür entschieden einen möglichst niedrigschwelligen Zugang für einen größtmöglichen Nutzungskreis zu ermöglichen. Konkret wurden zeitlich begrenzte Zugänge für Studierende und Mitarbeitende geplant und umgesetzt, die mittels einem Online-Formular über das Intranet der Hochschule Coburg beantragt und verlängert werden können. Die Nutzenden bekommen dadurch Zugang zu dem jeweiligen GPU und CPU Systemen. Der Rollout des Systems wurde durch dedizierte Schulungsmaterialien (Schritt-für-Schritt Anleitungen als Videos und schriftliche Dokumente) sowie Workshops unterstützt.