

bwForCluster NEMO

Forschungscluster für die Wissenschaft

Michael Janczyk Dirk von Suchodoletz Bernd Wiebelt 

eScience Abteilung, Albert-Ludwigs-Universität, Freiburg, Deutschland

In den ersten zweieinhalb Jahren seiner Betriebszeit entwickelte sich der bwForCluster NEMO zu einem signifikanten Baustein in den landesweiten Forschungsinfrastrukturen für das »High Performance Computing«. Der in der Zwischenzeit erhebliche Ausbau und die Erweiterung des Systems durch Shareholder ist ein Beleg für die Tragfähigkeit seines Betriebsmodells und das Vertrauen in das landesweite HPC-Konzept. Hierzu steuert nicht nur die lokale und landesweite Governance bei, sondern ebenfalls der enge Austausch innerhalb der NEMO-Community. Mit dem System wird eine stabile Umgebung für die diversen Bedürfnisse der Wissenschafts-Communities bereitgestellt. Parallel dazu werden neue Betriebs- und Monitoring-Konzepte entwickelt und getestet. Aktuelle und neuartige Herausforderungen liegen in der Unterstützung von »Virtualisierten Forschungsumgebungen« und zukünftigen digitalen Workflows ebenso wie in der Containerisierung und der Implementierung effektiver Betriebsmodelle gemeinsam mit den am Standort Freiburg betriebenen Cloud-Infrastrukturen.

1 Einleitung

Der bwForCluster NEMO¹ adressiert Forscher*innen auf Tier-Ebene 3, dem Einstiegssegment des »High Performance Computings« (HPC). Prinzipbedingt durch die Unterstützung verschiedener wissenschaftlicher Communities muss mit einer Mischung aus unterschiedlichen Benutzerprofilen und entsprechend heterogenen Erwartungshaltungen geplant und gearbeitet werden. Zur Versorgung der Fach-Com-

¹Zum Zeitpunkt des Verfassens ist der Cluster zweieinhalb Jahre (08/2016 – 01/2019) im Produktivbetrieb und damit bereits bei der Hälfte seiner auf 5 Jahre ausgelegten Betriebszeit.

munitys kommt hinzu, dass es für einige Arbeitsgruppen die erste Berührung mit Rechnen jenseits des Desktops ist, während andere Forschende bereits auf (eigenen) Clustern Erfahrungen sammeln konnten. Zusätzlich wurden Arbeitsgruppen akquiriert, welche die Forschungsinfrastruktur durch eigene finanzielle Beteiligungen vergrößert und sich damit als Shareholder erweiterte Nutzungsrechte erworben haben. Aus Betreibersicht muss eine sich ausdehnende Landschaft von Compute- und Storage-Systemen in komplexer werdenden wissenschaftlichen Workflows, die beispielsweise ein Pre-Processing in der Cloud und eine spätere Visualisierung großer Datenmengen vorsehen, effektiv gemanagt werden. Um ein Austarieren der vielfältigen Interessen und einen harmonischen Betrieb zu gewährleisten, wurden entsprechende Governance- und Betriebsstrukturen geschaffen, die sich in den zweieinhalb Jahren Laufzeit bewährt haben (Wesner u. a., 2016; Wiebelt u. a., 2016).

2 Die beteiligten Wissenschafts-Communitys

Der bwForCluster NEMO am Standort Freiburg bedient im Landesverbund von bwHPC die Bedürfnisse der Wissenschaft aus den Bereichen Elementarteilchenphysik, Neurowissenschaft, Mikrosystemtechnik und Materialwissenschaft (ENM). Im neuen »Umsetzungskonzept der Universitäten des Landes Baden-Württemberg für das High Performance Computing (HPC), Data Intensive Computing (DIC) und Large Scale Scientific Data Management (LS²DM)« von Schneider u. a. (2019) werden die Schwerpunkte noch detaillierter nach der DFG-Fachsystematik bis auf die Ebene der einzelnen Fächer aufgeteilt.² Das NEMO zugeordnete HPC-Kompetenzzentrum ENM unterstützt nach dieser die in Tabelle 1 dargestellten Schwerpunkte (vgl. auch die grafische Darstellung in Abbildung 1a).

Die Schwerpunkte spiegeln sich ebenfalls in den Rechenvorhaben der Nutzer*innen wider. Ein Rechenvorhaben stellt dabei einen Projektantrag für Rechenressourcen dar. Es dient der Auswahl des Forschungsschwerpunkts für das Rechenvorhaben seitens des Antragstellers und der Einteilung zu einem Forschungscluster durch ein über alle Clusterstandorte agierendes Clusterauswahlteam (3 Governance). Abbildung 2a zeigt die angemeldeten Rechenvorhaben auf dem bwForCluster NEMO in den ersten zweieinhalb Betriebsjahren. Die Rechenvorhaben teilen sich auf die ursprünglichen drei Schwerpunkte Elementarteilchenphysik, Neurowissenschaft und

² Fächer nach DFG-Fachsystematik: http://www.dfg.de/dfg_profil/gremien/fachkollegien/liste/index.jsp (besucht am 18.07.2018).

Schwerpunkte	DFG-Fachsystematik	Fächer
Neurowissenschaft	206 Neurowissenschaft	206-01 – 206-11
Elementarteilchenphysik	309 Teilchen, Kerne und Felder	309-01
Mikrosystemtechnik	407 Systemtechnik 408 Elektrotechnik und Informationstechnik	407-03, 407-06 408-01
Materialwissenschaft	405 Werkstofftechnik 406 Materialwissenschaft	405-01 – 405-05 406-01 – 406-05

Tabelle 1: NEMO Schwerpunktbildung nach DFG-Fachsystematik.

Mikrosystemtechnik auf. Je ein Rechenvorhaben aus den Fachgebieten Materialwissenschaft und Geowissenschaften wurden von Shareholdern gestellt. Investitionen von Forschungsgruppen waren bereits beim Antrag des Clusters ein wichtiger Erfolgsfaktor, um eine Konsolidierung der ursprünglich dezentralen und in Eigenregie betriebenen Ressourcen der Fach-Communitys am Rechenzentrum der Universität Freiburg zu erreichen.³

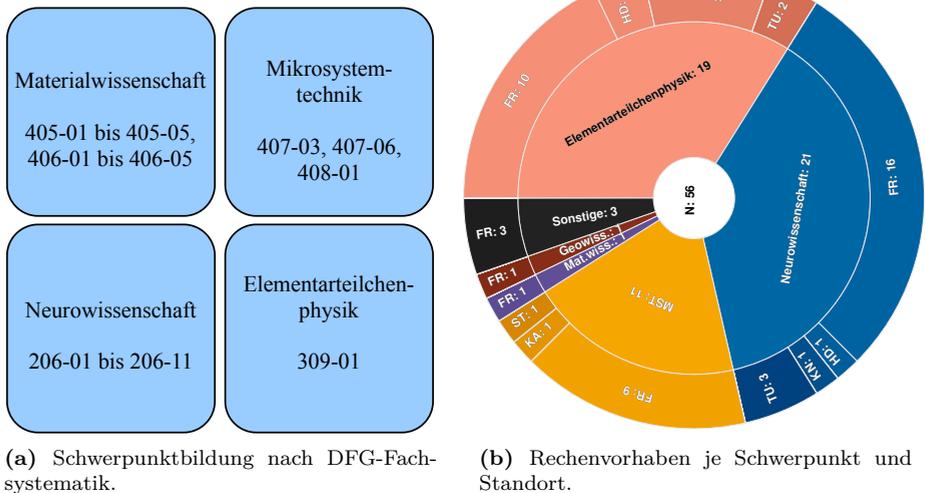
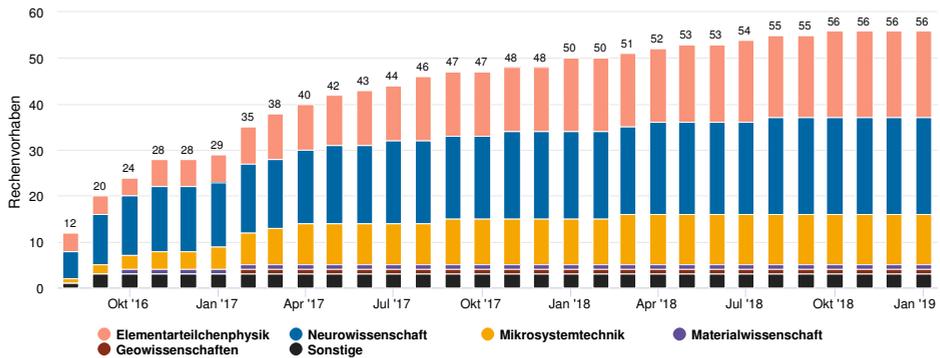
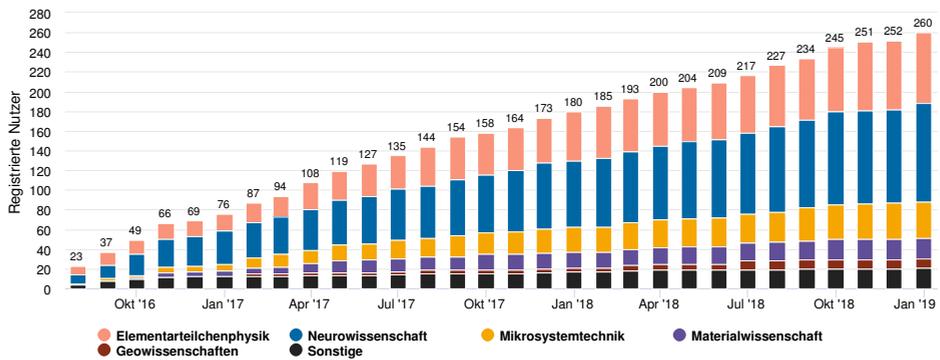


Abbildung 1: Schwerpunktbildung des bwForClusters NEMO.

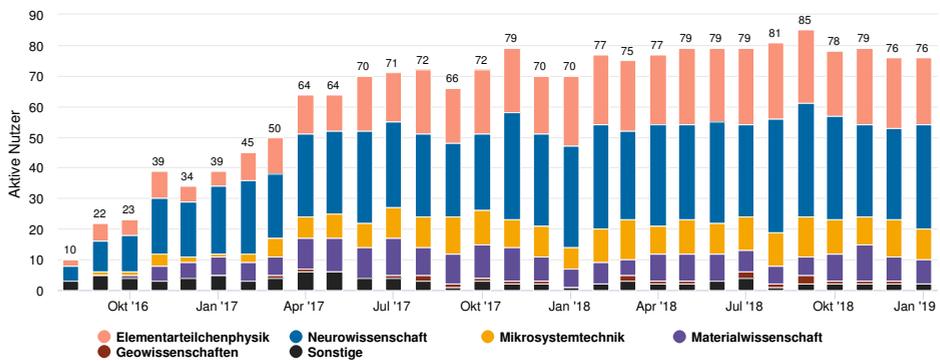
³Shareholder können alle Arbeitsgruppen aus dem Land werden und profitieren von einem für ihre Fachbereiche optimierten Forschungscluster.



(a) Angemeldete Rechenvorhaben.



(b) Registrierte Nutzer*innen.



(c) Aktive Nutzer*innen pro Monat.

Abbildung 2: Rechenvorhaben- und Nutzerentwicklung in den ersten zweieinhalb Jahren aufgeteilt nach Fachbereichen auf dem Forschungscluster NEMO.

Die Materialwissenschaft wie auch in kleinerem Maße die Geowissenschaften hatten, obwohl nur je ein Rechenvorhaben angemeldet wurde, einige registrierte (Abbildung 2b) und aktive (Abbildung 2c) Nutzer*innen. In der Materialwissenschaft lasteten die aktiven Nutzer*innen die Cluster-Ressourcen wahrnehmbar aus (Abbildung 8a). Dieses Profil führte dazu, dass das für den bwForCluster NEMO zuständige HPC-Kompetenzzentrum ENM im »Umsetzungskonzept II« um den Schwerpunkt Materialwissenschaft ergänzt wurde (Schneider u. a., 2019).

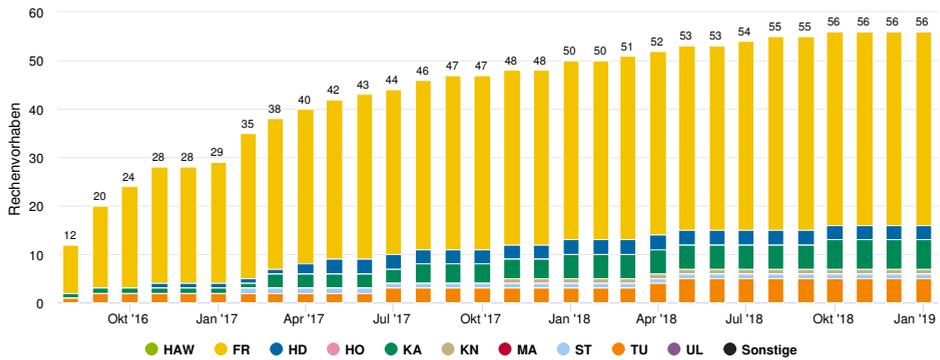
In Abbildung 2b wird deutlich, dass seit der offiziellen Inbetriebnahme des Clusters stetig neue Nutzer*innen dazu gekommen sind. Diese sind in einem hohen Maße aktiv. Abbildung 2c zeigt nur die Anzahl der Forschenden an, die mindestens einen Job im betreffenden Monat abgeschickt haben. In den ersten Monaten stieg diese Anzahl, bis sie sich in einem stabilen Fenster zwischen 70 und 80 aktiven Nutzer*innen pro Monat einpendelt. Diese Zahl dürfte wohl noch größer ausfallen, wenn Nutzer*innen innerhalb der Virtualisierten Forschungsumgebungen (VFU) einbezogen würden.⁴

Die Schwerpunktbildung in Baden-Württemberg erfolgte im Vorfeld des Clusterantrags nach den am Standort aktivsten wissenschaftlichen Communitys im HPC-Umfeld (Hartenstein u. a., 2013). Dies erklärt zudem die starke Konzentration der Rechenvorhaben aus Freiburg, wie in Abbildung 3a dargestellt. Die registrierten (Abbildung 3b) und aktiven (Abbildung 3c) Forschenden verteilen sich ähnlich. Die genaue Aufteilung der Rechenvorhaben auf die jeweiligen Felder und Standorte lässt sich Abbildung 1b entnehmen.

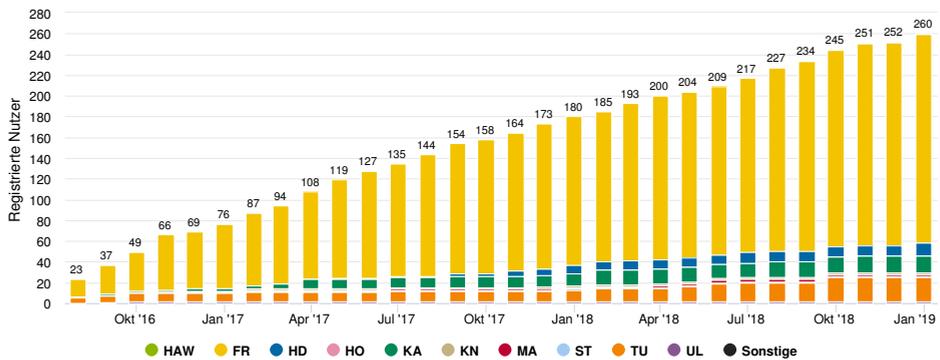
Die in Abbildung 8b gezeigte Auslastung verhält sich weniger eindeutig. Zwar stammt hier ebenfalls ein großer Teil der Auslastung aus Freiburg, aber insbesondere in den ersten Monaten überwiegen Wissenschaftler*innen aus Karlsruhe. Vergleicht man die Anzahl der aktiven Karlsruher Forschenden (Abbildung 3c) mit der Auslastung, fällt auf, dass teilweise ein einzelner Cluster-Nutzer bis zu einem Drittel des gesamten Clusters verwendet (Abbildung 4).

Diese Grafik stellt auch die Auslastung des Clusters der Nutzung durch Virtualisierte Forschungsumgebungen gegenüber. Diese decken sich insbesondere in den ersten Monaten mit der Auslastung durch Nutzer*innen aus Karlsruhe. Die VFU der »Compact Muon Solenoid Collaboration« (CMS) am CERN der Karlsruher Elementarteilchenphysik wurde bereits am vorherigen Testcluster aufgesetzt und lief

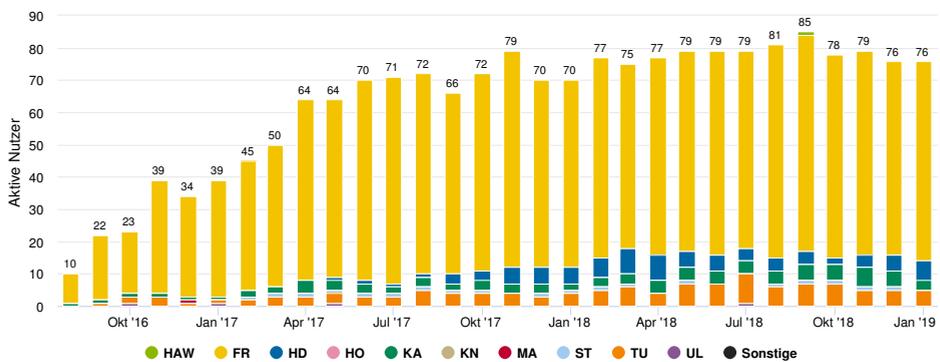
⁴Die VFU wird von einem Nutzer*innen der Arbeitsgruppe gestartet. Nutzer*innen innerhalb einer VFU sind für das System nicht sichtbar und werden von der Statistik nicht erfasst.



(a) Angemeldete Rechenvorhaben.



(b) Registrierte Nutzer*innen.

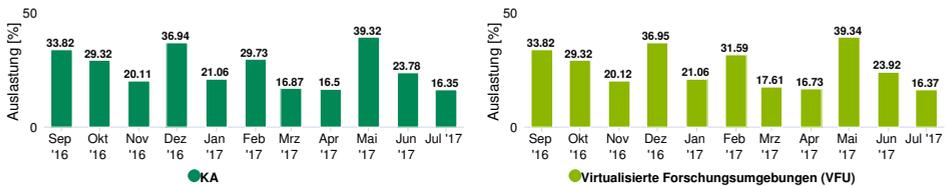


(c) Aktive Nutzer*innen.

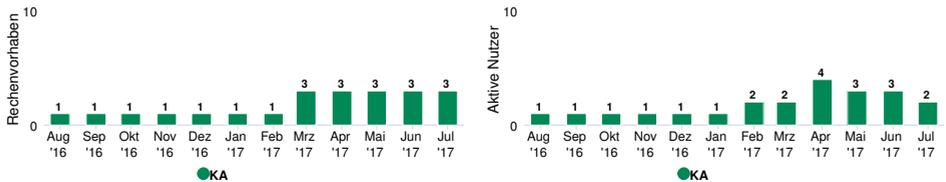
Abbildung 3: Rechenvorhaben- und Nutzerentwicklung in den ersten zwei-einhalb Jahren aufgeteilt nach Standorten auf dem Forschungscluster NEMO.

dort einige Monate stabil (Meier, 2017). Deshalb konnte die VFU CMS zu Beginn einen signifikanten Teil des Clusters auslasten (Abbildung 4a).

Die virtuelle Maschine, welche die VFU CMS zur Verfügung stellt, wird nur durch einen einzigen Nutzer gestartet. In der VFU können jedoch alle Karlsruher CMS-Gruppen rechnen. Deshalb ist der Vergleich aktive Nutzer*innen zu Auslastung des Clusters bei VFUs verzerrt. Die VFU wird mittelfristig auf einen VFU-Nutzer pro Wissenschafts- oder Arbeitsgruppe aufgeteilt werden. Eine direkte Abbildung der Cluster-Nutzer auf Forscher*innen in den jeweiligen VFUs ist zukünftig geplant. Spätestens bei der Nutzung von Containerisierungslösungen wie »Singularity« sind die Nutzer*innen in der Wirts- und Containerumgebung identisch.⁵



(a) Vergleich Virtualisierte Forschungsumgebung und Auslastung durch Karlsruher Nutzer.



(b) Rechenvorhaben und aktive Nutzer*innen aus Karlsruhe.

Abbildung 4: Vergleich der Statistiken für den Standort Karlsruhe mit der Auslastung durch VFUs im ersten Jahr von NEMO.

3 Governance

Für einen fairen Ausgleich der Interessen und einen reibungslosen Betrieb einer großen Forschungsinfrastruktur wie NEMO ist es wichtig, die Benutzer*innen frühzeitig, regelmäßig und in angemessener Art und Weise in Entscheidungsprozesse zu involvieren. Hierzu zählen sowohl anstehende Hardwareerweiterungen, Aufnahme neuer Shareholder, mögliche Erweiterungen der NEMO-Community oder Weiter-

⁵Singularity <https://www.sylabs.io/singularity> (besucht am 12.02.2019).

entwicklung des Betriebsmodells. Das Rechenzentrum als Betreiber des Forschungsclusters NEMO sieht sich in der Rolle des Dienstleisters der vier Fach-Communitys und greift hierfür auf deren Beratung und Vorschläge zurück. Aufgrund der hohen Anzahl an Beteiligten aus den ENM-Communitys wurde ein zweistufiges Modell aus großer Nutzerversammlung und kleinem Cluster-Beirat etabliert (Suchodoletz, Wiebelt und Janczyk, 2017). Die breit aufgestellte Nutzerversammlung, die einmal im Jahr tagen sollte, erlaubt es, ein Gesamtbild über Zufriedenheit und zukünftige Anforderungen aller involvierten Anwender*innen zu erhalten. In diesem Gremium erfolgen die Berichte durch das NEMO-Team, die Abstimmung der ENM-Communitys und die Vorstellung anstehender Entwicklungen.

Gleichzeitig werden aus den Reihen der wissenschaftlichen Communitys, der Betriebsgruppe und der Shareholder Vertreter in einen kleinen, handlungsfähigen Cluster-Beirat entsandt, der sich in halbjährlichen und bei Bedarf noch engeren Zyklen trifft und operative Belange des Clusters erörtert. Mitglieder des Cluster-Beirats sind Forschende der aktuell rechnenden Gruppen der ENM-Communitys, ein Vertreter des Landesnutzerausschusses (LNA-BW) sowie des NEMO Technical Advisory Boards (TAB), Vertreter der Shareholder sowie die operative Leitung des bwForCluster NEMO und bei Bedarf zusätzliche Expert*innen in beratender Funktion. Der Cluster-Beirat unterstützt die operative Leitung des bwForClusters NEMO in Belangen des Berichtswesens und der lokalen Governance sowie das HPC-Kompetenzzentrum ENM (Barthel u. a., 2019) inklusive dessen Entscheidungen im Cluster Auswahl Team (CAT). In den ersten zweieinhalb Jahren Laufzeit des Clusters tagte die Nutzerversammlung zwei Mal und der Cluster-Beirat fünf Mal.

Die NEMO-Governance funktioniert über eine enge Rückkopplungsschleife mit den Fach-Communitys, wie sie ähnlich bereits in der Antragsphase sowie zur Ausschreibung und Beschaffung erfolgte. Das dient gleichzeitig der Entlastung übergeordneter Ebenen wie Landesnutzerausschuss oder bwHPC-Lenkungskreis von ENM-spezifischen Belangen. Gleichwohl bleiben die übergeordneten Ebenen die letzte Instanz bei Problemen, die das bwHPC-Konzept als Ganzes betreffen, wie beispielsweise im Fall längerer Wartezeiten in der Queue (Wesner u. a., 2016).

Zu den Empfehlungen des Cluster-Beirats zählen die Einführung eines Technical Advisory Boards, die Einführung von sogenannten Memory-Knoten mit 256 bzw. 512 GiB RAM oder den Verzicht auf den Ausbau der XEON-Phi-Kapazitäten

zugunsten klassischer Rechenknoten. Das NEMO-TAB umfasst die Administratoren beziehungsweise technikaffinen Mitglieder der einzelnen Forschungsgruppen und bespricht betriebliche Belange des Clusters, um diese dann kompakt in den Cluster-Beirat zu tragen.

4 Die Ausrichtung des Forschungsclusters

Bei der Beschaffung des Clusters wurden die Wünsche der im Förderantrag gesammelten Forschungs-Communitys als Grundlage genommen. Daraus wurde eine sehr einheitliche Hardwarekonfiguration destilliert, die bisher bei den Erweiterungen beibehalten wurde. Der Forschungscluster NEMO ist für sich gesehen die größte zusammenhängende Maschineninstallation am Rechenzentrum der Universität Freiburg. Hinzu kommen inzwischen weitere Compute-Systeme, die ebenfalls von der Abteilung eScience administriert werden. Bei den nutzenden Communitys der weiteren Systeme bestehen eine Reihe von Überschneidungen, so dass sowohl bei der Auswahl der Hardware als auch der Nutzung von Ressourcen wie Speichersystemen gemeinsame Interessen bestehen, die im Betriebsmodell abgebildet werden.

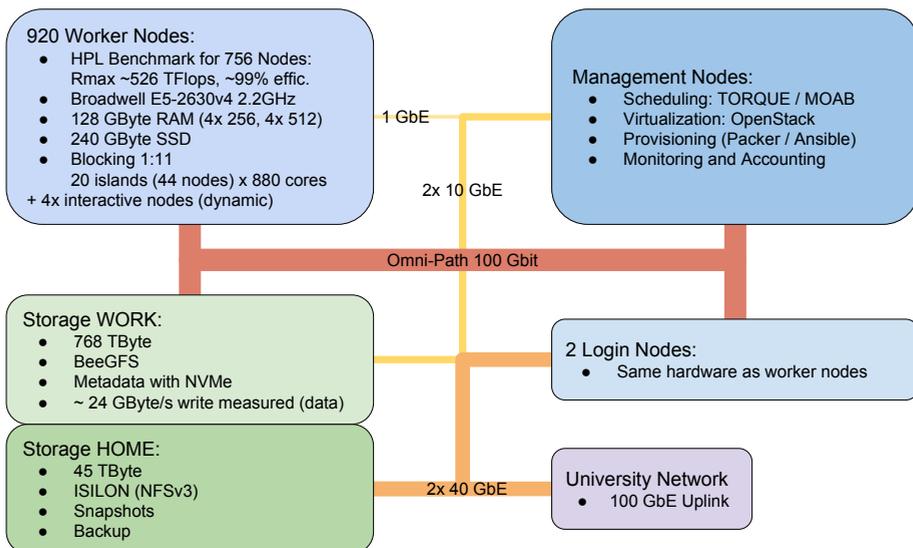


Abbildung 5: Aufbau und Netzwerkanbindung des bwForClusters NEMO.

Abbildung 5 zeigt die schematische Darstellung des bwForClusters NEMO. In der derzeitigen Konfiguration besteht NEMO inzwischen aus 920 reinen Rechenkno-

ten. Zu Beginn hatte der Cluster 748 Rechenknoten und wurde dann schrittweise auf 920 Knoten vergrößert. Vor diesen Erweiterungen wurde der MKL-optimierte »High Performance Computing Linpack Benchmark« (Intel Corporation, 2018) auf 756 identischen Rechenknoten durchgeführt und erreichte den Wert 525,714 TFlop/s bei einem theoretischen Maximalwert von 532,224 TFlop/s und einer Effizienz von etwa 98,78 %. Mit diesem Wert erreichte der Cluster im Juni des Jahres 2016 den Platz 214 der TOP500-Liste.⁶

Die Rechenknoten »Worker« verfügen jeweils über 128 GiB Hauptspeicher, einer SSD mit 240 GB Speicherplatz und einen 100 Gbit/s-Adapter für das Omni-Path-Hochleistungsnetzwerk. Jeweils 44 Maschinen sind zu einer Insel per Omni-Path und Gigabit-Ethernet verbunden. Jeder Omni-Path-Switch ist mit 4×100 Gbit/s an die Omni-Path-Spine-Ebene angebunden. Das ergibt einen Blocking-Faktor von 11:1. Pro Schrank sind zwei Inseln verbaut. Diese Konfiguration vermindert die Verkabelung, da nur wenige Kabel schrankübergreifend gezogen werden müssen.⁷ Die meisten Kabel verbleiben innerhalb eines Schrankes. Der Cluster besteht aus 20 Inseln mit je 880 Kernen, die jeweils bei Bedarf non-blocking verwendet werden können. In der Praxis wird das von den Nutzer*innen aber nicht genutzt, da dann die Wartezeiten in der Jobqueue steigen. Aber auch ohne Non-Blocking-Konfiguration werden Jobs mit teilweise über 2000 Kernen auf NEMO gerechnet.

Der zentrale Parallelspeicher hatte zu Beginn 576 TB Speicherplatz, wurde aber bereits bei der ersten Erweiterung auf 768 TB vergrößert (nutzbare Kapazität). Die Metadaten liegen hierbei für den schnellen parallelen Zugriff auf NVMEs. Eingesetzt wird BeeGFS vom Fraunhofer ITWM.⁸ Bei einem Test konnten vor der Erweiterung Nutzdaten mit über 24 GB/s übertragen und gespeichert werden.

4.1 Betriebsmodell

Das zugrundeliegende Betriebsmodell wurde auf eine effiziente Beschaffung, Inbetriebnahme und Erweiterbarkeit ausgelegt. Es nutzt hierzu das bereits länger etablierte Konzept des »Netzwerk-Bootens« (Schmelzer u. a., 2014), wodurch auf eine lokale Betriebssysteminstallation auf den einzelnen Knoten verzichtet werden kann.

⁶TOP500-Liste bwForCluster NEMO: <https://www.top500.org/system/178839> (besucht am 21.08.2018).

⁷Zwölf Glasfaserkabel werden pro Schrank herausgeführt, je vier Kabel pro Omni-Path-Switch und zwei pro Ethernet-Switch.

⁸Web-Präsenz des parallelen Cluster File Systems BeeGFS: <https://www.beegfs.io> (besucht am 12.02.2019).

Dem Laden und Starten des eigentlichen Betriebssystems ist ein Boot-Auswahl-Server vorgelagert, der es erlaubt in den Bootvorgang einzugreifen (Bauer, Messner u. a., 2019). Damit wird es möglich, Rechenknoten kurzfristig in einen anderen Betriebsmodus zu versetzen oder ohne großen Aufwand eine aktualisierte Softwareumgebung für NEMO auszuprobieren.

Im Rahmen des »ViCE-Projekts«⁹ (Bauer, Suchodoletz u. a., 2019) wurde die Unterstützung von Virtualisierung auf Basis von OpenStack (Meier, 2017; Suchodoletz, Wiebelt, Meier u. a., 2017) etabliert. Das beinhaltet die Bereitstellung geeigneter Cloud-Infrastruktur-Komponenten und die Provisionierung der entsprechenden Softwarepakete auf den Rechenknoten. Die Virtualisierten Forschungsumgebungen der CMS- bzw. ATLAS-Gruppen (»A Toroidal LHC ApparatuS« Experiment am CERN) der Experimentellen Elementarteilchenphysik nutzen die auf NEMO verfügbare OpenStack-Infrastruktur und steuern diese mittels des Ressourcebrokers ROCED (Suchodoletz, Wiebelt, Meier u. a., 2017; Bühler u. a., 2018). Die Auslastung des Clusters NEMO durch VFUs in den ersten zweieinhalb Jahren ist in Abbildung 6 dargestellt.

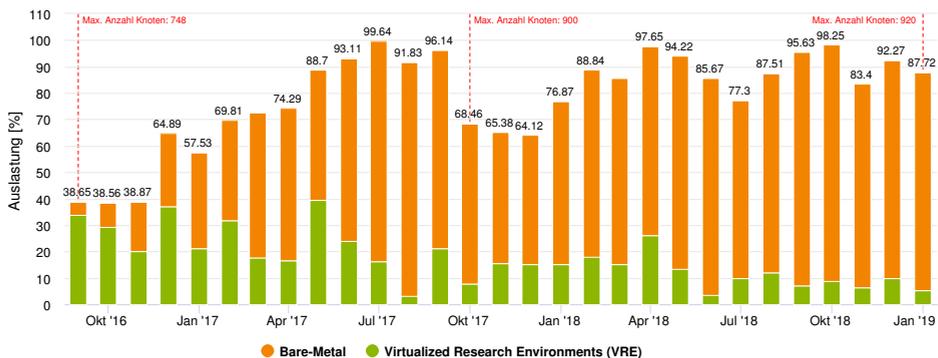


Abbildung 6: Auslastung des Clusters NEMO durch VFUs (Schätzung).

4.2 Stakeholder, Shareholder und Fairshare

Damit die Investitionen der Shareholder auf den Cluster abgebildet werden und gleichzeitig die Stakeholder aus den Forschungsschwerpunkten ihren Anteil nutzen können, muss die Clusternutzung kontingentiert werden. Hier spielt das »Fairshare-

⁹Gefördert im Rahmen der zweiten Linie der eScience-Initiative des Landes Baden-Württemberg.

Modell« eine Rolle (Wiebelt u. a., 2016).¹⁰ Die Stake- und Shareholdergruppen bekommen einen Anteil am Cluster zugeteilt. Dabei wird der Stakeholderanteil auf die Arbeitsgruppen der ENM-Schwerpunkte aufgeteilt. Die Shareholder bekommen den Teil ihres Investments am Gesamtcluster zugeteilt. Diese Einstellung wird im Accounting des Schedulers konfiguriert. Derzeit beträgt der Shareholderanteil am Cluster etwa 28 %. Die Stakeholderanteile verteilen sich gleichmäßig auf die aktiven Rechenvorhaben. Im Monat Januar 2019 waren 26 Arbeitsgruppen auf Seiten der Stakeholder aktiv und hatten je 2,77 % Anteil am Cluster (»Fairshare Value«, Abbildung 7a).

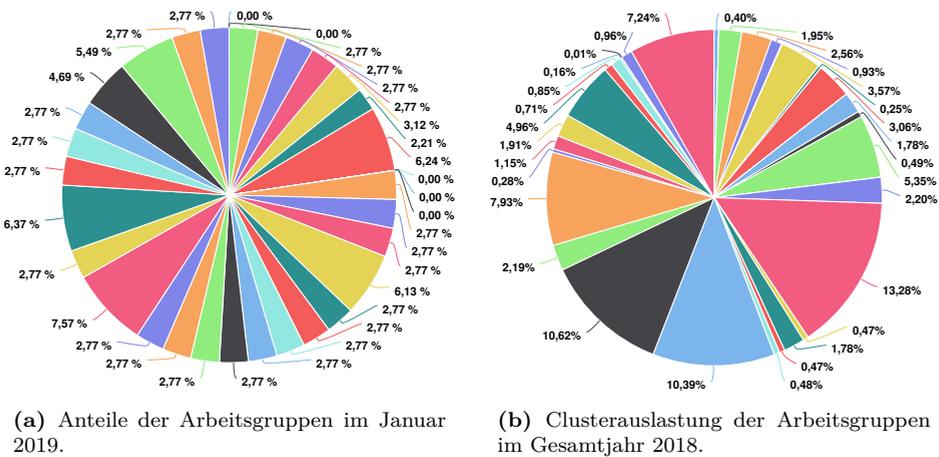


Abbildung 7: Die Anteile der jeweiligen Arbeitsgruppen an NEMO und deren Auslastung des Clusters im Jahr 2018.

Arbeitsgruppen können bis zur Höhe ihres Anteils den Cluster nutzen. Wenn sie unterhalb ihres Anteils liegen, bekommen die Jobs eine höhere Priorität in der Warteschleife und werden beim Scheduling bevorzugt. Liegen sie jedoch über ihrem Share, bekommen die Jobs negative Prioritäten, werden aber trotzdem in die Warteschlange eingefügt und können nach Abarbeitung der höher priorisierten Jobs anlaufen.

Die Anteilseigner geben der Gemeinschaft ihre Anteile. Diese kann damit diese Anteile mitbenutzen sogar über ihren eigenen Share hinaus. Die Anteilseigner können dann wiederum bei Bedarf über ihren Anteil hinaus den Cluster verwenden.

¹⁰Vgl. Abschnitt 5.3, Seite 419 ff. der Scheduler-Dokumentation von Adaptive Computing Enterprises, Inc (2018).

den. Damit ein langes Ansparen von Rechenzeit nicht möglich ist, wird der aktuelle Fairshare-Wert jeweils über die letzten drei Monate berechnet. Dabei werden 32 Zeitschritte zu je drei Tagen mit einem Verfallsfaktor von 0,95 je Schritt verwendet.

Ein Beispiel zur Berechnung ist in Auflistung 1 dargestellt. Das aktuelle Intervall wird voll eingerechnet, während die drei Tage zuvor nur noch zu 95 % eingerechnet werden. Dabei entspricht »Target« dem Anteil der Arbeitsgruppe und (%) dem derzeitigen Fairshare-Wert der Gruppe.

FSInterval	% Target	0	1	2	3	4	5	6	7	..	
FSWeight	---	---	1.00	0.95	0.90	0.85	0.81	0.77	0.73	0.69	..
AGx	9.92	2.77	9.50	11.21	10.73	9.86	9.75	9.68	9.88	9.71	..

Auflistung 1: Ausgabe eines beispielhaften Fairshare-Werts inkl. Intervallen mit dem Kommando `mdiag -f`.

Die aktuelle Verteilung der Stake- und Shareholder ist in Abbildung 7a dargestellt. Vergleicht man diese Aufteilung mit der tatsächlichen Verteilung bei der Auslastung im Jahr 2018 (Abbildung 7b), kann man das Prinzip des Fairshare gut erkennen. Arbeitsgruppen, die einen permanent hohen Rechenbedarf haben, profitieren davon, dass nicht alle Gruppen ständig ihren vollen Anteil ausschöpfen. Im Ausgleich dazu können Arbeitsgruppen mit stark schwankendem Rechenbedarf schneller bedient werden oder punktuell deutlich mehr Ressourcen bekommen, als es ihrem Anteil zusteht. Die inaktiven Rechenvorhaben werden ohne Anteile dargestellt (0,00 %).

Sollten in Zukunft zusätzliche Steuerungsmöglichkeiten notwendig werden, stehen mit »Preemption« (Adaptive Computing Enterprises, Inc, 2018, Kapitel 21) oder »Rollback Reservations« (Adaptive Computing Enterprises, Inc, 2018, Abschnitt 6.6.2) zwei Möglichkeiten zur Verfügung, Quality-of-Service-Anforderungen im Scheduling durchzusetzen.

4.3 NEMO Erweiterungen

NEMO wurde bereits in der Antragsphase um Eigenanteile von Forschungsgruppen erweitert, die am Standort Freiburg ebenfalls in die vorgenannte Zuordnung fielen. Zudem gab und gibt es erneute Beteiligungen nach Inbetriebnahme, wie beispiels-

weise das FIT¹¹ oder die ATLAS-Arbeitsgruppen, welche die Hochleistungsrechenressource durch eigene finanzielle Beteiligungen vergrößert und sich damit erweiterte Nutzungsrechte erworben haben (Suchodoletz, Wesner u. a., 2016).

Eine weitere, geplante Entwicklungsrichtung besteht in der Einrichtung von Visualisierungsknoten, die eine entfernte grafische Ausgabe auf den Geräten der Forschenden erlauben, ohne hierzu signifikante Anforderungen an diese zu stellen. Um ein leichtes Deployment und die Koexistenz verschiedener Gruppen zu erlauben, werden Container-basierte Ansätze im Rahmen des ViCE-Projekts entwickelt und erprobt.

Neu seit 2019 ist eine Maschine mit zwei Höheneinheiten einer AMD-CPU mit 32 Kernen sowie acht NVIDIA Tesla V100 Grafikkarten. Diese wird nach einer kurzen Testphase der NEMO-Community für Machine-Learning-Applikationen zur Verfügung gestellt werden.

4.4 Betriebsstatistiken

Ein wesentlicher Qualitätsmaßstab für die Forscher*innen ist die Wartezeit bis zum Start der eigenen Jobs. Durch ein strategisches Herangehen bei der Beschaffung und die Nutzung der Option der Aufnahme neuer Shareholder durch Aufwuchsfiananzierung konnte eine langfristig gute Auslastung mit regelmäßigen Erweiterungen des Systems verbunden werden. So wurden schrittweise in den letzten zweieinhalb Jahren zusätzliche Kapazitäten geschaffen, welche die Auslastung in einem für die Forschenden vorteilhaften Rahmen hielt.¹²

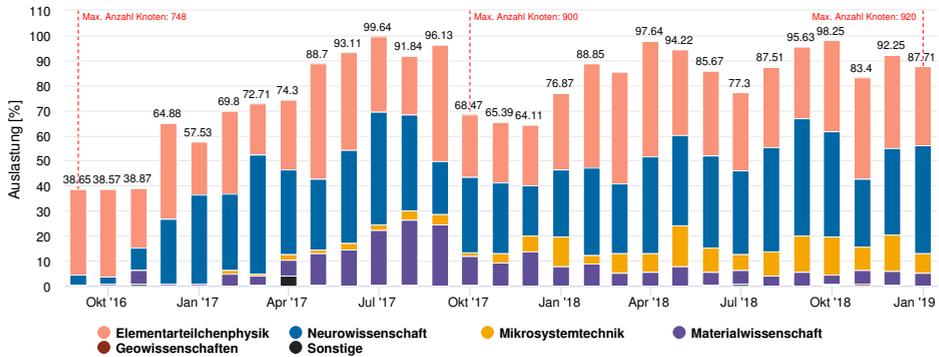
Abbildung 8 stellt die Auslastung des Clusters der ersten zweieinhalb Jahre aufgeschlüsselt nach Fachbereichen (8a) und nach Standorten (8b) dar. Zu Beginn bestand der Cluster aus 748 Rechenknoten. Im Oktober 2017 wurde der Cluster in einer ersten Erweiterung um 152 auf insgesamt 900 Rechenknoten vergrößert. Die Auslastung fiel dabei wieder auf ein niedrigeres Niveau, da mehr Knoten zur Verfügung standen, die die Forschenden zunächst wieder auslasten mussten.¹³ Eine weitere kleinere Erweiterung erfolgte schließlich im Januar 2019 um 20 weitere Rechenknoten. Alle Erweiterungen erfolgten mit Rechenknoten des gleichen Typs, um

¹¹Freiburger Zentrum für interaktive Werkstoffe und bioinspirierte Technologien, <http://www.fit.uni-freiburg.de> (besucht am 12.02.2019).

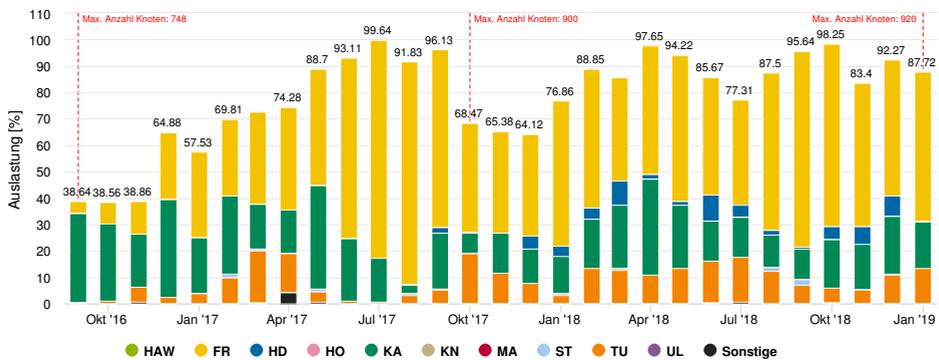
¹²Durch Erweiterungen konnte die Warteschlange der Jobs jeweils verkürzt werden.

¹³Bei der Erweiterung im Oktober 2017 kam es beim Update der BeeGFS-Server zu Problemen, so dass der Cluster mehrere Tage offline war.

die Aufgaben auf der administrativen Seite zu vereinfachen. Es gab lediglich eine Erweiterung des Hauptspeichers bei acht Rechenknoten und die Beschaffung des Machine-Learning-Knotens mit acht NVIDIA Tesla V100 Grafikbeschleunigern.¹⁴



(a) Auslastung bwForCluster NEMO nach Fachgebieten.



(b) Auslastung bwForCluster NEMO nach Standorten.

Abbildung 8: Auslastung des bwForClusters NEMO in den ersten zweieinhalb Jahren.

5 Aktuelle und zukünftige Entwicklungen

Dynamische Entwicklungen im »Scientific Computing« generieren fortlaufend neue Anforderungen an den Betrieb großer Forschungsinfrastrukturen wie NEMO. Der

¹⁴Eine Insel mit 44 Rechenknoten wurde bereits zu Beginn mit einem Mainboard gekauft, das doppelt so viele Speicherbausteine aufnehmen kann, um diese Erweiterungen zu ermöglichen. Dabei wurden alle RAM-Riegel wiederverwendet (4×512 GiB ergibt 4×256 GiB ohne Zusatzkosten).

sich reduzierende technologische Abstand und die Wahrnehmung seitens der Nutzer*innen lösen die klassische Dichotomie von Cloud und HPC zusehends auf. Die Softwareumgebungen gleichen sich vermehrt und für spezielle Aufgaben, wie Pre- oder Post-Processing oder »Remote Visualization« wird die Cloud als Compute-Plattform zunehmend relevant.

Das steigende Interesse am maschinellen Lernen in immer mehr Fachdisziplinen führt dazu, dass das »General-purpose computing on graphics processing units« (GPGPU) immer stärker nachgefragt wird, weshalb Grafikbeschleuniger einer breiteren Nutzerschicht – zu Beginn auch zu Evaluations- und Testzwecken – zugänglich gemacht werden sollten. Deshalb wurde 2019 ein erster Knoten für das Machine-Learning beschafft.

Mit dem steigenden Umfang von Forschungsdaten in einzelnen Projekten und im Wissenschaftsbetrieb insgesamt wird die Lokalität der Daten wieder relevant, da es sehr ineffizient sein kann, große Datenmengen für verhältnismäßig kleine Berechnungen über lange Strecken zu kopieren. Dieses unter dem Stichwort »Data Intensive Computing« beschriebene Phänomen erfordert ein verstärktes »Zusammendenken« der Forschungsinfrastrukturen Compute und Datenhaltung (Schneider u. a., 2019) und wird am NEMO-Standort durch die Beteiligung am Infrastrukturprojekt bwSFS (»Storage for Science«) gemeinsam mit dem BinAC-Standort Tübingen vorangetrieben (Suchodoletz, Hahn u. a., 2019).

Durch Anbinden zusätzlicher lokaler Wissenschaftsspeicher wie bwSFS lassen sich cluster-lokale Parallelspeicher wie BeeGFS als schnelle, kurzfristige Zwischenspeicher verwenden. Arbeitsordner können mehr als bisher rein zu Berechnungen genutzt werden und nach Jobende mit Metadaten versehen an weitere Speicher weitergeleitet werden. Dadurch muss der am Cluster direkt angeschlossene Parallelspeicher nicht mehr so groß bemessen werden und kann bereits bei kleinen und mittelgroßen Clustern aus Solid State Disks bestehen.

Da sich auf Tier-Ebene 3 die Architektur auf X86 beschränkt und konkrete Vorhersagen für langfristige Bedarfe einzelner Compute-Umgebungen schwer zu treffen sind, arbeitet das Betriebsteam am Standort Freiburg an zukünftigen Deployment-Modellen (Bauer, Messner u. a., 2019), die von einem sehr einfachen Basis-Setup eines Rechenknotens ausgehen. Hierzu wird dieser wie bisher über das lokale Netzwerk »gebootet« und mit einer sehr schlanken Softwareausstattung versehen. Gleichzeitig wird der Knoten in die jeweiligen Netzwerkkumgebung mit Zugriff auf die relevanten

Ressourcen versetzt. Davon abhängig sollen die gebooteten Knoten in einem weiteren Schritt für eine HPC- oder Cloud-Nutzung konfiguriert werden. Zusätzliche Softwaremodule für konkretere Nutzungsszenarien, wie der Einsatz von GPUs oder die Verwendung als Login-Knoten werden bedarfsbezogen in einem weiteren Schritt dynamisch nachgezogen.

Für die HPC-Knoten ist angedacht, in weiteren Schritten die von den Forschenden erwarteten Softwareumgebungen sowohl über das traditionelle Modulsystem jedoch zunehmend auch durch Containerisierung verfügbar zu machen. Gerade durch Letzteres können individuelle Forschungsgruppen ihre Rechenumgebungen selbst zusammenstellen. Sie können diese einer eigenen Versionskontrolle unterstellen und damit die Reproduzierbarkeit ihrer Forschungsworkflows verbessern.

Weiterhin wird für HTC-Jobs auf einen verstärkten Einsatz von Cloud-Ressourcen gesetzt. Dieses erlaubt deutlich längere Walltimes¹⁵ als sie in der aktuellen HPC-Umgebung von NEMO gewährt werden können. Ebenso bieten sich Cloud-Ressourcen für interaktive Jobs an. Aus der verstärkten administrativen Zusammenführung der Ressourcen entstehen neue Herausforderungen im Scheduling und in der Abrechnung. Hier besteht noch Forschungs- und Entwicklungsbedarf, der im Zuge des Projektes bwHPC-S5 angegangen wird (Barthel u. a., 2019).

In den aktuellen Überlegungen wird dabei noch nicht an eine automatische Rekonfiguration des Gesamtsystems gedacht. Jedoch soll eine deutlich höhere Dynamisierung der Ressourcenzuteilung je nach aktuellen Projekten und Anforderungen seitens der Forschenden erreicht werden. Nicht genutzte Cloud-Ressourcen können beispielsweise dafür verwendet werden, um eine sich aufgestaute Queue im HPC abzuarbeiten. Zusätzlich erlaubt die dynamische Ressourcenzuteilung eine verstärkte Berücksichtigung von Green-IT-Elementen. So lässt sich nicht nur eine schnelle Integration neuer Knoten ins Gesamtsystem erreichen, sondern auch eine verbesserte »Packung« der verschiedenen Compute-Jobs auf dem Gesamtsystem. Partielle Unterauslastungen können vermieden werden und die durch Konsolidierung frei gezogenen Knoten temporär bei Nicht-Nutzung außer Betrieb nehmen. So wurden bereits temporär zusätzliche Ressourcen NEMO zur Verfügung gestellt, um längere Queues unter der Woche abzuarbeiten.¹⁶ Dabei müssen die Geldgeber und deren Anforderungen jeweils berücksichtigt werden, so dass in der Endabrechnung die Kontingente der einzelnen Parteien wieder stimmen.

¹⁵Gesamtlaufzeit eines Jobs.

¹⁶Vergleiche beispielsweise Monate Juli 2017 oder Oktober 2018 in Abbildung 8.

6 Fazit und Ausblick

Mit der Zunahme verteilter Landesinfrastrukturen (Schneider u. a., 2019) – Freiburg arbeitet gemeinsam mit den Kollegen aus Tübingen an der Bereitstellung eines größeren an die HPC-Cluster angedockten Speichersystems mit Forschungsdatenmanagementkomponente (bwSFS) – werden Fragen der Steuerung und Abstimmung relevanter. Hier kann die erfolgreiche Governance auf Landesebene, insbesondere auch die von NEMO eine Vorlage bieten.

Das Basissystem des Clusters wurde nicht nur als stabile Softwareumgebung für die rechnenden Communitys genutzt, sondern bildete ebenso die Grundlage für Experimente und das Sammeln von Erfahrungen im Umgang mit Virtualisierten Forschungsumgebungen im Rahmen von ViCE (Meier, 2017). Die VFUs erlauben Forschenden eine komplett eigene Softwareumgebung zu nutzen, wie sie beispielsweise für bestimmte Rechnungen der ATLAS- oder CMS-Gruppen notwendig ist. Als Alternative zur vollständigen Virtualisierung bietet sich die Containerisierung an, die ebenfalls eigene Softwareerweiterungen oder komplette Forschungsumgebungen ermöglicht. Diese könnten in zukünftigen Betriebsmodellen eine stärkere Konvergenz von HPC und Cloud für das High Throughput Computing erlauben. Diese Überlegungen lassen sich zudem auf zukünftige wissenschaftliche Workflows anwenden, in denen verschiedene Forschungsinfrastrukturen nacheinander genutzt werden und beispielsweise ein Pre- oder Post-Processing in der Cloud ebenso vorsehen wie die Remote-Visualisierung auf einem spezialisierten System.

Mit der Entwicklung des Boot-Auswahl-Servers werden die bereits mit ViCE angefangenen Optionen flexibler Betriebsmodelle weitergeführt (Bauer, Messner u. a., 2019). Die Basiskonfiguration nutzt dabei Entwicklungen aus anderen Landesprojekten wie bwLehrpool (Suchodoletz, Münchenberg u. a., 2014) nach. Das »Distributed Network Block Device Version 3« (DNBD3) bietet spezielle Funktionalität für die performante Versorgung einer großen Zahl von Cluster-Knoten mit einem Root-Filesystem als auch für das Failover für den Fall der Nichterreichbarkeit eines DNBD3-Servers im Verbund (Rettberg u. a., 2019).

Auch für die Restlaufzeit von NEMO ist weiterhin eine schrittweise Erneuerung durch abgestimmte Ersatzinvestitionen und Äquivalenztausch¹⁷ angedacht. Die ur-

¹⁷In dem Sinne, dass Erweiterungen beispielsweise in Form von Arbeitsspeicher für bestehende Systeme erfolgen, wenn hier der größte Bedarf und gemeinsame Nutzen besteht. Hierfür erhält der Investierende Anteile aus dem bestehenden Gesamtpool statt beispielsweise neue Knoten hinzuzufügen.

sprünglich mehr als zwei Jahre stabil gehaltene Hardware-Landschaft wird dabei diverser; sowohl die Bereitstellung des GPGPUs als Compute-Alternative als auch die Aufnahme der AMD-Plattform und eine Anzahl von Memory-Knoten macht die Auswahl für die Forschenden größer. Anstehende Technologiewechsel werden bei zukünftigen Erweiterungen entsprechend berücksichtigt und mit den Share- und Stakeholdern abgestimmt.

Danksagungen

An dieser Stelle möchten die Autoren dem Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg und der DFG für die finanzielle Förderung des bwForClusters NEMO und die hervorragende Unterstützung der Projekte ViCE und bwHPC-S5 danken.

Korrespondenzautor

Michael Janczyk: michael.janczyk@rz.uni-freiburg.de
eScience Abteilung, Rechenzentrum Albert-Ludwigs-Universität Freiburg,
Hermann-Herder-Str. 10, 79104 Freiburg, Deutschland

ORCID

Michael Janczyk  <https://orcid.org/0000-0003-4886-736X>
Dirk von Suchodoletz  <https://orcid.org/0000-0002-4382-5104>
Bernd Wiebelt  <https://orcid.org/0000-0003-2771-4524>

Lizenz    4.0 <https://creativecommons.org/licenses/by-sa/4.0>

Literatur

Adaptive Computing Enterprises, Inc (2018). *Moab Workload Manager. Administrator Guide 9.1.3*. Administrator Guide. Version 9.1.3. Adaptive Computing Enterprises, Inc. URL: <http://docs.adaptivecomputing.com/9-1-3/MWM/Moab-9.1.3.pdf> (besucht am 12.02.2019).

- Barthel, R. und J. Salk (2019). »bwHPC-S5: Scientific Simulation and Storage Support Services. Unterstützung von Wissenschaft und Forschung beim leistungsstarken und datenintensiven Rechnen sowie großskaligem Forschungsdatenmanagement«. In: *Proceedings of the 5th bwHPC Symposium. HPC Activities in Baden-Württemberg*. Freiburg, September 2018. 5th bwHPC Symposium. Hrsg. von M. Janczyk, D. von Suchodoletz und B. Wiebelt. TLP, Tübingen, S. 17–28. DOI: [10.15496/publikation-29039](https://doi.org/10.15496/publikation-29039).
- Bauer, J., M. Messner u. a. (2019). »A Sorting Hat For Clusters. Dynamic Provisioning of Compute Nodes for Colocated Large Scale Computational Research Infrastructures«. In: *Proceedings of the 5th bwHPC Symposium. HPC Activities in Baden-Württemberg*. Freiburg, September 2018. 5th bwHPC Symposium. Hrsg. von M. Janczyk, D. von Suchodoletz und B. Wiebelt. TLP, Tübingen, S. 217–229. DOI: [10.15496/publikation-29055](https://doi.org/10.15496/publikation-29055).
- Bauer, J., D. von Suchodoletz, J. Vollmer und H. Rasche (2019). »Game of Templates. Deploying and (re-)using Virtualized Research Environments in High-Performance and High-Throughput Computing«. In: *Proceedings of the 5th bwHPC Symposium. HPC Activities in Baden-Württemberg*. Freiburg, September 2018. 5th bwHPC Symposium. Hrsg. von M. Janczyk, D. von Suchodoletz und B. Wiebelt. TLP, Tübingen, S. 245–262. DOI: [10.15496/publikation-29057](https://doi.org/10.15496/publikation-29057).
- Bührer, F. u. a. (2018). »Dynamic Virtualized Deployment of Particle Physics Environments on a High Performance Computing Cluster«. In: *Computing and Software for Big Science*. arXiv: [1812.11044](https://arxiv.org/abs/1812.11044) [physics.comp-ph].
- Hartenstein, H., T. Walter und P. Castellaz (2013). »Aktuelle Umsetzungskonzepte der Universitäten des Landes Baden-Württemberg für Hochleistungsrechnen und datenintensive Dienste«. In: *PIK – Praxis der Informationsverarbeitung und Kommunikation* 36.2. DOI: [10.1515/pik-2013-0007](https://doi.org/10.1515/pik-2013-0007).
- Intel Corporation (2018). *Intel Math Kernel Library for Linux. Intel MKL 2019 – Linux*. Developer Guide. Version 2019, Revision 065. Intel Corporation. URL: <https://software.intel.com/sites/default/files/mkl-2019-developer-guide-linux.pdf> (besucht am 08.02.2019).
- Meier, K. (2017). »Infrastrukturkonzepte für virtualisierte wissenschaftliche Forschungsumgebungen«. Diss. Albert-Ludwigs-Universität Freiburg im Breisgau.
- Rettberg, S., D. von Suchodoletz und J. Bauer (2019). »Feeding the Masses: DNBD3. Simple, efficient, redundant block device for large scale HPC, Cloud and PC pool installations«. In: *Proceedings of the 5th bwHPC Symposium. HPC Activities in Baden-Württemberg*. Freiburg, September 2018. 5th bwHPC Symposium. Hrsg. von M. Janczyk, D. von Suchodoletz und B. Wiebelt. TLP, Tübingen, S. 231–243. DOI: [10.15496/publikation-29056](https://doi.org/10.15496/publikation-29056).

- Schmelzer, S., D. von Suchodoletz, M. Janczyk und G. Schneider (2014). »Flexible Cluster Node Provisioning in a Distributed Environment«. In: *Hochleistungsrechnen in Baden-Württemberg. Ausgewählte Aktivitäten im bwGRiD 2012*. Beiträge zu Anwenderprojekten und Infrastruktur im bwGRiD im Jahr 2012. Hrsg. von J. C. Schulz und S. Hermann. KIT Scientific Publishing, Karlsruhe, S. 203–219. ISBN: 978-3-7315-0196-1. DOI: 10.5445/KSP/1000039516. URN: urn:nbn:de:0072-395167.
- Schneider, G. u. a. (2019). »Umsetzungskonzept der Universitäten des Landes Baden-Württemberg für das High Performance Computing (HPC), Data Intensive Computing (DIC) und Large Scale Scientific Data Management (LS²DM)«. Gekürzte Fassung. In: *Proceedings of the 5th bwHPC Symposium. HPC Activities in Baden-Württemberg*. Freiburg, September 2018. 5th bwHPC Symposium. Hrsg. von M. Janczyk, D. von Suchodoletz und B. Wiebelt. TLP, Tübingen, S. 3–16. DOI: 10.15496/publikation-29040.
- Suchodoletz, D. von, U. Hahn, B. Wiebelt, K. Glogowski und M. Seifert (2019). »Storage infrastructures to support advanced scientific workflows. Towards research data management aware storage infrastructures«. In: *Proceedings of the 5th bwHPC Symposium. HPC Activities in Baden-Württemberg*. Freiburg, September 2018. 5th bwHPC Symposium. Hrsg. von M. Janczyk, D. von Suchodoletz und B. Wiebelt. TLP, Tübingen, S. 263–279. DOI: 10.15496/publikation-29058.
- Suchodoletz, D. von, J. Münchenberg u. a. (2014). »bwLehrpool – ein landesweiter Dienst für die Bereitstellung von PC-Pools in virtualisierter Umgebung für Lehre und Forschung«. In: *PIK – Praxis der Informationsverarbeitung und Kommunikation* 37.1, S. 33–40. DOI: 10.1515/pik-2013-0046.
- Suchodoletz, D. von, S. Wesner und G. Schneider (2016). »Überlegungen zu laufenden Cluster-Erweiterungen in bwHPC«. In: *Kooperation von Rechenzentren: Governance und Steuerung – Organisation, Rechtsgrundlagen, Politik*. Hrsg. von D. von Suchodoletz, J. C. Schulz, J. Leendertse, H. Hotzel und M. Wimmer. De Gruyter, S. 331–342. ISBN: 978-3-11-045888-6. DOI: 10.1515/9783110459753.
- Suchodoletz, D. von, B. Wiebelt und M. Janczyk (2017). »bwHPC Governance of the ENM community«. In: *Proceedings of the 3rd bwHPC-Symposium*. (2016). Hrsg. von S. Richling, M. Baumann und V. Heuveline. Heidelberg: heiBOOKS. DOI: 10.11588/heibooks.308.418.
- Suchodoletz, D. von, B. Wiebelt, K. Meier und M. Janczyk (2017). »Flexible HPC: bwForCluster NEMO«. In: *Proceedings of the 3rd bwHPC-Symposium*. (2016). Hrsg. von S. Richling, M. Baumann und V. Heuveline. Heidelberg: heiBOOKS. DOI: 10.11588/heibooks.308.418.

- Wesner, S., T. Walter, B. Wiebelt, D. von Suchodoletz und G. Schneider (2016). »Strukturen und Gremien einer bwHPC-Governance – Momentaufnahmen und Perspektiven«. In: *Kooperation von Rechenzentren Governance und Steuerung – Organisation, Rechtsgrundlagen, Politik*. Hrsg. von D. von Suchodoletz, J. C. Schulz, J. Leendertse, H. Hotzel und M. Wimmer. de Gruyter, S. 315–329. ISBN: 978-3-11-045888-6. DOI: 10.1515/9783110459753-027.
- Wiebelt, B. u. a. (2016). »Strukturvorschlag für eine bwHPC-Governance der ENM-Community«. In: *Kooperation von Rechenzentren Governance und Steuerung – Organisation, Rechtsgrundlagen, Politik*. Hrsg. von D. von Suchodoletz, J. C. Schulz, J. Leendertse, H. Hotzel und M. Wimmer. de Gruyter, S. 343–354. ISBN: 978-3-11-045888-6. DOI: 10.1515/9783110459753-029.