Anhang 07: Vorlage für die Vereinbarung über Monitoring und Protokollierung der Verfügbarkeit des Systems während der Abnahmephase

1. Nutzbarkeit des Systems

Die folgende Regelung der Nutzbarkeit gilt ausschließlich für die Abnahmephase, nicht für den Regelbetrieb.

Das System (als Ganzes) gilt als nutzbar, wenn alle diese Bedingungen zutreffen:

- Mindestens 90% der GPU-Knoten sind vollumfänglich nutzbar.
- Mindestens sieben Serviceknoten sind vollumfänglich nutzbar.
- Alle Netzwerke (HPC-Interconnect, Ethernet) sind vollumfänglich verfügbar.
- Alle für die Nutzer zugreifbaren Dateisysteme sind verfügbar. Nicht vom Auftragnehmer verursachte Einschränkungen werden dabei nicht berücksichtigt.
- Das System-Monitoring funktioniert vollumfänglich.

2. Testperiode

Die Funktionsprüfung beginnt mit dem auf die Erklärung der Betriebsbereitschaft folgenden Arbeitstag. Der Test endet nach 30 Kalendertagen. Diese Zeit wird im Folgenden Testperiode genannt. Gelingt der Nachweis der Verfügbarkeit nicht, so kann bei Konsens zwischen Auftraggeber und Auftragnehmer der Test der Verfügbarkeit während der Aufnahme des Regelbetriebs fortgesetzt werden, bis die geforderte Verfügbarkeit an 30 aufeinander folgenden Kalendertagen erreicht wird. Sollte die Betriebsruhe der TU Dresden zum Jahreswechsel in den benötigten Zeitraum von 30 Kalendertagen fallen, so wird der Test mit Beginn der Betriebsruhe unterbrochen und im Anschluss an die Betriebsruhe fortgesetzt. Der Test verlängert sich in dem Fall um die entsprechende Anzahl von Tagen.

Für die Testperiode gelten Ausfälle, die nicht vom Auftragnehmer zu vertreten sind, nicht als Unterbrechung der Funktionsprüfung. Die Testperiode verlängert sich auf Verlangen des Auftraggebers um diese Zeit, es sei denn, dass der Auftraggeber die Unterbrechung zu vertreten hat.

Als Prüfzeit in der Testperiode gelten die Bürozeiten von 8:00 bis 16:00 Uhr an ortsüblichen Arbeitstagen. Die Nutzungszeit ist der Teil der Prüfzeit, in der das System verfügbar (im Sinne von 1.) ist. Als Ausfallzeit in der Testperiode gilt der Teil der Prüfzeit, in der das System nicht verfügbar ist. (Prüfzeit=Ausfallzeit+Nutzungszeit)

Als Neutralzeiten gelten die Zeiten der Testperiode außerhalb der Prüfzeit.
(Testperiode=Prüfzeit+Neutralzeit)

Parallele Jobs, die auf vergleichbaren Anlagen erfolgreich getestet wurden, müssen über das Batchsystem abgesetzt werden können. Ist dies während der Testperiode begründet gefährdet, hat der Auftraggeber das Recht, die Testperiode zu unterbrechen, bis die Mängel behoben sind. Diese Unterbrechung zählt nicht als Unterbrechung im Sinne des folgenden Abschnitts, die benötigte Zeit geht als Neutralzeit in keine Wertung ein, die Testperiode wird jedoch um diese Zeit verlängert.

Der Auftragnehmer sowie – mit Zustimmung des Auftraggebers eingesetzte – Subunternehmer des Auftragnehmers haben an jedem Arbeitstag der Testperiode von mindestens 8:00 bis 16:00 Uhr Zugang zum Maschinenraum. Eine Verlängerung des Zugangs bis ca. 18:00 Uhr ist in Absprache mit dem Betreiber in Einzelfällen realisierbar, kann aber nicht garantiert werden. Entsprechender Bedarf ist möglichst frühzeitig anzukündigen. Wenn der Zugang zum Rechnerraum eine Stunde vor und eine Stunde nach der Prüfzeit nicht gewährleistet werden kann (z.B. durch Krankheit oder Urlaub des Vor-

Ort Personals) und dadurch Wartungsarbeiten verhindert werden, so verkürzt sich die Prüfzeit entsprechend. Zugang zum Untergeschoss (Infrastruktur Elektro/Kälte) besteht i.d.R. nur bis 15:45 Uhr.

3. Betriebsunterbrechungen

Sämtliche Betriebsunterbrechungen und Fehler werden gemeinsam protokolliert. Als Ablageort wird der Ordner „Abnahme“ in einem gemeinsam geführten Repository benutzt.

Planmäßige Unterbrechungen

Der Auftragnehmer kann innerhalb der Prüfzeit pro Kalenderwoche eine Wartung des Systems vornehmen. Die Inanspruchnahme dieser Wartung muss mit einer Frist von 24 Stunden angekündigt werden. Die Testperiode verlängert sich bei Inanspruchnahme der Wartung entsprechend. Die fristgerechte Inanspruchnahme der Wartung zählt nicht als Unterbrechung.

Der Abbruch von zum Beginn einer planmäßigen Unterbrechung noch laufenden Jobs wird nicht gewertet. Darüber hinaus können weitere Unterbrechungen einvernehmlich zwischen Auftraggeber und Auftragnehmer vereinbart werden.

Der Auftragnehmer hat das Recht, außerhalb der Prüfzeit proaktive Wartungsmaßnahmen durchzuführen, wenn dadurch laufende Jobs nicht unterbrochen werden. Es dürfen zum Beispiel Knoten proaktiv ausgetauscht werden, falls Diagnose-Tools, die zum Lieferumfang gehören, über Unregelmäßigkeiten in Bezug auf diese Knoten berichten.

Außerplanmäßiger Unterbrechungen

Ungewollte Betriebsunterbrechungen, die auf die Hardware und den Software-Stack des Auftragnehmers zurückzuführen sind, werden als Ausfälle bezeichnet. Ausfälle werden für jedes Teilsystem (GPU-Knoten, Serviceknoten, Monitoring, Dateisystem, Netzwerk, Management-Knoten) getrennt erfasst.

In dem Fall, dass ein Fehler den Absturz sowohl eines Dateisystems als auch einer Rechenkomponente (oder Teilen davon) verursacht, wird er nur beim Dateisystem und nicht bei der Rechenkomponente erfasst, wenn ein Absturz des Dateisystems ursächlich für den Ausfall der Rechenknoten ist. In allen anderen Fällen zählt der Ausfall für beide Komponenten. Diese Kausalität ist vom Auftragnehmer plausibel darzulegen.

Ist eine Unterbrechung auf eine Komponente des Auftraggebers zurückzuführen (z.B. externe Netzwerk-Komponenten), werden die Ausfälle erfasst, aber nicht gewertet.

In der Prüfzeit ist die folgende Anzahl von Betriebsunterbrechungen zulässig:

- 15 außerplanmäßige Unterbrechungen in den Rechenkomponenten (Ausfall von Beschleunigern, GPU- oder Service-Knoten, Verluste der Verbindung zu den Dateisystemen, Batchsystem, etc.)
- 2 Ausfälle der Monitoring-Knoten, auch bei funktionierendem Failover,
- 1 Unterbrechung des Dateisystems.

4. Berechnung der gewichteten Ausfallzeiten

Ebenso wie die Anzahl der Ausfälle werden auch die Ausfallzeiten des Systems in zwei getrennten Listen erfasst, eine für die Rechenkomponente und eine für das Speichersystem. Dabei spielt es keine Rolle, ob ein Ausfall planmäßig oder außerplanmäßig ist.

Ausfälle, die beide Teilsysteme beeinflussen, werden in beiden Listen erfasst. Fallen Systeme aus, deren Funktionalität ohne Einschränkung von redundanten Systemen automatisch übernommen wird, so wird die Ausfallzeit nicht gewertet. Dies gilt jedoch nicht für die Service- und Compute-Knoten. Falls die Beschaffung von Ersatzteilen zur Behebung eines Fehlers notwendig ist, kann der Auftragnehmer dies dem Auftraggeber melden. Stimmt der Auftraggeber dem zu, so kann die Zeit für die Beschaffung der Ersatzteile als Neutralzeit für das gesamte System bewertet werden. Die Testperiode wird in diesem Fall um diese Länge der Neutralzeit verlängert.

Das System (Compute-Knoten, Switch, Server etc.) gilt als ausgefallen, wenn es nicht mehr vollumfänglich Benutzeraufträge bearbeiten kann. Das schließt auch Komponenten ein, die zwar für sich betrachtet voll funktionsfähig sind, jedoch aufgrund anderer Ausfälle nicht genutzt werden können (z.B. Compute-Knoten die aufgrund eines defekten Switches nicht mehr erreichbar sind). Als Ausfallzeit gilt die Zeit von der Fehlermeldung durch den Auftraggeber bis zur Meldung der erneuten Betriebsbereitschaft durch den Auftragnehmer (siehe Abschnitt Meldewege). Die Wiederherstellung von Compute-Knoten endet mit der Freischaltung des Knotens für Slurm (resume) durch den Auftragnehmer.

Ein Speichersystem gilt als nicht funktionsbereit, wenn Anfragen der Rechenkomponente an einen der Speicherbereiche nicht in angemessener Zeit beantwortet werden. Als nicht mehr angemessen gilt, wenn – während in dem System kein I/O-Benchmark läuft – das Kommando "ls -la" auf einem Rechenknoten nicht innerhalb von 15 Sekunden zurückkehrt. Die Zeit für die Instandsetzung eines Speicherbereichs (fsck, etc.) nach einem Fehler gilt ebenfalls als Ausfallzeit des Speichersystems. Berücksichtigt werden die jeweiligen Meldezeiten.

Ausfälle von Teilen des Systems werden anteilig berücksichtigt, falls das entsprechende Teilsystem weiterhin Benutzeraufträge bearbeiten kann. Sind Verzeichnisse des Speichersystems nicht erreichbar, zählt das als kompletter Ausfall. Bei Teilausfällen wird die reale Ausfallzeit t der betroffenen Komponenten mit einem Faktor ω gewichtet und die gewichtete Ausfallzeit T_A der Ausfallzeit des Systems zugeschlagen: $T_A = \omega * t$

Ausgefallene Komponente	Teilsystem	Gewichtsfaktor ω
GPU-Knoten	A - GPU-Knoten	1/<Anzahl der GPU-Knoten>
Service-Knoten	B - Service-Knoten	1/<Anzahl der Service-Knoten>
Storage-Server oder -VM	C - Storage	1/<Anzahl der Server bzw. VMs im Redundanz-Verbund>
Switch	D - Netzwerke	1/<Anzahl der Switche>
Monitoring-Knoten	E - Management-Knoten	1/2

Der vom Auftragnehmer zu verantwortende Zeitanteil T eingeschränkter Verfügbarkeit für das Gesamtsystem während der Prüfzeit berechnet sich aus: $T=T_A \,/\, (T_A+T_N)\cdot 100\ \%$, wobei T_A die Summe der gewichteten Ausfallzeiten des Teilsystems innerhalb der Prüfzeit und T_N die Summe der Nutzungszeiten innerhalb der Prüfzeit ist.

5. Erfolgreicher Zuverlässigkeitstest

Der Zuverlässigkeitstest gilt als erfolgreich bestanden, wenn alle folgenden Bedingungen erfüllt, sind:

- Die Nutzbarkeit des Gesamtsystems gemäß den in Abschnitt 1 definierten Bedingungen, während der Prüfzeit beträgt mindestens 99 %.

- Die Dateisysteme arbeiten frei von Inkonsistenzen und Datenkorruptionen.
- Der vom Auftragnehmer zu verantwortende Zeitanteil T eingeschränkter Verfügbarkeit für das Gesamtsystem während der Prüfzeit ist kleiner oder gleich 1 %.
- Während der Testperiode ist nicht mehr als 10 % der insgesamt zur Verfügung stehenden Rechenzeit durch Ausfälle verloren gegangen.
- Die Anzahl der vom Auftragnehmer zu verantwortenden außerplanmäßigen Unterbrechungen überschreitet nicht die oben angegebenen Werte in Abschnitt 3.

Vorlage des Protokolls

Beginn der Testperiode: __.__.____

Unterbrechungen:

Von	Bis	Grund	Verlängerung d. Testperiode

Ende der Testperiode: __.__.____

Nutzbarkeit des Gesamtsystems

Zeiten eingeschränkter Verfügbarkeit:

Von	Bis	Grund	gewichtete Ausfallzeit

Die erzielte Verfügbarkeit gemäß den Bedingungen

in Abschnitt 1 beträgt: ___ %.

Die geforderte Zuverlässigkeit von mindestens 99 % wurde

- ☐ erreicht
- ☐ nicht erreicht

Anzahl der Ausfälle

Ausgefallene Komponente	Anzahl
A - GPU-Knoten	
B - Service-Knoten	
C - Storage	
D - Netzwerke	
E - Management-Knoten	

Die maximale Anzahl erlaubter Unterbrechungen wurde

- ☐ überschritten
- ☐ nicht überschritten

Verfügbarkeit der Teilsysteme

Teilsystem	T_A
A - GPU-Knoten	
B - Service-Knoten	
C - Storage	
D - Netzwerke	
E - Management-Knoten	

Der zulässige Zeitanteil eingeschränkter Verfügbarkeit je Teilsystem wurde

- ☐ überschritten
- ☐ nicht überschritten

Verlorene Rechenzeit

Durch Ausfälle sind ___ % der Rechenzeit verloren gegangen.
Die maximale Ausfallzeit von 10 % wurde

- ☐ überschritten
- ☐ nicht überschritten

Bestehen der Zuverlässigkeitsprüfung

Kriterium	bestanden
Nutzbarkeit des Gesamtsystems gemäß Abs. 1 von 99 %	
Zulässige Gesamtanzahl von Ausfällen	
Verfügbarkeit des Gesamtsystems >=99 %	
Durch Ausfälle verlorene nutzbare Rechenzeit < 10 %	
Dateisysteme arbeiten frei von Inkonsistenzen und Datenkorruptionen	

Die Zuverlässigkeitsprüfung ist

- ☐ bestanden
- ☐ nicht bestanden

Meldewege

- Ausfälle werden gemeldet über das Ticketsystem oder über die E-Mail-Adresse _____
- Die Meldung über wiederhergestellte Funktionsfähigkeit von Komponenten geht an die E-Mail-Adresse _____