

HPC Ausfall: Zusammenfassung zu den Ereignissen der letzten Wochen

Die Ereignisse der vergangenen Wochen waren für alle Beteiligten – HPC-Nutzerinnen und -Nutzer und die HPC-Mitarbeiter des ZIM – eine schwere Belastung und haben unter den HPC-Nutzerinnen und -Nutzern vielleicht Zweifel an der Qualität des IT-Services „HPC“ aufkommen lassen. Daher wollen wir im Folgenden aufzeigen, was passiert ist und erklären, welche Maßnahmen wir ergriffen haben, damit sich solch ein katastrophaler Ausfall nicht wiederholen kann.

Was ist passiert?

Unsere Daten sind uns wichtig! Daher werden Daten an der HHU im HPC-Bereich mehrfach redundant gesichert: einerseits durch ein RAID-System, bei dem im laufenden Betrieb bis zu 2 von 10 Festplatten gleichzeitig ausfallen können, ohne dass der Betrieb eingeschränkt wird, und andererseits durch das Spiegeln der kompletten Daten zu regelmäßigen Zeitpunkten. Im Herbst letzten Jahres hat sich gezeigt, dass die damals eingesetzte Lösung zum Spiegeln der Daten mit den massiv gestiegenen Anforderungen hinsichtlich Anzahl der Dateien und Volumen der Daten nicht mehr zuverlässig funktionierte. Seit diesem Zeitpunkt wurden die Daten effektiv nur noch über das RAID-System gesichert. Um wieder die volle Absicherung zu erhalten, sollte auf ein anderes System zum Spiegeln der Daten umgestellt werden. Hierfür war vorher ein Firmware-Update der Speicherinfrastruktur nötig. Dieses Update wurde sorgfältig mit dem Hersteller der Speicherinfrastruktur geplant und abgestimmt. Schon vor dem Update lagen im System jedoch Hardware-Defekte vor, die vom Überwachungssystem der Hardware nicht erkannt bzw. nicht angezeigt wurden. So ist es beim Update-Vorgang zu Problemen gekommen, die einerseits zu Datenverlusten und andererseits zu dem langen Ausfall der HPC-Infrastruktur geführt haben. Ursache der Hardwaredefekte waren nach aktuellem Kenntnisstand unzureichende Vorkehrungen und Schutzmaßnahmen gegen Überspannungen in der Stromversorgung, wie sie bei Notstromtests auftreten können. Die genaue Fehleranalyse, sowohl für die Schäden an der Hardware als auch für die Probleme beim Firmware-Update, dauert derzeit noch an. Wir werden darüber berichten, sobald diese vollständig aufgearbeitet sein werden.

Was ist der aktuelle Status?

Die HPC-Infrastruktur ist wieder in Betrieb. Das Firmware-Update ist durchgeführt. Es gab Datenverluste. Diese konnten minimiert werden, was allerdings für die betroffenen Nutzerinnen und Nutzer selbstverständlich kein Trost ist. Grundsätzlich ist jede Art von Datenverlust inakzeptabel. In Kürze wird das neue System zur Spiegelung der Daten in Betrieb genommen. Derzeit werden die Daten im HPC-Bereich ausschließlich durch das RAID-System abgesichert. Weitere kritische Arbeiten am Speichersystem werden so lange nicht stattfinden, bis das neue System zum Spiegeln der Daten im Einsatz ist.

Was sind die unmittelbaren Konsequenzen, damit sich so etwas nicht wiederholt?

- Wir haben eine neue, sehr leistungsfähige Synchronisations-Software zum Spiegeln der Daten beschafft. Diese wird umgehend in Betrieb genommen. Damit wird es wieder ein vollständiges Redundanzsystem für das GPFS geben.
- Hardwareseitig wird das Backup-System für den Storage sowie ein Teil der Compute-Nodes in einem separaten Serverraum (in der ULB) untergebracht werden, damit diese Komponenten räumlich vom HHU-Serverraum, in dem das HPC-Hauptsystem untergebracht ist, getrennt sind (Schutz bei Brand, Wassereintrich, etc.). Die in dem separaten Serverraum untergebrachten Compute-Nodes werden auch bei Wartungsarbeiten am Hauptsystem in Betrieb bleiben, um durchgängigen Zugriff auf die Daten und in besonders dringenden Fällen auch ein Arbeiten auf dem HPC-System zu ermöglichen.
- Wir werden die Systeme so absichern, dass sie gegen Schäden aufgrund von Stromausfall, Überspannung und Wartungsarbeiten an der Stromversorgung besser geschützt sind.
- Wir werden unseren Kommunikationsprozess weiter verbessern. Insbesondere werden wir alle Nutzerinnen und Nutzer sofort informieren, falls es Probleme mit dem RAID-System oder beim Spiegeln der Daten geben sollte, damit besonders wertvolle Daten zusätzlich außerhalb des HPC-Storage-Systems gesichert werden können.
- Wir haben auch eine Bitte an unsere Nutzerinnen und Nutzer: Wenden Sie sich bei Problemen und Nachfragen nicht direkt, sondern stets über das Ticket-System an uns, da dort mehrere Mitarbeiterinnen und Mitarbeiter antworten können. An uns persönlich gerichtete E-Mails können insbesondere in Zeiten hohen Arbeitsanfalls oder auch bei Urlaub oder Krankheit längere Zeit unbeantwortet bleiben.

Was sind die mittelfristigen Konsequenzen, damit sich so etwas nicht wiederholt?

Mittelfristig braucht die HHU einen weiteren großen Serverraum, in dem dann auch redundante HPC-Systeme untergebracht werden können. Die Planungen dafür laufen. Wenn alles gut geht, wird dieser Serverraum 2025 zur Verfügung stehen. Zusätzlich arbeiten wir weiter an der Optimierung des bestehenden HHU-Serverraums in Gebäude 25.41. So wird beispielsweise die Stromversorgung so umgestellt werden, dass künftig weniger Beeinträchtigungen aufgrund von Stromausfällen und Wartungsarbeiten an der Stromversorgung auftreten können.

Die Details für alle, die es genauer wissen wollen

Als im Frühjahr 2018 das parallele Hochleistungs-Storagesystem des Herstellers Data-Direct-Networks (kurz DDN) auf Basis von IBMs GPFS (heute „Spectrum Scale“) geplant und aufgebaut wurde, wurde parallel dazu auch ein Backup-System aufgebaut. Dieses System soll einspringen, falls es mit dem primären System Probleme gibt und somit den Betrieb des Clusters möglichst schnell wieder oder gar unterbrechungsfrei gewährleisten. Kernkomponente für eine solche Architektur ist eine Synchronisation der Daten zwischen beiden Systemen, die auch Teil der Abnahmekriterien für das System war. Zum damaligen Zeitpunkt funktionierte dieser Mechanismus (nach einer von uns eingeforderten Nachbesserung) einwandfrei.

Im Laufe der letzten zwei Jahre stieg jedoch die Datenmenge von knapp 500 TB bei Inbetriebnahme des Storage-Systems stetig auf nun mehr als das Fünffache an. Dies führte dazu, dass der Backup-Mechanismus immer instabiler wurde. Seit Herbst 2019 bereitete das HPC-Team ein großes Update vor, um die ursprüngliche Stabilität wieder zu gewährleisten. Bei einem ersten Versuch Ende Oktober traten dabei erhebliche Komplikationen mit der Hardware auf, sodass dieser Prozess nach knapp einer Woche Laufzeit abgebrochen werden musste. Da solch tief greifende Arbeiten nur bei komplett leerem Cluster durchgeführt werden können, sollte ein neuer Versuch erst wieder ab dem 12.3.2020 unternommen werden. Denn um die Ausfallzeiten des HPC Systems zu minimieren, nutzen wir immer externe Vorgaben für solche Termine. Eine solche Gelegenheit bot sich bei den jährlichen Redundanztests der Stromversorgung durch Dezernat 6 der HHU. Da es in den letzten Jahren erfahrungsgemäß immer wieder zu Problemen durch diese Tests kam, wurde dafür vorsorglich ohnehin der Compute-Teil des Clusters heruntergefahren. So konnte zumindest die restliche zentrale IT der HHU (inkl. Storage) - abgesichert durch die Notstromversorgung - weiterlaufen. Im Anschluss an die Tests sollten dann die Updates gemacht werden, damit der Cluster nicht mehrfach heruntergefahren werden muss. Für diese Arbeiten waren von DDN, dem Hersteller des Systems, 17 Stunden eingeplant gewesen. Das System wäre also eigentlich bereits nach einem Tag wieder online gewesen.

Die folgenden Erklärungen sind eine Zusammenfassung und kein detaillierter Ablaufplan. Einige der Schritte und Prozesse wurden mehrfach wiederholt oder dauerten teilweise Tage.

Nach dem Stromtest war lediglich ein einziges der Netzteile in einem Fehlerzustand und musste neu gestartet werden. Dies ist für ein großes System wie unser HPC System nichts Ungewöhnliches und mehr Anzeichen für Hardware-Schäden gab es auch erstmal nicht. Der Hersteller (DDN) teilte uns demzufolge mit, dass die Updates wie geplant stattfinden könnten. Nach 6 Stunden waren bereits viele Arbeitsschritte vom Update-Plan geschafft, sodass es so aussah, als würden wir sogar schon deutlich früher fertig werden.

Nach dem Update der Enclosures, also den Festplatten-Systemen, kam es zu einer Reihe von unterschiedlichsten Fehlermeldungen. Diese deuteten an, dass es bei den Verbindungen zwischen Enclosures und Storage-Servern Probleme gab. Diese Meldungen gab es vor dem Update nicht, sonst hätten wir die Arbeiten erst gar nicht begonnen. Zusätzlich wurden plötzlich über 35 Festplatten des Systems gleichzeitig als fehlerhaft markiert. Dies war zwar schon besorgniserregend, aber noch nicht kritisch, da immer 10 Festplatten zusammen genutzt werden (ein sog. RAID-Pool), um sowohl die Daten parallel zu speichern, als auch durch eine Parität und doppelte Parität die Datenintegrität auch bei Festplattenausfällen aufrechtzuerhalten. Letztendlich können dadurch bis zu zwei der 10 Festplatten gleichzeitig ausfallen, ohne dass es zu Datenverlust kommt. Sofern die defekten Festplatten also einigermaßen gleichmäßig auf unser GPFS-Speichersystem verteilt sind, wären dies durchschnittlich weniger als eine Festplatte pro Pool gewesen.

Auffällig war hier jedoch, dass die defekten Platten mehrheitlich einen einzigen Pool (den berüchtigten Pool-49) betrafen und so wurde dieser als "Non-Redundant" markiert. Das System begann automatisch eine der Spare-Platten (welche im System immer bereitstehen) zu nutzen, um die fehlende Parität dieses Pools und damit die Sicherheit gegen den Ausfall von weiteren Festplatten, wiederherzustellen. Dieser Prozess schlug jedoch nach knapp 18 Stunden bei 96% fehl.

4% von 80 TB entspricht immer noch über 3 TB an Daten, welche von dem Problem betroffen sein können!

An diesem Punkt wurde uns und dem Hersteller klar, dass es noch weitere unentdeckte Fehler im System gab. Daher sprang nun ein Team aus Spezialisten von DDN rund um den Globus ein, um zusammen mit dem HPC-Team zu planen, welche Schritte die Daten retten könnten.

Wir kamen nach weiteren Analysen zu dem Schluss, dass es also bei einer ganzen Menge von hochwertigen elektronischen Komponenten zu physikalischen Defekten gekommen war. Also entschieden wir, dass direkt mehrere "IOM-Module" (I/O-Multiplexer), die die Verbindungen zwischen Enclosures und Storage-Servern herstellen, ausgetauscht werden müssten und zudem einige Kabel vorsichtshalber ebenfalls. Diese Teile wurden per Express-Lieferung aus den Niederlanden innerhalb von wenigen Stunden zum ZIM geliefert.

Da sowohl Herr Raub als auch Herr Rehs unter Quarantäne wegen des Corona-Virus standen, musste der neue Kollege Herr Siebert, welcher eigentlich für Anwendungs-Support zuständig ist und das Storage-System und den physikalischen Aufbau gar nicht kannte, die Komponenten des Systems austauschen. Mittels Remote-Anleitung per Videokonferenz konnte er diese Austausche dennoch erfolgreich durchführen und die erkannten Fehler wurden so behoben. Alle schöpften Hoffnung, dass das System am Montag, den 30.3.2020 wieder online gehen würde.

Da trotz Reparatur und Wiederherstellung des Pools 49 ein paar Dateien unwiederbringlich beschädigt waren, wurden die betroffenen Nutzer zu diesem Zeitpunkt darüber informiert.

Als das System am 30.3.2020 wieder online ging, häuften sich jedoch schon nach wenigen Stunden die Meldungen der Nutzer, dass es weiterhin zu I/O-Fehlern kam, welche bedeuteten, dass immer noch nicht-korrigierbare Defekte vorliegen mussten. Dies wurde sofort an den Hersteller gemeldet und es wurden erneut mit den Spezialisten die Ursachen gesucht und letztendlich auch gefunden. Damit solche Probleme zukünftig nicht erneut unentdeckt bleiben, arbeitet der Hersteller DDN an einer Verbesserung seiner eigenen Überwachung der Hardware.

Erneut wurden Teile kurzfristig aus den Niederlanden an die HHU geschickt und dieses Mal durch Herrn Rehs donnerstags Abends und durch Herrn Raub am Karfreitag verbaut. Danach war klar, dass noch weitere Dateien betroffen waren und die noch heilen Dateien schnell aus Pool 49 gerettet werden müssen, bevor dieser komplett zusammenbrechen würde. Bei einer Kapazität von über 60 TB in einem Pool musste dies auf jeden Fall verhindert werden. Daher wurde beschlossen, die Daten auf Filesystem-Ebene aus dem betroffenen Pool zu migrieren und auf die anderen Pools zu verteilen. Dies klappte zwar für die ersten 10% der Daten gut, jedoch traten auch dort dann IO-Fehler auf, was zu einem Abbruch des Prozesses führt.

Nun mussten jedes Mal die fehlerhaften Blöcke auf den Festplatten des Pool 49 identifiziert werden und den Dateien im Filesystem zugeordnet werden, bevor ein erneuter Versuch gestattet werden konnte. Diese Suche nach den betroffenen Dateien dauerte immer zwischen 6 und 8 Stunden, da die Metadaten von über 100 Millionen Dateien im Filesystem überprüft werden mussten. Danach konnten die betroffenen Blöcke gelöscht werden und der Prozess für Migration von Neuem gestartet werden, in der Hoffnung, dass er dieses Mal durchlief. So ging dieser Prozess über Tage hinweg, um immer weiter Dateien zu retten, die noch zu retten waren. Glücklicherweise wurde bei jeder Iteration die Zahl der defekten Blöcke kleiner und es gab erneut etwas Hoffnung, dass man innerhalb der nächsten Tage fertig werden würde.

Leider war auch dies wieder nicht der Fall. Die Geschwindigkeit des Prozesses reduzierte sich gegen Ende schlagartig und geschätzt wäre dann noch ein Jahr benötigt worden. Dies war natürlich nicht akzeptabel, also musste nach einem alternativen Weg gesucht werden. Da das Filesystem auf Enclosure-Ebene und -Verbindung zu den Storage-Controllern durch die Austausche immerhin stabil zu laufen schien und die Anzahl der defekten Blöcke immer kleiner wurde, entschied sich das HPC-Team zusammen mit dem Hersteller, die Daten nicht zu migrieren, sondern den Festplattenpool aus dem Filesystem zu nehmen und damit eine Migration der Daten - auch mit Lesefehlern - zu erzwingen. Am Ende des Prozesses hätten wir erneut eine Liste mit Dateien, die von diesen Lesefehlern betroffen waren, gehabt und man hätte die Nutzer informieren können.

Leider brach auch dieser Prozess aufgrund von zu vielen Lesefehlern - entgegen den Erwartungen - mehrfach ab. Erst beim dritten Versuch wurden tatsächlich endgültig alle Daten aus dem Pool migriert und auf andere Festplatten verteilt. Nun konnten die Listen mit allen defekten Dateien gesammelt und die Nutzer informiert werden; dies geschah am 18.04.2020. Insgesamt wurden bei dem Ausfall knapp 1000 von 100 Millionen Dateien beschädigt. Dies sind zwar leider immer noch viel zu viele, aber immer noch viel besser als 80 TB von 2,5 PB zu verlieren!

Seit dem 20.04.2020 kann der Cluster wieder ohne Einschränkungen genutzt werden. Im Hintergrund wird immer noch daran gearbeitet, die nun verlorene Speicherkapazität des Pools 49 wieder dem System hinzuzufügen, aber dies beeinträchtigt lediglich die Performance und nicht die generelle Verfügbarkeit des Systems.