

GWDG NACHRICHTEN 11|16

Extension of the
HPC Capacity

HDC-Anwendungs-
konservierung

Backup großer
Datenbereiche mit TSM

33. DV-Treffen der
Max-Planck-Institute

ZEITSCHRIFT FÜR DIE KUNDEN DER GWDG





GWDG NACHRICHTEN

11|16 Inhalt

.....

4 GWDG extends High Performance Computing Capacity 7 Die HDC-Anwendungskonservierung – ein Dienst zur Archivierung und Bereitstellung komplexer Forschungsergebnisse

9 Kurz & knapp 10 Backup großer Datenbereiche mit TSM 16 33. DV-Treffen der Max-Planck-Institute in Göttingen 18 Personalien 19 Kurse

Impressum

.....

Zeitschrift für die Kunden der GWDG

ISSN 0940-4686
39. Jahrgang
Ausgabe 11/2016

Erscheinungsweise:
monatlich

www.gwdg.de/gwdg-nr

Auflage:
550

Fotos:

© Sashkin - Fotolia.com (1)
© momius - Fotolia.com (6)
© alain wacquier - Fotolia.com (10)
© pterwort - Fotolia.com (15)
© Edelweiss - Fotolia.com (23)
© MPLbpc-Medienservice (3, 18)
© GWDG (2, 5, 16, 17, 19)

Herausgeber:

Gesellschaft für wissenschaftliche
Datenverarbeitung mbH Göttingen
Am Faßberg 11
37077 Göttingen
Tel.: 0551 201-1510
Fax: 0551 201-2150

Redaktion:

Dr. Thomas Otto
E-Mail: thomas.otto@gwdg.de

Herstellung:

Maria Geraci
E-Mail: maria.geraci@gwdg.de

Druck:

Kreationszeit GmbH, Rosdorf



Prof. Dr. Ramin Yahyapour
ramin.yahyapour@gwdg.de
0551 201-1545

Liebe Kunden und Freunde der GWDG,

vom 4. bis 6. Oktober fand das diesjährige 33. DV-Treffen der Max-Planck-Institute in Göttingen statt. Es hat uns gefreut, die IT-Verantwortlichen wieder einmal bei uns zu Gast gehabt zu haben. Auch bei diesem Treffen zeigte sich wieder, dass IT-Versorgung zunehmend in Kooperation erfolgt und gerade aus der Zusammenarbeit ein hoher Nutzen für die Beteiligten resultiert. Nicht alle Probleme müssen mehrfach erfahren und individuell gelöst werden. Community-Arbeit kostet zwar Anstrengungen; diese lohnen sich jedoch auf lange Sicht.

In dieser Ausgabe der GWDG-Nachrichten finden Sie weiterhin einen Artikel zum vor Kurzem erfolgten Ausbau unserer HPC-Ressourcen. In regelmäßigen Abständen erneuern und erweitern wir unser Angebot hierfür. Auch hier darf ich alle Nutzer ermuntern, sich an der jeweils vorgeschalteten Bedarfsplanung zu beteiligen. Nur so kann sichergestellt werden, dass die Systeme auch für ihre Anforderungen und Bedürfnisse geeignet sind.

Ich wünsche Ihnen viel Spaß beim Lesen dieser Ausgabe.

Ramin Yahyapour

GWDG – IT in der Wissenschaft

GWDG extends High Performance Computing Capacity

Text und Kontakt:

Dr. Christian Boehme
christian.boehme@gwdg.de
0551 201-1839

The GWDG Scientific Compute Cluster, our high performance computing (HPC) system, has recently been extended by 121 nodes and one frontend gwdu103. This increases its peak performance by 104 TFlop/s (104 x 10¹² floating point operations per second) to 356 TFlop/s. Furthermore, the extension also adds nodes with up to 1.5 TB main memory for applications with extreme memory requirements and graphics processing units (GPUs) as compute accelerators to the system.

OBJECTIVES AND TECHNICAL DETAILS

Providing state-of-the-art HPC resources to researchers is an important service in the GWDG portfolio. In order to accommodate the steadily increasing hardware requirements of scientific applications, it is necessary to regularly upgrade and extend these resources. Following the typical HPC technology lifecycle and recommendations of the DFG ("Deutsche Forschungsgemeinschaft"), GWDG provides a major upgrade to its HPC systems every five years. The last of these upgrades happened in 2013. In-between major upgrades GWDG acquires smaller extensions, both to provide more specialized capabilities to HPC users, and to use technological advances to achieve a better performance vs cost ratio than achievable with a strictly five-year upgrade period, especially considering that even a few years often yield significant improvements of the energy efficiency of HPC systems.

The tender for the 2016 extension was won by Hamburg based company Delta Computing. The system offered and installed by Delta has these key components:

- 76 base nodes with two Intel E5-2650 v4 ("Broadwell") CPUs each. The E5-2650 v4 has 12 processor cores at 2.2 GHz clock speed. Each base node is equipped with 128 GB main memory. These nodes mainly extend the capacity of our general purpose "mpi" queue. They also replace the AMD "Woodcrest" based nodes acquired in 2011, which also were equipped with 128 GB main memory. These nodes are co-financed by the Institute of Astrophysics of the University of Göttingen.
- 15 nodes with 512 GB main memory, otherwise identical to the base nodes, which extend the capacity of the "fat" queue for memory intensive jobs. As they have twice as much memory as the existing "fat" nodes they are also used in the newly established "fat+" queue for extreme memory requirements.

- 10 nodes with two NVIDIA Tesla K40 GPUs each. The Tesla K40 has a comparatively large memory of 12 GB with ECC error correction, and a balanced single and double precision performance. It is therefore suited for a broad range of GPU accelerated applications.
- 15 nodes with two NVIDIA GTX 980 GPUs each (see Figure 1). The GTX 980 is a consumer GPU and offers a very good price / performance ratio for single precision GPU accelerated applications with medium memory requirements. An example for such applications is GRO-MACS, a molecular dynamics software seeing much use in Göttingen. The GTX nodes are co-financed by the Max Planck Institute for Biophysical Chemistry.
- 5 nodes with 1.5 TB main memory each meant for applications with extreme memory requirements, for example in the area of bioinformatics. These nodes are equipped with four Intel E5-4620 v3 ("Haswell") CPUs each, as Intel could not offer a "Broadwell" option for this kind of system in time. The E5-4620 v3 has 10 processor cores at 2 GHz clock speed.

Die GWDG erhöht die Leistung ihrer Hochleistungsrechner

Der „Scientific Compute Cluster“, die Hochleistungsrechner der GWDG, wurde jetzt um 121 Rechenknoten und ein Frontend gwdu103 erweitert. Dadurch erhöht sich die maximale Gesamtleistung des Systems um 104 TFlop/s (104 x 10¹² floating point operations per second) auf nun 356 TFlop/s. Außerdem wird der Cluster durch Knoten mit bis zu 1,5 TB Hauptspeicher für Anwendungen mit extremen Speicheranforderungen und durch GPUs (Graphics Processing Units) zur Anwendungsbeschleunigung ergänzt.

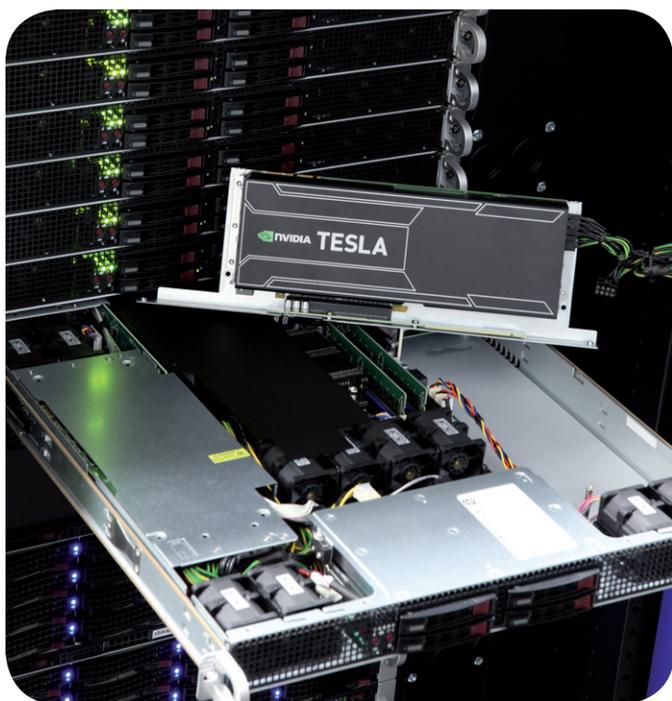


Figure 1: View of an NVIDIA Tesla equipped node with the GPU removed

- A parallel scratch file system with 108 TB capacity.

USING THE CLUSTER EXTENSION

The following paragraphs explain how to use the new nodes and are meant for users already familiar with the GWDG Scientific Compute Cluster. If you need a general introduction to using the cluster, please refer to our documentation [1] or consider participating in our course “Using the GWDG Scientific Compute Cluster – an Introduction” [2].

Scratch usage

The new scratch file system is mounted on the new nodes as `/scratch2`. The new nodes do not have access to the old `/scratch` file system and vice versa. For your convenience `/scratch2` is linked to `/scratch` on the new nodes, but the new nodes cannot access any files stored in the existing `/scratch`.

Requesting nodes with access to the existing `/scratch` works as before:

```
#BSUB -R scratch
```

To request the new `/scratch2` use:

```
#BSUB -R scratch2
```

CPU architecture selection

With the upgrade our cluster provides four generations of Intel CPUs and two generations of AMD CPUs. However, the main difference between these CPU types is whether they support Intel's AVX2 or not. For differentiating these types, we have introduced the `x64inlvl` (for x64 instruction level) label:

- `x64inlvl=1`: Supports only AVX
- `x64inlvl=2`: Supports AVX and AVX2

In order to choose an AVX2 capable node you therefore have to include

```
#BSUB -R "x64inlvl=2"
```

in your submission script.

Selecting nodes by their CPU slot count

The `np xx` resource requirement used for selecting nodes by CPU slot count is deprecated and should be substituted by `ncpus= xx` . For example, in order to choose a node with 20 CPU slots, your submission script should contain

```
#BSUB -R "ncpus=20"
```

The old syntax (`-R np20`) will continue to work, and will also be supported on the new nodes. However, it will be removed with the next cluster extension.

Selecting nodes by the size of their main memory

You can select a node by its maximum available memory (`maxmem`) in MB. To select a node with more than about 500 GB memory use:

```
#BSUB -R "maxmem>500000"
```

Note that this only **selects** nodes with more than 500 GB memory. If you also want to **reserve** this memory, the easiest way is to combine `-R "maxmem>..."` with `-x` for an exclusive job. See our documentation for further information.

To select a node by memory per core you can use basic math in the resource requirement:

```
#BSUB -R "maxmem/ncpus>6000"
```

selects nodes with more than about 6 GB per core. You can then use `-n` to reserve $n \times 6$ GB of memory for your job.

Using the fat+ queue

The `fat+` queue is our new queue for jobs with extreme memory requirements. It provides nodes with 512 GB, 1.5 TB, and 2 TB main memory. The `fat+` queue should not be used for jobs which could also run in other queues. Therefore, please note the following policies for the `fat+` queue:

- Your job must use more than 250 GB RAM.
- Your job must use at least a full 512 GB node or half a 1.5 TB or 2 TB node.

In order to request a full 512 GB node use:

```
#BSUB -x
#BSUB -R "maxmem < 600000"
```

The 512 GB nodes are also available in the `fat` queue, without these restrictions. However, `fat` jobs on these nodes have a lower priority compared to `fat+` jobs.

In order to request half a 1.5 TB node, when your job needs more than 500 GB memory, use:

```
#BSUB -n 20
#BSUB -R span[hosts=1]
#BSUB -R "maxmem < 1600000 && maxmem > 600000"
```

Requesting full 1.5 TB or 2 TB nodes works similar to selecting a full 512 GB node.

GPU selection

In order to use a GPU, you need to submit your job to the *gpu* queue, and request GPU shares. Each node equipped with a GPU provides as many GPU shares as it has cores, independent of how many GPUs are built in. So, on the new nodes, which have 24 cores, the following would give you exclusive access to GPUs:

```
#BSUB -R "rusage[ngpus_shared=24]"
```

Note that you need not necessarily also request 24 cores with *-n 24*, as jobs from the MPI queue may utilize free CPU cores if you do not need them. The new nodes have two GPUs each, and you should use both, if possible. If you request less shares than available, other jobs may also utilize the GPUs. However, selecting a specific GPU to use needs to be handled in the application and requires some experience, so we generally do not recommend to share GPUs.

Currently we have two generations of NVIDIA GPUs in the cluster, selectable with the *nvgen* resource requirement:

- *nvgen=1*: Kepler
- *nvgen=2*: Maxwell

If you need their double precision performance, or error correcting memory (ECC RAM), you can select the Tesla GPUs with

```
#BSUB -R tesla
```

Our Tesla K40 are of the Kepler generation (*nvgen=1*). If you want to make sure to run on a node equipped with two GPUs use:

```
#BSUB -R "ngpus=2"
```

As always, if you have any issues with the new nodes please write to support@gwdg.de.

LINKS

- [1] https://info.gwdg.de/docs/doku.php?id=en:services:application_services:high_performance_computing:start
- [2] <https://www.gwdg.de/de/general-services/courses>



FTP-Server

Eine ergiebige Fundgrube!

Ihre Anforderung

Sie möchten auf das weltweite OpenSource-Softwareangebot zentral und schnell zugreifen. Sie benötigen Handbücher oder Programmbeschreibungen oder Listings aus Computerzeitschriften. Sie wollen Updates Ihrer Linux- oder FreeBSD-Installation schnell durchführen.

Unser Angebot

Die GWDDG betreibt seit 1992 einen der weltweit bekanntesten FTP-Server, seit sieben Jahren mit leistungsfähigen Ressourcen für schnellen Service

Ihre Vorteile

- > Großer Datenbestand (40 TByte), weltweit verfügbar
- > Besonders gute Anbindung im GÖNET

- > Aktuelle Software inkl. Updates der gebräuchlichsten Linux-Distributionen
- > Unter pub befindet sich eine aktuell gehaltene locatedb für schnelles Durchsuchen des Bestandes.
- > Alle gängigen Protokolle (http, ftp, rsync und nfs) werden unterstützt.

Interessiert?

Wenn Sie unseren FTP-Server nutzen möchten, werfen Sie bitte einen Blick auf die u. g. Webseite. Jeder Nutzer kann den FTP-Dienst nutzen. Die Nutzer im GÖNET erreichen in der Regel durch die lokale Anbindung besseren Durchsatz als externe Nutzer.

>> www.gwdg.de/ftp-server



Die HDC-Anwendungs-konservierung – ein Dienst zur Archivierung und Bereitstellung komplexer Forschungsergebnisse

Text und Kontakt:

Dr. Sven Bingert
sven.bingert@gwdg.de
0551 201-2164

Stefan Buddenbohm
buddenbohm@sub.uni-goettingen.de
0551 39-20555

Forschungsdaten bestehen zunehmend aus komplexeren Repräsentationen, die ganz verschiedene Inhalte, Formate und Strukturen annehmen und miteinander verwoben sind. Die Archivierung und Referenzierung derartig komplexer Forschungsdaten erfordert auf Seiten der Infrastrukturanbieter neue Lösungen und Angebote, die über eine konventionelle Archivierung in Repositorien hinausgeht. Als ein Beispiel für neue Angebote im Kontext des Forschungsdatenmanagements wird im Folgenden die Anwendungskonservierung des Humanities Data Centre, eines geisteswissenschaftlichen Forschungsdatenzentrums, vorgestellt.

NEUE FORMEN KOMPLEXER FORSCHUNGSDATEN

Forschung wird in vielen verschiedenen Disziplinen mit unterschiedlichen Methoden betrieben. Genauso vielfältig sind Formate und Strukturen digitaler Forschungsergebnisse und -daten. Die Nachnutzung von Ergebnissen und Daten ist dabei eine wesentliche Voraussetzung für zukünftige Forschung, aber auch um Nachvollziehbarkeit und Transparenz der Forschung zu gewährleisten. Dies erfordert von Seiten der Wissenschaft (idealerweise) die Dokumentation ihrer Arbeit und die Aufbereitung der Forschungsergebnisse und -daten in einer entsprechenden Qualität. Für Bereitstellung und Nachnutzung sind dafür wiederum entsprechende Infrastrukturen und Dienste bereitzustellen. Was mit Blick auf die konventionellen Forschungsergebnisse, sprich textbasierte Publikationen, bereits etablierte Praxis ist, stellt sich hinsichtlich der Forschungsdaten wesentlich komplexer und komplizierter dar. Für Zeitschriftenartikel, Monografien und Sammelbände sind Publikationsrepositorien als etablierte Archivierungsorte mit einer darauf aufsetzenden Struktur an Mehrwertdiensten wie bspw. Meta-suchen oder Registries vorhanden, die durch etablierte Metadaten- und Schnittstellenstandards ermöglicht werden. Anders sieht die Situation für insbesondere geisteswissenschaftliche Forschungsdaten aus. Hier stehen Standardisierung sowie die selbstverständliche Handhabung der Archivierungspraxis und Nachnutzung noch am Anfang [1].

DAS HUMANITIES DATA CENTRE

Die erfolgreichen und etablierten Forschungsaktivitäten der Georg-August-Universität Göttingen im Bereich der Geisteswissenschaften legten es nahe, den Blick auch auf begleitende und neue Forschung ermöglichende Infrastruktur zu richten. Ein Beispiel dafür ist das Humanities Data Centre (HDC), das als Forschungsdatenzentrum für die Geisteswissenschaften zum

The HDC Application Preservation – a Service for Archiving and Providing complex Research Results

Research data today becomes more and more complex, moving away from file-based structures and content to more and more complex representations of data. These complex representations of research data usually convey a more complete picture of a research project or result different than when only looking at the various raw data on a file-based level. Therefore the demand for archiving and referencing solutions for this kind of complex research data is necessary and has to be addressed by the research infrastructure provider. The article describes the HDC application preservation as an example for this correlation.

1. August 2016 seinen Betrieb aufgenommen hat. Es wird von der GWDG und der Niedersächsischen Staats- und Universitätsbibliothek Göttingen (SUB) getragen. Vorausgegangen war eine zweijährige Designphase, deren Ziel es unter anderem war, ein initiales Serviceportfolio zu entwickeln. Hierfür wurden auf der einen Seite die Voraussetzungen und Bedürfnisse der Wissenschaft erhoben, auf der anderen Seite Erfahrungen und vorhandene Dienste auf Seiten der Infrastrukturanbieter, in diesem Fall in erster Linie die GWDG und die SUB, ausgewertet. Über diese dienstorientierten Anbieter hinaus waren auch wissenschaftliche Partner Teil des Projektkonsortiums. Das initiale Serviceportfolio als Ergebnis wird Gegenstand konstanter Weiterentwicklung und Anpassung sein. Als ein Beispiel für die innovativen Aspekte dieses Angebotes wird im Folgenden die HDC-Anwendungskonservierung vorgestellt.

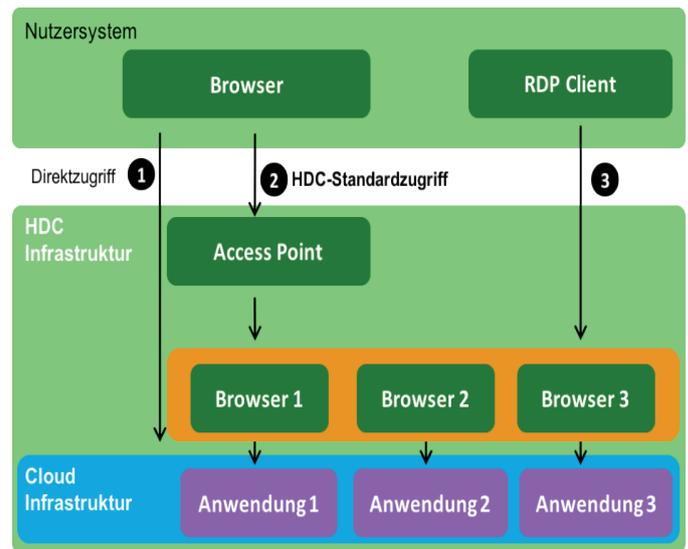
ANWENDUNGSKONSERVIERUNG FÜR KOMPLEXE FORSCHUNGSDATEN

Ein Beispiel für die oben angesprochenen komplexen Forschungsdaten sind komplexe Softwareumgebungen (Complex Software Environments, CSEs). Diese bestehen nicht nur aus einem Datenformat, sondern aus verschiedenen Softwareschichten, die der Verarbeitung, Verknüpfung und Darstellung von Daten dienen [2]. Eine Digitale Edition in den Geisteswissenschaften kann bspw. aus einer Datenbank und einer grafischen Oberfläche als Schnittstelle zum menschlichen Nutzer bestehen. Um eine Archivierung und Präsentation solcher Anwendungen zu ermöglichen, wurde im Rahmen der HDC-Designphase die „Angebotskomponente Anwendungskonservierung“ entwickelt.

Ziel der Anwendungskonservierung ist es, CSEs zu archivieren und über einen möglichst langen Zeitraum der Wissenschaft zugänglich zu machen. Verschiedene technische und organisatorische Aspekte sind dabei zu berücksichtigen. Zum einen muss die Sicherheit der Anwendung und der dazugehörigen Daten gewährleistet sein. Aber auch der Betrieb über mehrere Generationen der zugrundeliegenden Infrastruktur muss bedacht werden. Im Folgenden soll nicht auf alle diese Bereiche eingegangen werden, sondern der Fokus auf die technische Bereitstellung und Architektur gelegt werden.

Das Grundprinzip der Anwendungskonservierung ist die Überführung einer CSE, bspw. einer Digitalen Edition, zum Ende eines Forschungsprojektes in das Forschungsdatenzentrum. Mittels einer Cloud-Infrastruktur wird die CSE im Forschungsdatenzentrum gespeichert, betrieben und bereitgestellt. Idealerweise gestaltet sich das Nutzungserlebnis mit der CSE genauso wie vorher im Rahmen der Bereitstellung durch das Forschungsprojekt. Das heißt, der Nutzer hat einen möglichst einfachen Zugang zur CSE, idealerweise über einen herkömmlichen Browser, und erhält durch andere Angebote des Forschungsdatenzentrums Zugang zu Dokumentation, Rohdaten oder Ähnlichem.

Damit ein sicherer Betrieb ohne aufwendige Pflege der Anwendung seitens des Datenzentrums gewährleistet werden kann, muss der Zugriff auf die Anwendung in einem kontrollierten Rahmen geschehen. Dies wird durch die Verwendung eines Browsers innerhalb der HDC-Infrastruktur erreicht, der nur über einen Zugangspunkt erreichbar ist (Zugang 2 in der Abbildung 1). Die im HDC-Browser angezeigten Inhalte werden dann im Browser des Nutzers ohne Verluste dargestellt. Der Zugangspunkt für den Nutzer dient dabei als HTML-Schnittstelle für die eigentliche



1_Architektur der HDC-Anwendungskonservierung

Remote Desktop Protocol-Verbindung (RDP) zum HDC-Browser. Dieses Vorgehen hat den Vorteil, dass auf Seiten des Nutzers keine weiteren Programme installiert oder konfiguriert werden müssen (Zugang 3).

Die im Datenzentrum konservierten Anwendungen besitzen besondere Regeln für die Firewall, die einen Direktzugriff (Zugang 1) und damit die mögliche Gefahr einer Beschädigung oder Missbrauch der Anwendung verhindern.

Ein weiterer Vorteil dieses Verfahrens ist die Darstellung der Inhalte über einen archivierten Browser. Oftmals werden Webanwendungen für bestimmte Browserhersteller und -versionen optimiert. Durch Weiterentwicklungen können dann die Darstellung und damit auch der Nutzen der Forschungsergebnisse beeinträchtigt werden. In der Anwendungskonservierung wird zu jeder Anwendung ein optimierter Browser ausgewählt, verknüpft und innerhalb der Infrastruktur betrieben. Dadurch ermöglicht die Anwendungskonservierung eine längerfristige Darstellung und Referenzierung der Anwendung [3]. Dies ist eine wesentliche Anforderung der Wissenschaft, wie sich in der HDC-Angebots-erhebung herausgestellt hat. Darunter ist zu verstehen, dass die in einem Forschungsvorhaben entwickelten Anwendungen – das können Webseiten ebenso wie komplexere Webanwendungen oder Datenbanken sein – nach Beendigung des Vorhabens üblicherweise nur für einen begrenzten Zeitraum weiterbetrieben werden können. Auf der anderen Seite referenzieren die beteiligten Wissenschaftler aber derartige Anwendungen sehr intensiv; die Frage nach einem entsprechenden Angebot von Seiten der Forschungsdatenzentren liegt also nahe.

Die Anwendungskonservierung nutzt gängige Technologien und Standards, die aufgrund ihrer Verbreitung gute Voraussetzungen für Interoperabilität und Nutzbarkeit seitens der Wissenschaft besitzen. Dadurch sind eine Weiterentwicklung und der sichere Betrieb gewährleistet sowieso eine Migration zu einem anderen Produkt und eine Nachhaltigkeit der Anwendungskonservierung im Forschungsdatenzentrum möglich.

FAZIT

Die hier vorgestellte Anwendungskonservierung ist eine mögliche Dienstkomponente eines (geisteswissenschaftlichen)

Forschungsdatenzentrums. Die Anwendungskonservierung ist von ihrer Zielsetzung, der Aufbewahrung und Referenzierung vollständiger Anwendungen wie Digitale Editionen oder Datenvisualisierungen, für ein Forschungsdatenzentrum insofern innovativ, als dass sie die sich verändernden Forschungsdaten reflektiert und auf Anforderungen aus der Wissenschaft eingeht, für die es bisher kein derartiges Angebot gibt: die Präsentation und der granulare Nachweis komplexer Forschungsdaten. Die Anwendungskonservierung bettet sich dabei in das Serviceportfolio des HDC ein. Die dort dargestellten Dienste verdeutlichen zwei Punkte: zum einen die Berücksichtigung der vielfältigen neuen Formen komplexer Forschungsdaten, zum anderen die Erkenntnis, dass ein Forschungsdatenzentrum daher über konventionelle Repositorien zur Aufbewahrung dateibasierter Forschungsdaten hinausgehen muss. Das HDC versucht, dem nicht nur mit der Anwendungskonservierung zu entsprechen, sondern auch mit einer starken Gewichtung des Beratungsanteils sowie bei der möglichen Integration von Mehrwertdiensten zur Nutzung der archivierten Daten. Für Letzteres ist die Zitation von konservierten Anwendungen mittels Persistent Identifier (PID) ein Beispiel.

Weitere Informationen zur Anwendungskonservierung und eine Implementierung finden Sie auf den Webseiten des Humanities Data Centre <http://www.humanities-data-centre.de>. Anfragen

können Sie gerne per E-Mail an kontakt@humanities-data-centre.de richten.

REFERENZEN

- [1] Andreas Aschenbrenner, Stefan Buddenbohm, Claudia Engelhardt, Ulrike Wuttke (2015): Humanities Data Centre – Angebote und Abläufe für ein geisteswissenschaftliches Forschungsdatenzentrum. HDC-Projektbericht: <http://resolver.sub.uni-goettingen.de/purl?gs-1/13740>
- [2] Stefan Buddenbohm, Claudia Engelhardt, Ulrike Wuttke (2016): Angebotsgenese für ein geisteswissenschaftliches Forschungsdatenzentrum. In: Zeitschrift für digitale Geisteswissenschaften (DOI: 10.17175/2016_003)
- [3] Sven Bingert, Stefan Buddenbohm, Daniel Kurzawe (2016): Referencing of Complex Software Environments as Representations of Research Data. In: Proceedings of the 20th International Conference on Electronic Publishing (DOI:0.3233/978-1-61499-649-1-58) sowie Sven Bingert, Stefan Buddenbohm (2017, im Erscheinen): Research Data Centre Services for Complex Software Environments in the Humanities. In: Information Services and Use Journal. ●

Kurz & knapp

GWDG-Stand bei der Immatrikulationsfeier am 17. Oktober 2016

Wie schon seit vielen Jahren, war die GWDG auch in diesem Jahr wieder mit einem Informationsstand bei der offiziellen Immatrikulationsfeier der Universität Göttingen am 17.10.2016 im Zentralen Hörsaalgebäude (ZHG) vertreten, um sich den neuen Studierenden vorzustellen.

Zusammen mit zahlreichen anderen Einrichtungen der Universität, des Studentenwerks und der Stadt sowie Göttinger Kulturinstitutionen präsentierte sich die GWDG im „Forum Studium – Beratung, Betreuung, Kontakte rund ums Studium“ und gab damit den Studienanfängerinnen und -anfängern eine gute Gelegenheit zum ersten Kennenlernen der GWDG in ihrer Funktion als gemeinsames Rechen- und IT-Kompetenzzentrum für die Universität Göttingen und die Max-Planck-Gesellschaft.

Dieses Angebot wurde auch intensiv genutzt. Mitarbeiter der GWDG beantworteten viele Fragen zum umfangreichen Leistungsangebot der GWDG. Zu den üblichen Fragen zu Rechnern und Internetzugängen, Mailing sowie Kursen, dem WLAN am Göttingen Campus und Datensicherheit kamen dieses Jahr vermehrt Fragen zu Cloud Services und Mobile Computing, zum Lehrangebot und zur Möglichkeit von Praktika und Hilfskrafttätigkeiten bei der GWDG.

Otto

Kursprogramm 2017 erschienen

Das Kursprogramm der GWDG für das Jahr 2017 ist erschienen. Es bietet wieder ein umfangreiches Angebot an Kursen zur effizienten Nutzung von Hardware, Software und Netzen.

Nähere Informationen zum Kursangebot der GWDG sind unter <http://www.gwdg.de/kurse> zu finden. Wenn Sie Wünsche nach weiteren Kursen haben, die nicht im aktuellen Kursprogramm enthalten sind, können Sie gerne eine entsprechende E-Mail an support@gwdg.de senden.

Otto

Neue Busfahrzeiten am Faßberg

Im Liniennetz der Göttinger Verkehrsbetriebe wurden zum 17.10.2016 einige Änderungen vorgenommen, die auch die Buslinie 23 betreffen, die an der Haltestelle „Faßberg“ am Max-Planck-Campus verkehrt, auf dem sich auch die GWDG befindet. Der Takt der Linie 23 wurde verdoppelt, da immer mehr Fahrgäste zwischen Bahnhof und Uni-Nord unterwegs sind. Daher fährt die Linie 23 im Wintersemester während der Vorlesungszeit doppelt so oft wie bisher. Viermal stündlich verkehrt ein Bus der Linie 23 nun in jede Richtung. Zusammen mit den vier Abfahrtszeiten der Linien 21 und 22 ergeben sich somit werktags zu den Hauptverkehrszeiten zurzeit insgesamt acht Fahrmöglichkeiten pro Stunde vom Faßberg ins Zentrum.

Otto



Backup großer Datenbereiche mit TSM

Text und Kontakt:
Björn Nachtwey
bjoern.nachtwey@gwdg.de
0551 201-2181

Das permanente Wachstum von Daten stellt die Filesystem-Administratoren vor die große Herausforderung, das Backup der Daten innerhalb des vorgegebenen Zeitfensters zu bewältigen und so die zugesagte Absicherung gegen Manipulation und Datenverlust einzuhalten. Der nachfolgende Artikel beleuchtet anhand der Backuplösung „IBM Spectrum Protect (ISP)“ – ehemals „Tivoli Storage Manager (TSM)“ – verschiedene Möglichkeiten und Ansätze. Diese werden sowohl kurz erläutert als auch die Grenzen und Beschränkungen aufgezeigt. Der zweite Teil des Artikels entwickelt auf Basis der Grundidee „Parallelisierung mittels mehrerer Backup-Threads“ verschiedene Varianten, das Backup zu beschleunigen. Dieser Artikel basiert auf einem entsprechenden Vortrag auf dem 33. DV-Treffen der Max-Planck-Institute, das vom 4. bis 6. Oktober 2016 in Göttingen stattgefunden hat.

AUSGANGSLAGE

Die Datenmengen wachsen, durchschnittlich um 20 % pro Jahr [1,2]. Neben den Herausforderungen, diese Daten sinnvoll und performant zu speichern, fällt häufig ein Aspekt aus dem Fokus: Wie werden diese wachsenden Datenmengen gesichert?

Nominell wächst die Leistung der Tape-Systeme schneller als das Datenvolumen [1,2], aber auch diese Sicht ist leider unvollständig, da der Prozess des Sicherns, das eigentliche Backup, häufig den Flaschenhals darstellt.

„IBM Spectrum Protect (ISP)“ (bzw. zuvor schon als „WDSF/VM“, „DFDSM“, „ADSM“ und „TSM“) verfolgt bereits seit Langem den Ansatz, beim Backup nur die Änderungen seit dem letzten Backup zu sichern und dabei auf regelmäßige Vollsicherungen zu verzichten, also einen „Incremental forever“-Ansatz.

Der Vorteil liegt auf der Hand: Gerade bei großen Datenmengen (> 10 TB) ist die täglich geänderte Datenmenge relativ gering, so dass auch bei sehr vielen Versionen (bei der GWDG in der Regel bis zu 90) das Volumen der täglichen Sicherungen im Vergleich zur abgesicherten Datenmenge nach – IBM-Wortwahl „Front End

Capacity“ (FEC) – eher gering ist [3]. Damit ist auch die benötigte Kapazität für das Vorhalten der Versionen im Vergleich zur FEC nur geringfügig größer. Sichert man hingegen regelmäßig die Daten vollständig, so muss die Backupkapazität das Mehrfache der FEC umfassen. Bei gemischten Ansätzen, z. B. dem „Grandfather-Father-Son“-Prinzip mittels

- monatlicher Vollsicherungen (Grandfather),
- wöchentlicher Differenzsicherungen (Father) und
- täglicher Sicherungen (Son)

erhöht sich die notwendige Backupkapazität nochmal.

Aber das „Incremental forever“ löst ein wesentliches Problem einer jeden inkrementellen Sicherung nicht, nämlich die Antwort auf die Frage, welche Daten überhaupt gesichert werden müssen. ISP identifiziert diese zu sichernden Daten („Backupkandidaten“) dadurch, dass es alle Verzeichnisse und Dateien auf dem zu sichernden Rechner mit denen aus dem letzten Backup vergleicht und sich geänderte Dateien merkt. Dieser Prozess läuft in der Regel mit einer Geschwindigkeit von 1 – 2 Mio. Objekten pro Stunde.

Damit dauert allein das Durchsuchen eines 100-TB-Filesystems

mit rund 100 Mio. Objekten zwischen 50 und 100 Stunden; ein **tägliches Backup eines solchen Filesystems ist daher nicht möglich**.

Hinzu kommt das Problem, dass sich innerhalb dieser erheblichen Suchzeit wohl wieder nennenswert viele Daten neu ändern oder sogar gelöscht werden. Als Ergebnis wirft ISP reichlich Fehlermeldungen (*ANS4037E Object '<NAME>' changed during processing. Object skipped, ANS4005E Error processing '<NAME>': file not found*). Was kann man also tun?

LÖSUNGEN, DIE KEINE SIND

Eine Variante ist wohl die „Vogel-Strauß-Methode“: Man passt die Dienstbeschreibung der Fileserver an und sichert dem Nutzer keine täglichen Backups, sondern (zunächst) nur alle zwei Tage zu. Im Rahmen des Datenwachstums wird die Backuphäufigkeit fortlaufend angepasst. Bei spätestens 150 Mio. Objekten dürfte es nur noch eine monatliche Sicherung sein.

An diesem Punkt sollte über die zweite „Null-Lösung“ nachgedacht werden: Das Backup entsprechender Filesysteme komplett aufzugeben.

Da das Suchen der „Backupkandidaten“ also das Problem des Backups ist, könnte man also überlegen, in den sauren Apfel zu beißen und doch nur Vollsicherungen zu machen, im Vergleich zu DISK-Storage sind Tapes ja auch relativ günstig. Leider ist dies nach unseren Erfahrungen definitiv auch „keine Lösung“.

Für eine Vollsicherung von 100 TB benötigt man nämlich bei einer 10GE-Anbindung zwar theoretisch nur knapp 24 Stunden [4], praktisch wirken aber wahrscheinlich einige Hemmnisse [5], so dass man effektiv nur etwa 2 – 4 TB pro Tag sichert und für jede Vollsicherung rund 25 – 50 Tage benötigt – also in etwa die gleiche Zeit wie beim „incremental“.

HYBRIDER ANSATZ MIT SNAPSHOTS

Zahlreiche Filesysteme und auch die meisten Filer bieten die Möglichkeit, Snapshots anzulegen. In der Kombination von Snapshots und dem ISP-Backup lässt sich ein hybrider Ansatz umsetzen: Backups so häufig wie möglich, also beispielsweise wöchentlich, zwischendurch erfolgen Snapshots.

Meist gibt es neben der erheblichen Ausweitung des Backupzeitfensters noch den positiven Nebeneffekt, dass die Endnutzer direkt auf die Snapshots zugreifen können und die Administratoren von zahlreichen Restore-Anfragen entlastet werden. Stützt man zudem das Backup auf einen Snapshot, löst sich auch das Problem der geöffneten Files (Fehlermeldung *ANE4987E Error processing '<NAME>': the object is in use by another process*).

Voraussetzung für diesen Ansatz ist natürlich, dass die Filesysteme Snapshots unterstützen – und zwar in ausreichender Menge.

BESCHLEUNIGUNG MIT ISP-BORDMITTELN

IBM bietet für die Beschleunigung des Backups einige Bordmittel an:

Vereinfachte Identifikation ausschließlich über das Änderungsdatum

Üblicherweise vergleicht der ISP-Client zahlreiche Meta-Daten, um Objekte für das erneute Backup auszuwählen. Neben

dem Datum der letzten Änderung sind dies auch Dateigröße, Prüfsumme und Zugriffsrechte/ACLs. Im interaktiven Aufruf *dsmsc i* und/oder als „Object“ im Client-Schedule kann durch die Option *-INCRbydate* die Prüfung auf den Vergleich des Änderungsdatums des Objektes mit dem letzten Backup reduziert und somit erheblich beschleunigt werden. Allerdings birgt die Option auch einige Probleme: Insbesondere wenn keine Snapshots genutzt werden oder falls das Backup abbricht, fallen Dateien, die während des Backups in bereits abgearbeiteten Ordnern geändert oder erstellt werden, beim nächsten Lauf mit *-incrbydate* heraus, wenn sie nicht nochmal verändert wurden. IBM empfiehlt daher ausdrücklich, regelmäßig ein normales „incremental“ auszuführen [6]. Ähnliche Probleme können auftreten, wenn Client und Server unterschiedliche Systemzeiten haben.

Ein weiterer wichtiger Punkt: Gelöschte Dateien werden nicht erkannt, verbleiben im Backup, und Dateien, die mit einem alten Datum ins System kommen, z. B. Softwareinstallation, werden nicht gesichert!

Ausschalten von ACLs und Prüfsummen

Die Verarbeitung (und damit zuvor die Prüfung) von ACLs und das Erstellen von Prüfsummen verlangsamt den Identifikationsprozess und kann über mehrere Optionen beeinflusst werden. Es sollte aber gut abgewogen werden, ob der Geschwindigkeitsgewinn gegenüber dem Verlust an Informationen ausreichend überwiegt.

- *skipacl* unterbindet die ACL-Verarbeitung vollständig, wahrscheinlich werden die ACLs aber auch nicht gesichert (Option nur für UNIX und Mac OS).
- *skipaclupdatecheck* schaltet auch die Checksummenermittlung ab (Option nur für UNIX und Mac OS).
- *skipntsecuritycrc* unterbindet die Berechnung einer CRC-Prüfsumme (Option nur für Windows).

Parallelisierung von Backups bei mehreren Filespaces

Liegen die zu sichernden Daten auf mehreren Partitionen, lässt sich der Backupprozess über die Option *RESSOURCEUTILIZATION* auf bis zu 100 parallele Streams verteilen und damit sowohl die Bandbreite besser ausnutzen, als auch durch die

Backup of large Data Areas with TSM

As data grows, filesystems do so, too. This growth also has an impact for the backup of the data because the time for backup does not increase. Using IBM Spectrum Protect „ISP“ (formerly known as „TSM“) data can be backed up incrementally – forever without any additional full backup once started. But the amount of data is just one problem that will be solved by the increasing speed and capacity of the backup hardware. Even more important becomes the time to identify those data, that has to be backup up (so called „backup candidates“). In this article several approaches for speeding up this „seeking the candidates“ are described, starting with some options for improvement ISP offers by itself. Further more a short view follows looking on special filesystems that supports directly an backup with ISP. In the last part a generic approach for a parallelization of seeking-threads is shown. Starting from a simple idea some improvements are discussed.

Parallelisierung die Suchzeit erheblich verringern. Da durch diese Parallelisierung auch zusätzliche Sessions auf Seiten des ISP-Servers erzeugt werden, muss gegebenenfalls die Anzahl der MAX-SESSIONS erhöht werden.

Explizites Backup nur der geänderten Dateien

Sofern Informationen vorliegen, welche Dateien sich seit dem letzten Backup ändern haben und welche Dateien seitdem gelöscht wurden, kann ISP auch ausschließlich eine Sicherung dieser Dateien vornehmen. Statt einem „incremental Backup“ ist dann ein „selective Backup“ mit der expliziten Angabe dieser Dateien möglich:

- `dsmc sel –filelist=<DATEI mit Namen geänderter Dateien>`
- `dsmc expire –filelist=<DATEI mit Namen gelöschter Dateien>`

Das Grundprinzip des „selective backup“ wird auch in den nachfolgenden Ansätzen genutzt.

JournalBasedBackup / FilepathDemon

Bereits seit TSM 5 bietet IBM die Methode des „Journal Based Backup“ (JBB) an. Hierbei wird über den „JBB-Demon“ (bzw. „FilepathDemon“) das zu sichernde Filesystem überwacht und Informationen zu neuen, geänderten und gelöschten Dateien gesammelt. Beim Backup werden diese Informationen vom TSM/ISP-Client analog dem „selective backup“ genutzt. Der Aufwand für die Identifikation der Backupkandidaten entfällt und das Backup reduziert sich auf die Übertragung der neuen/geänderten Daten.

Tests bei der GWDC mit einem Linux-Fileserver mit ca. 150 TB Kapazität verteilt über 22 Filespaces waren aber nicht erfolgreich: Obwohl der zusätzliche Ressourcenbedarf für den JBB erheblich war, fiel der Zeitgewinn, vor allem durch regelmäßige Neuindizierungen, doch eher gering aus. In anderen Konstellationen mag der JBB durchaus deutliche Vorteile bringen.

Daneben gibt es noch eine wichtige Einschränkung: „Journal Based Backup“ geht nur mit lokalen Filesystemen. CIFS/NFS und Cluster-Filesysteme gehen nicht.

Hinweis: Optimierungen für die Datenübertragung finden sich im Performance Tuning-Guide [6].

FILESYSTEME, DIE EIN SCHNELLES BACKUP UNTERSTÜTZEN

Einige Filesysteme/Filer unterstützen ein schnelles Backup mittels ISP, indem sie die notwendigen Backupkandidaten ermitteln und dem ISP-Client zur Verfügung stellen. (Diese Liste ist nur eine Auswahl):

IBM Spectrum Scale (ehemals GPFS)

IBMs Cluster-Filesystem unterstützt natürlich das Backup mit ISP und bietet sogar ein eigenes Skript „mmbackup“ an. Dieses nutzt nicht nur die Informationen zu den Backupkandidaten, sondern kann auch die Datenübertragung über mehrere (ISP-)Knoten und GPFS-Server parallelisieren. Die Max Planck Computing and Data Facility (MPCDF) kann die mancherorts gemachten Erfahrungen, dass „mmbackup“ nicht einfach „out-of-the-box“ läuft, nicht bestätigen: Die anfängliche Erstellung der Konfiguration erfordert zwar ein wenig Einarbeitung, danach läuft „mmbackup“ sowohl stabil wie performant. Für den Erfahrungsaustausch steht die

MPCDF gern zur Verfügung. IBM Spectrum Scale bietet außer zu ISP auch eine enge Verzahnung mit HPSS als HSM-System, so dass sich das Problem auch durch (teilweise) Verlagerung der Daten ins HPSS verringern lässt – wobei ISP/ISS auch sehr große Datenmengen in vergleichsweise kurzer Zeit sichern kann.

NetApp SnapDiff

Auch NetApp unterstützt das Backup der eigenen NAS-Filer bereits seit TSM 5 in vielfältiger Art und Weise. Neben NDMP gehört über die Funktion „SnapDiff“ auch eine Beschleunigung des inkrementellen Backups dazu. Beim SnapDiff werden die Änderungen an Dateien und Verzeichnissen zwischen zwei Snapshots an den ISP-Client übergeben. Die Integration geht soweit, dass der ISP-Client die benötigten Snapshots sogar auf dem Filer auslösen kann und nach einem erfolgreichen Backup den vorletzten selbstständig löschen kann. Die Anleitung zum „SnapDiff“ liest sich zunächst einfacher, als es in der praktischen Umsetzung dann der Fall ist.

Grundsätzlich sollte jedes Cluster/ScaleOut-Filesystem eine Liste mit neuen, geänderten und gelöschten Dateien bereitstellen können, da diese (Meta-)Informationen für die Konsistenz der Daten (und ganz besonders der Caches) auf den Clusterknoten notwendig sind. In der Praxis bestehen die Probleme darin, dass diese Informationen nicht einfach zugänglich sind und es seitens der Hersteller keine Werkzeuge zum Zugriff auf diese Daten gibt. Quantum hat den Bedarf auf Kundenseite wahrgenommen und prüft derzeit, wie für das StorNext-Filesystem diese Informationen bereitgestellt werden können.

EINE IDEE FÜR ALLE FILESYSTEME

Damit stellt sich für alle Nutzer, die weder IBM Spectrum Scale noch eine NetApp in Betrieb haben („mmbackup“ und „SnapDiff“ sind hierfür jeweils die besten Lösungen!) und weder Vollsicherungen noch NDMP machen wollen, die Frage: Was nun?

Wie zuvor ausgeführt, benötigt das Identifizieren der Backupkandidaten beim ISP-Backup die meiste Zeit. Dieser Prozess untersucht den gesamten Dateibaum des zu sichernden Filesystems – sequentiell in einem einzelnen Thread. Die Lösung liegt also darin, aus diesem einen Prozess mehrere parallele zu machen.

Nutzer lassen sich meist in Gruppen einteilen (z. B. Arbeitsgruppen oder Institute), besonders in Universitäts-Umgebungen findet sich diese Einteilung auch in Filesystemen wieder, da dort zur einfacheren Zugriffssteuerung häufig zunächst eine Orderebene mit Fakultäten oder Instituten existiert, und unterhalb dieser Ebene liegen dann die Nutzer- und Arbeitsgruppenverzeichnisse.

Variante 1

Eine Parallelisierung des Backups ist möglich, wenn man statt eines ISP-Knotens für das gesamte Filesystem jeweils jeder Fakultät bzw. jedem Institut einen eigenen Knoten einrichtet und das Backup „fakultätsweise“/„institutsweise“ erfolgt. Statt eines einzelnen Prozesses durchsuchen nun mehrere parallel das Filesystem (Fileserver sollten in der Lage sein sogar mehrere Hundert parallele Prozesse zu verarbeiten), die Suchzeiten sollten sich erheblich reduzieren. In der Praxis offenbaren sich bei diesem Ansatz mindestens zwei Probleme:

- Nicht alle Fakultäten/Institute haben gleich viele Daten. Meist gibt es ein oder zwei, die allein nahezu die gesamte

Kapazität des Filesystems nutzen. Daher laufen die Backups der übrigen wesentlich schneller, für die „Großnutzer“ verringert sich die Backupzeit im schlimmsten Falle nur geringfügig und das Backupzeitfenster reicht weiterhin nicht aus.

- Kommen weitere Fakultäten (wohl nicht so häufig) oder Institute hinzu, muss der Backup-Administrator rechtzeitig seine Konfiguration anpassen, sonst bleiben die neuen außen vor.

Unter UNIX lassen sich die Knoten relativ elegant über „VIRTUALMOUNT“s trennen, für Windows muss man für jeden Knoten entsprechenden *exclude.dir*-Regeln erstellen.

Variante 2

Häufig sind die Benutzer auf den Filesystemen aber nicht in Gruppen organisiert, sondern alle Verzeichnisse liegen flach nebeneinander auf der Einstiegsebene. Für jedes Benutzerverzeichnis einen eigenen ISP-Knoten anzulegen, wiederholt zum einen das zweite zuvor genannte Problem und ist in Anbetracht der Anzahl von Nutzern sehr aufwändig.

Einfacher ist es daher, die Verzeichnisse anhand eines Musters zu unterscheiden, beispielsweise nach dem (oder den) ersten Zeichen: „^[a,A]“, „^[b,B]“, ..., „^[z,Z]“, „^[0-9]“ (ISP versteht an dieser Stelle sogar „regular Expressions“!)

Man erhält also 27 bzw. 729 ISP-Knoten, die automatisch auch alle neuen Verzeichnisse umfassen. Unglücklicherweise erfassen die RegEx-Formulierungen nur die Verzeichnisse, die existieren, nicht aber die gelöschten. Abhilfe ist möglich, wenn man zusätzlich alle Verzeichnisse des Startpfades ohne Unterverzeichnisse sichert.

Obwohl diese Variante häufig besser als die erste ist, erfüllt sie nicht alle Erwartungen:

- Das Lösen des „gelöschte Verzeichnisse“-Problems ist umständlich.
- Die Konfiguration wird – besonders, wenn man nach den ersten beiden Buchstaben unterscheidet, sehr umfangreich.
- Änderungen der Verzeichnisnamen verteilen die Daten über mehrere ISP-Knoten und besonders der Restore wird aufwändig.

EIN ANSATZ FÜR ALLE FILESYSTEME

Bereits im letzten Jahrzehnt stand die (damalige) Generali Versicherungs-AG vor dem eingangs skizzierten Problem und Rudolf Wüst als Backup-Administrator erweiterte die zuvor genannte Idee um einen entscheidenden Ansatz. Hieraus entwickelte er eine Lösung, die erfolgreich mit bis zu 2.000 Threads das „Such-Problem“ parallelisierte. Freundlicherweise teilte Herr Wüst seine Erweiterung gern und der Autor dieses Artikels hat sie aufgegriffen. Im Rahmen seiner Tätigkeit bei der GWDG entwickelt er seine ursprünglich einfachen Skripte nun weiter.

Ziel einer praktikablen Lösung muss es sein, sowohl alle Verzeichnisse zu erfassen, sie in einem ISP-Knoten zu speichern und dennoch das Suchen zu parallelisieren. Dies gelingt, wenn man statt des einfachen Backupaufrufs ein Skript ausführt, das seinerseits mehrere parallele Threads zum Sichern der Verzeichnisse startet. Der Kern des Skriptes besteht also in einer Schleife der folgenden Form (siehe Beispiel 1).

Für alle (finde alle Verzeichnisse im Startpfad)

```
{
    Starte einen Backup-Thread für das aktuelle Verzeichnis
}
```

Beispiel 1: PseudoCode

Statt eines „incremental backups“ werden also viele „partial incremental backups“ ausgeführt und zwar für jedes Verzeichnis einer.

Die gelöschten Verzeichnisse erfasst man mit einem anschließenden Backup des Startpfades ohne Unterverzeichnisse – die letzte Angabe ist enorm wichtig, sonst macht man ein normales „incremental backup“ auf dem gesamten Filesystem.

Als Quelltext für die BASH sieht dies wie in Beispiel 2 aus.

```
startpath=<Startpfad>;
folderlist=<Pfad zu einer Datei mit Ordernamen>;
find $startpath -xdev -mindepth 1 -maxdepth 1 -type d \
-print > $folderlist;
for $i in $(cat $folderlist)
do
    dsmc -i $i -subdir=yes &
done
dsmc -i $startpath -subdir=no;
rm $folderlist;
```

Beispiel 2: Quellcode BASH

Im Rahmen der ersten Tests wird man herausfinden, dass das Skript in der vorliegenden Form tatsächlich ebenso viele Threads wie vorhandene Verzeichnisse startet. Zum einen zwingt man damit den Rechner, der das Backup ausführt, in die Knie, zum zweiten wird wahrscheinlich quasi sofort die „MaxSessions“-Einstellung des ISP-Servers erreicht und dieser verweigert weitere Verbindungen.

Abhilfe schafft ein Zähler, der bei Erreichen der erlaubten Anzahl von Threads einfach wartet. Da die abgespaltenen Backup-Threads als „parent process ID“ die „process ID“ des Skriptes selbst haben, lassen sich diese Threads sogar dann zählen, wenn man das Skript für mehrere Filesysteme gleichzeitig laufen lässt. Als BASH-Code sieht die Schleife damit wie in Beispiel 3 aus.

In der erweiterten Form werden nun wesentliche Ziele erreicht, aber vollständig zufrieden kann man dennoch nicht sein:

- Für das Reporting über den ISP-Server fehlt ein Rückgabewert. Im einfachsten Fall kann man eine Zeile *return 0* am Ende anhängen; dann ist das Schedule immer erfolgreich – unabhängig davon, ob Fehler auftreten oder nicht. Wie im Beispiel 3 schon ergänzt, sollte man also eher die Ausgabe der einzelnen „partial incremental backups“ einsammeln und zum Ende des Skriptes auswerten, beispielsweise nach Fehlern suchen oder die „Summaries“ zusammenfassen. Anhängig von der Art (ggfs. auch Anzahl) der Fehler und der „Files failed“ kann dann das Skript die passenden Rückgabewerte geben. (Diese Erweiterung ist im veröffentlichten Quellcode bereits enthalten.)

```

pid=$$; # parents process id
startpath=<Startpfad>;
folderlist=<Pfad zu einer Datei mit Ordernamen>;
maxthreads=<max. Anzahl paralleler Threads>;
find $startpath -xdev -mindepth 1 -maxdepth 1 -type d \
-print > $folderlist;

while [ -s $folderlist ]
do
  nthr=$(ps axo ppid,cmd | grep $ppid | grep -v grep | wc -l)
  if [ $nthr -le $maxthreads ]
  then
    # get new start path
    folder=$(head -n 1 < $folderlist);

    # backup actual folder
    dsmc i $folder/ -subdir=yes -quiet >> $ppid.log &

    # remove first line from folderlist
    sed -i '1 d' $folderlist
  else
    sleep 5; # wait to complete another thread
  fi;
done
dsmc -i $startpath -subdir=no

# Am Ende auf alle laufenden Threads warten
while [ $nthr -gt 1 ]
do
  >&2 echo "Waiting for $nthr threads to end"
  sleep 60;
  nthr=$(ps axo ppid,cmd | grep $ppid | grep -v grep | wc -l)
done
rm $folderlist;

```

Beispiel 3: Erweiterter BASH-Code

- Das bei der Idee genannte Problem, dass einzelne Verzeichnisse einen nennenswerten Anteil der Filesystemgröße nutzen und damit die Laufzeit des Backups maßgeblich beeinflussen, wird durch das Skript nicht gelöst. Die Ungleichheit der Datenmenge/Objektanzahl wird sicherlich kleiner, verlagert sich letztendlich aber nur. Als nächsten Schritt könnte man daher die Verzeichnisliste über mehrere Ebenen erstellen und die somit größere Anzahl partieller Backups vergrößern. Im Ergebnis sollte sich die Ungleichheit nochmal stärker ausgleichen.
- Im Reporting des ISP-Server lassen sich neben dem Return Code eines „Client-Schedules“ natürlich auch detailliert Fehlermeldungen und in Form der Summary ein Überblick auslesen. Dies ist mit dem angegebenen Skript (derzeit) nicht möglich, es liefert nur einen ampelhaften Status über den Return Code.

OFFENE PUNKTE / AUSBLICK

Wie im letzten Abschnitt genannt, sind noch einige Fragen

offen, beispielsweise nach dem Übertragen der Summary in das Serverlog. Ebenso ist das Skript bisher nur für Linux implementiert, eine Windows-Variante auf Basis der PowerShell wird aktuell erstellt.

Für eine „runde Sache“ sollte noch eine Fehlerbehandlung ergänzt werden. Ein Problem, das nicht gelöst werden kann, ist die Tatsache, dass „partial incremental backups“ nicht die „Last Backup“-Attribute der Knoten bzw. Filespaces ändern und dies natürlich auch nicht im Rahmen des skizzierten Skriptes erfolgt. Von einem schreibenden Eingriff in die DB2 der ISP-Server sollte man absehen, da dies die Gewährleistung durch IBM berührt. IBM untersagt **ausdrücklich** direkte Zugriffe auf die ISP-DB2 außerhalb von entsprechenden Anweisungen im Rahmen des Supports.

UND WIE BESCHLEUNIGT MAN DEN RESTORE?

Die zuvor genannten Ansätze mit ISP-Bordmitteln und der skizzierte Ansatz für die Parallelisierung funktionieren erst einmal nur für das Backup. Sollen viele Dateien aus dem Backup restored werden, ist dies bei den Ansätzen mit mehreren Knoten für einen Filespace sehr einfach, da sowieso für jeden Knoten ein eigener Restore laufen muss und die Prozesse hierbei parallel laufen. Für den Ansatz der parallelen Threads ist eine Anpassung für den Restore auf Basis einer Filelist einfach möglich: Statt einer „Folderliste“ wird beim Restore über eine Dateiliste parallelisiert.

Zu bedenken ist aber, dass in einer Umgebung mit einer Tapelibrary als Storage-Backend die Leistung des Restore in der Regel durch die Anzahl der Laufwerke begrenzt ist. Außerdem organisiert ISP den Restore normalerweise (also ohne die Option `-disablenqr=yes`) so, dass die Tapemounts optimiert werden. Wird eine Filelist parallel durch zahlreiche parallele Threads abgearbeitet, so kann der Server die Tape-Zugriffe nicht optimieren. Kommt allerdings ein disk-basierter FILE- oder Container-Pool zum Einsatz ist der parallele Restore über zahlreiche Threads schneller. Sofern die Daten über Serverreplikation auf zwei Servern vorgehalten werden, kann der Restore auch über beide Server verteilt und damit zusätzlich beschleunigt werden.

Erfahrungen zeigen leider, dass „Full Restores“ auch bei Parallelisierung einen enormen Aufwand bedeuten und sich nur unbefriedigend beschleunigen lassen.

VERFÜGBARKEIT / ZUGANG ZUM QUELLCODE

Es ist davon auszugehen, dass weder der genannte Ideengeber noch die GWDG den Anspruch erheben können, als einzige die skizzierte Idee gehabt und implementiert zu haben. Vielmehr dürften zahlreiche TSM/ISP-Nutzer vor dem gleichen Problem stehend ähnliche Lösungen gefunden haben.

Eine kommerzielle Implementierung des Ansatzes der Parallelisierung (derzeit nur für Windows) findet sich im Produkt „MAGS“ der General Storage [7]. Neben einem verbindlichen Support bietet „MAGS“ auch eine „hübsche“ webbasierte Oberfläche und eine intuitive Bedienung. Eine detailliertere Produktbetrachtung soll hier nicht erfolgen, den Mehrwert muss man auch individuell ermitteln.

Das im Rahmen dieses Artikels angesprochene Skript ist für registrierte Benutzer im GitLab [8] der GWDG frei verfügbar. Die Skripte dürfen uneingeschränkt eingesetzt und verändert werden. Über eine Rückmeldung und Vorschläge freuen wir uns natürlich.

ÜBERTRAGBARKEIT AUF ANDERE BACKUPLÖSUNGEN

Die dargestellten Ansätze adressieren das Problem der Identifikation von Dateien und sind somit natürlich auf alle anderen Fragestellungen übertragbar, wo eine Dateiliste erstellt werden soll. Ersetzt man den Aufruf der ISP-CLI durch ein weiteres *find*, lassen sich auch parallel alle Dateien finden, über entsprechende Parameter auch gefiltert nach allen von *find* unterstützten Attributen. Ebenso kann man eine weitere Schleife ergänzen, die beliebige Operationen mit allen Einträgen einer vollständigen Dateiliste macht. Damit lassen sich auch andere Backup-Lösungen optimieren, die eine Verzeichnis- oder Fileliste verarbeiten können.

DANKSAGUNG

Der Autor dankt den Herren Gerd Becker (Empalis GmbH), Wolfgang Hitzler (IBM) und Manuel Panea (MPCDF) für das Korrekturlesen des Artikels und die Änderungs- und Verbesserungsvorschläge. Besonderer Dank gilt Herrn Rudolf Wüst (Generali Shared Services S.c.a.r.l.) für die Freigiebigkeit, seine Ideen zu teilen.

FUSSNOTEN UND LINKS

- [1] <http://www.storageconference.us/2014/Presentations/Shimizu.pdf>
- [2] <http://www.insic.org/news/2015%20roadmap/15pdfs/2015%20Technical%20Roadmap.pdf>
- [3] Nach Auswertung der GWDG-ISP-Server zwischen 15 % und 84 % zusätzlich zu den aktiven Daten, wobei die 84 % ein Ausreißer ist; der Mittelwert liegt bei 39 %.
- [4] $10GE \pm 10 \text{ Gb/s} \approx 108 \text{ TB/d}$
- [5] Die tatsächliche Bandbreite ist deutlich kleiner als 10GE; der Durchsatz am Quellsystem muss 10 Gb/s liefern; es müssen ausreichend viele und schnelle Laufwerke und entsprechender Staging-Bereich am Backup-Server vorhanden sein.
- [6] https://www.ibm.com/support/knowledgecenter/SSGSG7_7.1.6/client/r_opt_incrbydate.html
- [7] http://www.general-storage.com/PRODUCTS/dsm/ISI-MAGS/body_dsmisi-mags.html
- [8] <https://gitlab.gwdg.de/bnachtw/dsmci>



Software und Lizenzverwaltung

Der einfache Weg zur Software!

Ihre Anforderung

Sie benötigen eine Software, für die es keine von Ihnen nutzbare Rahmenvereinbarung gibt. Die Anzahl der erforderlichen Lizenzen ist nicht genau festgelegt.

Unser Angebot

Wir verfügen über eine Reihe von Rahmen- und Campusvereinbarungen mit namhaften Softwareherstellern und -lieferanten, über die Software auch in geringerer Stückzahl bezogen werden kann. Wir wickeln für Sie die Beschaffung der erforderlichen Lizenzen ab. Wir können uns bei Vertragsverhandlungen und Bedarfsanalysen engagieren. Zugriffslizenzen können auch über Lizenzserver verwaltet werden.

Ihre Vorteile

- > Sie können die benötigte Software in vielen Fällen sofort nutzen.

- > Sie brauchen kein eigenes Ausschreibungs- und Beschaffungsverfahren durchzuführen.
- > Sie ersparen sich die zeitraubenden Verhandlungen mit den Softwareherstellern und -lieferanten.
- > Die Anzahl der benötigten Lizenzen wird Ihnen flexibel zur Verfügung gestellt.
- > Wir können die Nachfrage von verschiedenen Nutzern für neue Lizenzvereinbarungen bündeln.

Interessiert?

Informationen zu bestehenden Lizenzvereinbarungen sind auf der u. g. GWDG-Webseite zu finden. Falls Sie nach spezieller Software suchen, die noch nicht auf unserer Webseite erwähnt ist, kommen Sie bitte auf uns zu. Wir werden prüfen, ob wir eine Vereinbarung abschließen können und bündeln die Nachfrage mit anderen Nutzern.

>> www.gwdg.de/software



33. DV-Treffen der Max-Planck-Institute in Göttingen

Text und Kontakt:
Dr. Thomas Otto
thomas.otto@gwdg.de
0551 201-1828

Einmal im Jahr treffen sich IT-Spezialisten aus nahezu allen Max-Planck-Instituten und weiteren Einrichtungen der Max-Planck-Gesellschaft zu ihrem traditionellen DV-Treffen. Das diesjährige 33. Treffen fand vom 4. bis 6. Oktober auf dem Max-Planck-Campus in Göttingen statt und die GWDG fungierte wie immer, wenn das DV-Treffen dort stattfindet, als lokaler Organisator. Knapp 150 Teilnehmer erlebten ein abwechslungsreiches Programm mit vielen interessanten Beiträgen zu aktuellen IT-Themen.

ABWECHSLUNGSREICHES PROGRAMM

Das DV-Treffen der Max-Planck-Institute richtet sich an alle IT-Mitarbeiter und IT-Interessierte innerhalb der Max-Planck-Gesellschaft. Sinn und Zweck des Treffens sind vor allem der Austausch von Ideen und Erfahrungen, das Erkunden von Trends und neuen Entwicklungen und nicht zuletzt das Knüpfen von Kontakten zwischen den Instituten. Dem Programmkomitee war es auch in diesem Jahr wieder gelungen, ein abwechslungsreiches Programm zusammenzustellen, das ein breites Spektrum von aktuellen Themen abdeckte, die in zahlreichen Vorträgen und Workshops behandelt wurden.

VORTRÄGE

Ein Schwerpunktthema und „Dauerbrenner“ der DV-Treffen war auch diesmal die IT-Sicherheit bzw. der Betrieb einer sicheren IT-Infrastruktur, mit der sich mehrere Vorträge in unterschiedlichen Facetten befassten. Hierzu gehörten die E-Mail-Sicherheit wie auch Monitoring und Netzwerkkonzepte für Lastenverteilung und Ausfallsicherheit. Ein weiterer Themenblock widmete sich dem Bereich Backup und Archivierung.

Im Rahmen der Berichte aus den drei Sprecherkreisen „IT-Sprecherkreis“, „Haustechnik-Sprecherkreis“ und

„Bibliotheks-Sprecherrat“ wurden die Teilnehmer mit interessanten Informationen versorgt, die z. T. auch über den Tellerrand der eigentlichen IT hinausgingen und deutlich machten, dass die IT mittlerweile in nahezu allen Bereichen der Institute und Einrichtungen eine zentrale Rolle einnimmt. Das wurde auch in den beiden Berichten des „Lizenz-Arbeitskreises“ und des „IT4Science-Arbeitskreises“ deutlich.

Die MPCDF (früher bekannt als RZG) und die GWDG als zentrale IT-Dienstleister der Max-Planck-Gesellschaft informierten als fester Programmbestandteil der DV-Treffen auch diesmal wieder über ihre Neuigkeiten und stellten dabei insbesondere neue

33. IT Meeting of the Max Planck Institutes in Göttingen

Once a year, IT specialists from almost all Max Planck Institutes and other institutions of the Max Planck Society meet for their traditional IT meeting. This year's 33rd meeting took place from the 4th to the 6th of October on the Max Planck Campus in Göttingen. As always when the meeting is held there the GWDG acted as local organizer. Nearly 150 participants experienced a varied program with many interesting presentations on current IT topics.

Dienste für die Max-Planck-Gesellschaft vor.

Ebenso Tradition bei DV-Treffen sind auch die Berichte aus verschiedenen Abteilungen bzw. Bereichen der MPG-Generalverwaltung. Diesmal wurde neben den Neuigkeiten vom Beratenden Ausschuss für Rechenanlagen (BAR) und vom Datenschutz auch über den virtuellen Verwaltungsarbeitsplatz (AP 2016) sowie das Identity & Access Management (IAM) für Verwaltungs-IT-Dienste berichtet.

WORKSHOPS

Fester Bestandteil im Programm der DV-Treffen ist ein Workshop-Block. Diesmal fanden parallel in mehreren Räumen insgesamt acht einstündige Workshops zu verschiedenen Themen statt: Sicherheit in der Haustechnik, Windows 10, Forschungsdatenmanagement, Sicherheit in der TK-Anlage, Apple iOS – Client-/Softwaremanagement, Vergabeverfahren, „Hacken macht Spaß“ und VoIP/VC in der MPG. Kurzfristig wurde für Interessenten zudem eine Führung durch die Experimentierhalle des Max-Planck-Instituts für Dynamik und Selbstorganisation organisiert.

KEYNOTE

Als „Highlight“ und zum Abschluss des ersten Tages hielt Prof. Dr. Bruce Allen vom MPI für Gravitationsphysik (Albert-Einstein-Institut), Teilinstitut Hannover eine Keynote zum Thema „The Discovery of Gravitational Waves“. In seinem einstündigen Vortrag berichtete er auf kurzweilige und spannende Weise, wie sein Institut an der Entdeckung der Gravitationswellen beteiligt war, deren Existenz Albert Einstein schon vor 100 Jahren voraussagte und wonach Physiker 50 Jahre suchten.



POSTERSESSION UND LIVE-DEMONSTRATIONEN

Zum Abschluss des zweiten Tages konnten sich die Teilnehmer bei einer Postersession und Live-Demonstrationen ausführlicher über interessante Projekte aus verschiedenen Bereichen bzw. Instituten informieren. Dabei standen jeweils Ansprechpartner bereit und es entwickelte sich ein oft intensiver Gedankenaustausch.

WAHL DES IT-SPRECHERKREISES

Turnusmäßig wurde beim DV-Treffen auch der IT-Sprecherkreis neu gewählt. Er versteht sich als Sprachrohr der IT-Abteilungen und als Mittler zu und zwischen allen Schnittstellen der IT, internen und externen Gremien, anderen Forschungseinrichtungen

und Firmen. Folgende Personen wurden am 04.10.2016 in den IT-Sprecherkreis gewählt:

- Thomas Baumann (MPI für Menschheitsgeschichte)
- Helmut Hayd (MPI für Kognitions- und Neurowissenschaften)
- Jörg Herrmann (Gemeinsame Verwaltung, MPI für Informatik, MPI für Softwaresysteme)
- Heinz Junkes (Fritz-Haber-Institut)
- Rainer Kleinrensing (MPI für Mathematik in den Naturwissenschaften)
- Andreas Oberreuter (MPI für Radioastronomie)
- Gerry Truschkewitz (MPI für Dynamik komplexer technischer Systeme)

VERLEIHUNG DES IT COMMUNITY AWARDS

Seit dem Jahr 2010 wird bei den DV-Treffen der IT Community Award verliehen. Er würdigt Personen, die einen wesentlichen Beitrag zur Förderung des Community-Gedankens in der IT der MPG geleistet haben. Mit diesem Preis soll die Anerkennung der IT Community zum Ausdruck gebracht werden, dass ihr Beitrag vielen geholfen hat, ihre Arbeit besser und erfolgreicher zum Ziel zu führen. Die Preisträgerin in diesem Jahr ist Christa Hausmann-Jamin, die bis zu ihrem Eintritt in den Ruhestand im Mai diesen Jahres viele Jahre in verschiedenen Einrichtungen und Gremien der MPG tätig war, zuletzt mehrere Jahre als IT-Leiterin am MPI für Gravitationsphysik mit den Teilinstituten in Potsdam und Hannover. Sie erhält den Preis für ihr Engagement in der Organisation mehrerer DV- und IT-Verantwortlichen-Treffen als auch die vielen aktiven Diskussionsbeiträge in Mailinglisten und Arbeitskreisen.



FAZIT

Das diesjährige 33. DV-Treffen der Max-Planck-Institute machte einmal mehr deutlich, welche zentrale Rolle eine dezentrale leistungsfähige und innovative IT-Infrastruktur für die Spitzenforschung in allen Wissenschaftsbereichen der Max-Planck-Gesellschaft spielt. Die IT-Abteilungen der Max-Planck-Institute stellen zusammen mit den wissenschaftlichen Rechenzentren der MPG die dafür notwendige Infrastruktur bereit. Das DV-Treffen spielt dabei eine wichtige Rolle als Plattform für die Kommunikation und den Informationsaustausch zwischen allen beteiligten IT-Einrichtungen, um sich den ständig neuen Herausforderungen zu stellen und die passenden Lösungen dafür zu finden.

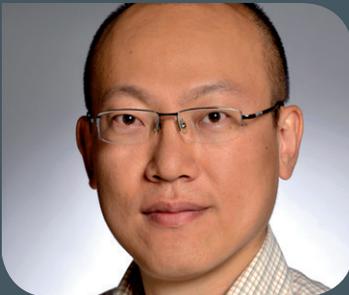
Unterlagen und Streams zum 33. DV-Treffen können unter dv-treffen@mpg.de angefordert werden. ■

NEUE MITARBEITERIN LENA STEILEN

Seit dem 1. Oktober 2016 ist Frau Lena Seilen in der Arbeitsgruppe „eScience“ (AG E) tätig. Sie arbeitet für die Göttingen eResearch Alliance (eRA), einem Service der Georg-August-Universität Göttingen – kooperativ betrieben von der GWDC und Niedersächsischen Staats- und Universitätsbibliothek Göttingen (SUB) – für alle Forschenden am Göttingen Campus, der IT- und informationswissenschaftliche Angebote und Aktivitäten am Standort bündelt und als zentraler Ansprechpartner fungiert. Frau Steilens Arbeitsschwerpunkte innerhalb der eRA liegen in den Bereichen Verwaltung, Gremien- und Öffentlichkeitsarbeit sowie Forschungsdatenmanagement. Sie bringt breit gefächerte Erfahrungen aus den Bereichen Digital Humanities und Forschungsdatenmanagement aus früheren Tätigkeiten in der SUB und ihrem erfolgreich abgeschlossenen Magisterstudium in Mittlerer und Neuerer Geschichte sowie Europäischer Ethnologie/Kulturanthropologie an der Georg-August-Universität Göttingen mit. Darüber hinaus ist sie durch diverse Tätigkeiten im öffentlichen Dienst und der privaten Wirtschaft versiert in den Feldern Organisation, Veranstaltungsmanagement und Kundenservice. Frau Steilen ist per E-Mail unter lena.steilen@gwdg.de zu erreichen.



Wieder



NEUER MITARBEITER DR. HAO WANG

Seit dem 1. Oktober 2016 ist Herr Dr. Hao Wang als wissenschaftlicher Mitarbeiter in der Arbeitsgruppe „eScience“ (AG E) tätig. Er arbeitet dort in dem gerade begonnenen Forschungsprojekt „SENDATE-Secure-DCI“ in den Bereichen „Network Function Virtualization“, „Software-Defined Networks“ und „Cloud Computing“. Vor seinem Eintritt in die GWDC war Herr Dr. Wang seit 2009 wissenschaftlicher Mitarbeiter in der Arbeitsgruppe „Distributed Computer Systems“ (DISCO Lab) im Department of Computer Science der Technischen Universität Kaiserslautern. Im Juli 2015 wurde ihm nach Abschluss seiner Dissertation zum Thema „Modeling and Performance Analysis of Networks with Flow Transformations“ der Titel Dr.-Ing. in Informatik verliehen. Sein Forschungsfokus im DISCO Lab lag in den Feldern Leistungsmodellierung und -analyse von verteilten und vernetzten Systemen. Herr Dr. Wang bringt ferner einen Master-Abschluss in Informatik der Technischen Universität Kaiserslautern und einen Bachelor-Abschluss in Informatik der Northeastern University Shenyang, China mit. Darüber hinaus kann er auf praktische Erfahrungen als Software-Engineer in der Industrie (Implementierung von Autonavigationssystemen) sowie aus Praktika bei den Telekom Innovation Laboratories (T-Labs), Berlin zurückgreifen. Herr Dr. Wang ist per E-Mail unter hao.wang@gwdg.de und telefonisch unter 0551 201-2118 zu erreichen.

Wieder

INFORMATIONEN:
support@gwdg.de
0551 201-1523



November 2016 bis
Dezember 2017

Kurse

KURS	VORTRAGENDE/R	TERMIN	ANMELDEN BIS	AE
INDESIGN - AUFBAUKURS	Töpfer	02.11. – 03.11.2016 9:30 – 16:00 Uhr	26.10.2016	8
EINFÜHRUNG IN DIE STATISTISCHE DATENANALYSE MIT SPSS	Cordes	08.11. – 09.11.2016 9:00 – 12:00 und 13:00 – 15:30 Uhr	01.11.2016	8
DIE SHAREPOINT-UMGEBUNG DER GWDG	Buck	10.11.2016 9:00 – 12:30 und 13:30 – 15:30 Uhr	03.11.2016	4
EINFÜHRUNG IN DAS IP-ADRESSMANAGEMENTSYSTEM DER GWDG FÜR NETZWERKBEAUFTRAGTE	Dr. Beck	15.11.2016 10:00 – 12:00 Uhr	08.11.2016	2
ADMINISTRATION VON PCS IM ACTIVE DIRECTORY DER GWDG	Quentin	17.11.2016 9:00 – 12:30 und 13:30 – 15:30 Uhr	10.11.2016	4
MAC OS X IM WISSENSCHAFTLICHEN ALLTAG	Bartels	21.11. – 22.11.2016 9:30 – 16:30 Uhr	14.11.2016	8
QUICKSTARTING R: EINE ANWENDUNGSORIENTIERTE EINFÜHRUNG IN DAS STATISTIKPAKET R	Cordes	23.11. – 24.11.2016 9:00 – 12:00 und 13:00 – 15:30 Uhr	16.11.2016	8
UNIX FÜR FORTGESCHRITTENE	Dr. Sippel	28.11. – 30.11.2016 9:15 – 12:00 und 13:15 – 15:30 Uhr	21.11.2016	12
OUTLOOK – E-MAIL UND GROUPWARE	Helmvoigt	01.12.2016 9:15 – 12:00 und 13:00 – 16:00 Uhr	24.11.2016	4

KURS	VORTRAGENDE/R	TERMIN	ANMELDEN BIS	AE
HIGH-LEVEL, HIGH-PERFORMANCE TECHNICAL COMPUTING WITH JULIA	Chronz	06.12.2016 9:15 – 16:30 Uhr	29.11.2016	4
ANGEWANDTE STATISTIK MIT SPSS FÜR NUTZER MIT VORKENNTNISSEN	Cordes	07.12. – 08.12.2016 9:00 – 12:00 und 13:00 – 15:30 Uhr	30.11.2016	8
DIE SHAREPOINT-UMGEBUNG DER GWDC	Buck	15.12.2016 9:00 – 12:30 und 13:30 – 15:30 Uhr	08.12.2016	4
STATISTIK MIT R FÜR TEILNEHMER MIT VORKENNTNISSEN – VON DER ANALYSE ZUM BE- RICHT	Cordes	17.01. – 18.01.2017 9:00 – 12:00 und 13:00 – 15:30 Uhr	10.01.2017	8
EINFÜHRUNG IN SHAREPOINT 2013 FÜR ANWENDER	Buck	19.01.2017 9:00 – 12:30 Uhr	12.01.2017	2
EINFÜHRUNG IN SHAREPOINT 2013 FÜR ADMINISTRATOREN	Buck	19.01.2017 13:30 – 17:00 Uhr	12.01.2017	2
INDESIGN – GRUNDLAGEN	Töpfer	24.01. – 25.01.2017 9:30 – 16:00 Uhr	17.01.2017	8
ADMINISTRATION VON PCS IM ACTIVE DIRECTORY DER GWDC	Quentin	09.02.2017 9:00 – 12:30 und 13:30 – 15:30 Uhr	02.02.2017	4
EINFÜHRUNG IN SHAREPOINT 2013 FÜR ANWENDER	Buck	16.02.2017 9:00 – 12:30 Uhr	09.02.2017	2
EINFÜHRUNG IN SHAREPOINT 2013 FÜR ADMINISTRATOREN	Buck	16.02.2017 13:30 – 17:00 Uhr	09.02.2017	2
GRUNDLAGEN DER BILDBEARBEITUNG MIT PHOTOSHOP	Töpfer	21.02. – 22.02.2017 9:30 – 16:00 Uhr	14.02.2017	8
NETZLAUFWERKE FÜR DIE GEMEINSAME DATENABLAG	Quentin	02.03.2017 9:00 – 12:30 und 13:30 – 15:30 Uhr	23.02.2017	4
INDESIGN – GRUNDLAGEN	Töpfer	14.03. – 15.03.2017 9:30 – 16:00 Uhr	07.03.2017	8
EINFÜHRUNG IN SHAREPOINT 2013 FÜR ANWENDER	Buck	16.03.2017 9:00 – 12:30 Uhr	09.03.2017	2
EINFÜHRUNG IN SHAREPOINT 2013 FÜR ADMINISTRATOREN	Buck	16.03.2017 13:30 – 17:00 Uhr	09.03.2017	2
PHOTOSHOP FÜR FORTGESCHRITTENE	Töpfer	28.03. – 29.03.2017 9:30 – 16:00 Uhr	21.03.2017	8
USING THE GWDC SCIENTIFIC COMPUTE CLUSTER – AN INTRODUCTION	Dr. Boehme Ehlers	03.04.2017 9:30 – 16:00 Uhr	27.03.2017	4
PARALLELRECHNERPROGRAMMIERUNG MIT MPI	Prof. Haan	04.04. – 05.04.2017 9:15 – 17:00 Uhr	28.03.2017	8
EINFÜHRUNG IN SHAREPOINT 2013 FÜR ANWENDER	Buck	06.04.2017 9:00 – 12:30 Uhr	30.03.2017	2
EINFÜHRUNG IN SHAREPOINT 2013 FÜR ADMINISTRATOREN	Buck	06.04.2017 13:30 – 17:00 Uhr	30.03.2017	2

KURS	VORTRAGENDE/R	TERMIN	ANMELDEN BIS	AE
EINFÜHRUNG IN DIE STATISTISCHE DATENANALYSE MIT SPSS	Cordes	25.04. – 26.04.2017 9:00 – 12:00 und 13:00 – 15:30 Uhr	18.04.2017	8
EINFÜHRUNG IN SHAREPOINT 2013 FÜR ANWENDER	Buck	04.05.2017 9:00 – 12:30 Uhr	27.04.2017	2
EINFÜHRUNG IN SHAREPOINT 2013 FÜR ADMINISTRATOREN	Buck	04.05.2017 13:30 – 17:00 Uhr	27.04.2017	2
INDESIGN – AUFBAUKURS	Töpfer	09.05. – 10.05.2017 9:30 – 16:00 Uhr	02.05.2017	8
ADMINISTRATION VON PCS IM ACTIVE DIRECTORY DER GWDG	Quentin	11.05.2017 9:00 – 12:30 und 13:30 – 15:30 Uhr	04.05.2017	4
QUICKSTARTING R: EINE ANWENDUNGSORIENTIERTE EINFÜHRUNG IN DAS STATISTIKPAKET R	Cordes	16.05. – 17.05.2017 9:00 – 12:00 und 13:00 – 15:30 Uhr	09.05.2017	8
EINFÜHRUNG IN DAS IP-ADRESSMANAGEMENTSYSTEM DER GWDG FÜR NETZWERKBEAUFTRAGTE	Dr. Beck	30.05.2017 10:00 – 12:00 Uhr	23.05.2017	2
NETZLAUFWERKE FÜR DIE GEMEINSAME DATENABLAGE	Quentin	01.06.2017 9:00 – 12:30 und 13:30 – 15:30 Uhr	25.05.2017	4
EINFÜHRUNG IN SHAREPOINT 2013 FÜR ANWENDER	Buck	08.06.2017 9:00 – 12:30 Uhr	01.06.2017	2
EINFÜHRUNG IN SHAREPOINT 2013 FÜR ADMINISTRATOREN	Buck	08.06.2017 13:30 – 17:00 Uhr	01.06.2017	2
ANGEWANDTE STATISTIK MIT SPSS FÜR NUTZER MIT VORKENNTNISSEN	Cordes	13.06. – 14.06.2017 9:00 – 12:00 und 13:00 – 15:30 Uhr	06.06.2017	8
OUTLOOK – E-MAIL UND GROUPWARE	Helmvoigt	15.06.2017 9:15 – 12:00 und 13:00 – 16:00 Uhr	08.06.2017	4
STATISTIK MIT R FÜR TEILNEHMER MIT VORKENNTNISSEN – VON DER ANALYSE ZUM BERICHT	Cordes	20.06. – 21.06.2017 9:00 – 12:00 und 13:00 – 15:30 Uhr	13.06.2017	8
DATENSCHUTZ - VERARBEITUNG PERSONENBEZOGENER DATEN AUF DEN RECHENANLAGEN DER GWDG	Dr. Grieger	21.06.2017 9:00 – 12:00 Uhr	14.06.2017	2
EINFÜHRUNG IN SHAREPOINT 2013 FÜR ANWENDER	Buck	10.08.2017 9:00 – 12:30 Uhr	03.08.2017	2
GRUNDLAGEN DER BILDBEARBEITUNG MIT PHOTOSHOP	Töpfer	15.08. – 16.08.2017 9:30 – 16:00 Uhr	08.08.2017	8
ADMINISTRATION VON PCS IM ACTIVE DIRECTORY DER GWDG	Quentin	24.08.2017 9:00 – 12:30 und 13:30 – 15:30 Uhr	17.08.2017	4
EINFÜHRUNG IN SHAREPOINT 2013 FÜR ANWENDER	Buck	07.09.2017 9:00 – 12:30 Uhr	31.08.2017	2

KURS	VORTRAGENDE/R	TERMIN	ANMELDEN BIS	AE
EINFÜHRUNG IN SHAREPOINT 2013 FÜR ADMINISTRATOREN	Buck	07.09.2017 13:30 – 17:00 Uhr	31.08.2017	2
INDESIGN – GRUNDLAGEN	Töpfer	12.09. – 13.09.2017 9:30 – 16:00 Uhr	05.09.2017	8
NETZLAUFWERKE FÜR DIE GE- MEINSAME DATENABLAG	Quentin	21.09.2017 9:00 – 12:30 und 13:30 – 15:30 Uhr	14.09.2017	4
EINFÜHRUNG IN SHAREPOINT 2013 FÜR ANWENDER	Buck	19.10.2017 9:00 – 12:30 Uhr	12.10.2017	2
EINFÜHRUNG IN SHAREPOINT 2013 FÜR ADMINISTRATOREN	Buck	19.10.2017 13:30 – 17:00 Uhr	12.10.2017	2
PHOTOSHOP FÜR FORTGE- SCHRITTENE	Töpfer	24.10. – 25.10.2017 9:30 – 16:00 Uhr	17.10.2017	8
ADMINISTRATION VON PCS IM ACTIVE DIRECTORY DER GWDC	Quentin	02.11.2017 9:00 – 12:30 und 13:30 – 15:30 Uhr	26.10.2017	4
EINFÜHRUNG IN DIE STATIS- TISCHE DATENANALYSE MIT SPSS	Cordes	07.11. – 08.11.2017 9:00 – 12:00 und 13:00 – 15:30 Uhr	31.10.2017	8
INDESIGN – AUFBAUKURS	Töpfer	14.11. – 15.11.2017 9:30 – 16:00 Uhr	07.11.2017	8
EINFÜHRUNG IN SHAREPOINT 2013 FÜR ANWENDER	Buck	16.11.2017 9:00 – 12:30 Uhr	09.11.2017	2
EINFÜHRUNG IN SHAREPOINT 2013 FÜR ADMINISTRATOREN	Buck	16.11.2017 13:30 – 17:00 Uhr	09.11.2017	2
QUICKSTARTING R: EINE AN- WENDUNGSORIENTIERTE EIN- FÜHRUNG IN DAS STATISTIK- PAKET R	Cordes	21.11. – 22.11.2017 9:00 – 12:00 und 13:00 – 15:30 Uhr	14.11.2017	8
EINFÜHRUNG IN DAS IP-AD- RESSMANAGEMENTSYSTEM DER GWDC FÜR NETZWERK- BEAUFTRAGTE	Dr. Beck	28.11.2017 10:00 – 12:00 Uhr 13:30-15:30 Uhr	21.11.2017	2
NETZLAUFWERKE FÜR DIE GE- MEINSAME DATENABLAG	Quentin	30.11.2017 9:00 – 12:30 und 13:30 – 15:30 Uhr	23.11.2017	4
OUTLOOK – E-MAIL UND GROUPWARE	Helmvoigt	07.12.2017 9:15 – 12:00 und 13:00 – 16:00 Uhr	30.11.2017	4
ANGEWANDTE STATISTIK MIT SPSS FÜR NUTZER MIT VOR- KENNTNISSEN	Cordes	12.12. – 13.12.2017 9:00 – 12:00 und 13:00 – 15:30 Uhr	05.12.2017	8
EINFÜHRUNG IN SHAREPOINT 2013 FÜR ANWENDER	Buck	14.12.2017 9:00 – 12:30 Uhr	07.12.2017	2
EINFÜHRUNG IN SHAREPOINT 2013 FÜR ADMINISTRATOREN	Buck	14.12.2017 13:30 – 17:00 Uhr	07.12.2017	2

Teilnehmerkreis

Das Kursangebot der GWDG richtet sich an alle Mitarbeiterinnen und Mitarbeiter aus den Instituten der Universität Göttingen und der Max-Planck-Gesellschaft sowie aus einigen anderen wissenschaftlichen Einrichtungen.

Anmeldung

Anmeldungen können schriftlich per Brief oder per Fax unter der Nummer 0551 201-2150 an die GWDG, Postfach 2841, 37018 Göttingen oder per E-Mail an die Adresse support@gwdg.de erfolgen. Für die schriftliche Anmeldung steht unter <https://www.gwdg.de/antragsformulare> ein Formular zur Verfügung. Telefonische Anmeldungen können leider nicht angenommen werden.

Kosten bzw. Gebühren

Unsere Kurse werden wie die meisten anderen Leistungen der GWDG in Arbeitseinheiten (AE) vom jeweiligen Institutskontingent abgerechnet. Für die Institute der Universität Göttingen und der Max-Planck-Gesellschaft erfolgt keine Abrechnung in EUR.

Absage

Sie können bis zu acht Tagen vor Kursbeginn per E-Mail an support@gwdg.de oder telefonisch unter 0551 201-1523 absagen. Bei späteren Absagen werden allerdings die für die Kurse berechneten AE vom jeweiligen Institutskontingent abgebucht.

Kursorte

Alle Kurse finden im Kursraum oder Vortragsraum der GWDG statt. Die Wegbeschreibung zur GWDG sowie der Lageplan sind unter <https://www.gwdg.de/lageplan> zu finden.

Kurstermine

Die genauen Kurstermine und -zeiten sowie aktuelle kurzfristige Informationen zu den Kursen, insbesondere zu freien Plätzen, sind unter <https://www.gwdg.de/kursprogramm> zu finden.



Servervirtualisierung

Der einfache Weg zum Server!

Ihre Anforderung

Sie benötigen zur Bereitstellung eines Dienstes einen Applikations- oder Datenbankserver. Ihnen fehlen Platz, Hardware, Infrastruktur oder Manpower. Gleichzeitig soll der Server möglichst hochverfügbar und performant sein.

Unser Angebot

Wir bieten Ihnen die Möglichkeit des Hostings von virtuellen Servern für Ihre Anwendungen basierend auf VMware ESX. Sie können Ihre eigenen virtuellen Maschinen verwalten, die in unserer zuverlässigen Rechnerinfrastruktur gehostet werden, die unterschiedliche Verfügbarkeitsgrade unterstützen. Unsere Installation hält die Best-Practice-Richtlinien von VMware ESX ein. Sie bleiben Administrator Ihres eigenen virtuellen Servers, ohne sich mit der physikalischen Ausführungsumgebung beschäftigen zu müssen.

Ihre Vorteile

- > Leistungsfähiges VMware-Cluster mit zugehörigem Massenspeicher

- > Hohe Ausfallsicherheit und Verfügbarkeit durch redundante Standorte und Netzwerkverbindungen sowie USV-Absicherung
- > Bereitstellung aller gängigen Betriebssysteme zur Basisinstallation
- > Umfassender administrativer Zugang zu Ihrem Server im 24/7-Selfservice
- > Möglichkeit der automatisierten Sicherung des Servers auf unsere Backupsysteme
- > Zentrales Monitoring durch die GWDG
- > Große Flexibilität durch Virtualisierungstechnologien wie Templates, Cloning und Snapshots
- > Schutz vor Angriffen aus dem Internet durch leistungsfähige Firewallsysteme sowie ein Intrusion Prevention System

Interessiert?

Jeder Nutzer mit einem gültigen Account bei der GWDG kann das VMware-Cluster nutzen. Um einen virtuellen Server zu beantragen, nutzen Sie bitte die u. g. Webadresse.

>> www.gwdg.de/virtuelle-server



Gesellschaft für wissenschaftliche
Datenverarbeitung mbH Göttingen