20 Jahre Langzeitarchivierung am RZG

Hartmut Reuter
reuter@rzg.mpg.de

- Komponenten der Langzeitarchivierung
- Technologiewechsel
- Langzeit-Archive am RZG
- Problem mit TSM-HSM
- Empfehlungen

Langzeitarchivierung am RZG

- Keine Erfahrung mit
  - Portalen
  - Metadaten-Datenbanken
  - geeigneten Formaten für digitalisierte Dokumente
- RZG stellt nur ein einigermaßen sicheres Filesystem zur Verfügung
  - anwendungsunabhängig
  - beliebig erweiterbar (skalierend)
  - weltweit sichtbar
  - Technologiewechsel für Benutzer unsichtbar
- Benutzer sind selbst verantwortlich für
  - ihre Metadaten-Datenbanken
  - Konvertierung ihrer Daten auf neue Formate

Was ist Langzeitarchivierung?

- heißt länger archivieren als Lebensdauer der Komponenten
  - Archivierungs-Hardware (Band-, Plattensysteme) 5 Jahre
  - HSM-Systeme (Software, Hardware) 10 Jahre
  - Filesysteme mit HSM Funktionen 20 Jahre
  - Metadaten-Server (Datenbanken) 20 Jahre
  - Dokumentformate (.tif, .pdf ...) ? Jahre

- Ersetzen von Komponenten kann Jahre dauern (Umkopieren)
  - deshalb kurzlebige Komponenten möglichst verstecken:
    - Hardware unter HSM-System
    - HSM-System unter
      - Filesystem mit HSM Funktionalität oder
      - Portal mit Datenbank
  - Gelegentlich müssen leider auch diese ersetzt werden!

Komponenten der Langzeitarchivierung am RZG

- Archivierungs-Hardware (Bänder+Platten)
  - Lebenszyklus < 5 Jahre
    - Wartungskosten, Kapazitätsentwicklung
    - IBM 3480 -> D2 -> D3 -> STK 9840 -> STK 9940 -> STK 9940B-> LTO4

- HSM (Hierachical Storage Management) Systeme
  - erwartete Lebensdauer ~ 10 Jahre
    - In der Regel Plattform-abhängig
    - HADES -> DMF (Cray) -> DMF (SGI) -> TSM-HSM (IBM)

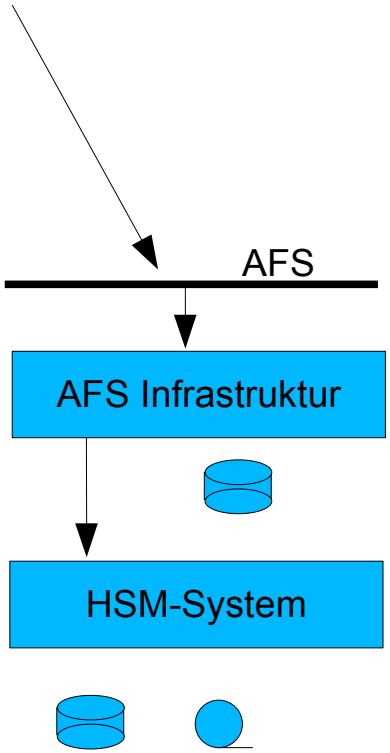
- Filesystem mit HSM-Funktionalität
  - erwartete Lebensdauer 15 – 20 Jahre
    - AMOS erreichte 18 Jahre, HADES 15 bzw. 21 Jahre
    - MR-AFS am RZG 1994-2008 (14 Jahre)
    - OpenAFS + Object Storage seit 2007

- Projekt-spezifische Metadaten Server (Datenbanken)
  - erwartete Lebensdauer > 20 Jahre
    - Oracle > 16 Jahre im Einsatz, bisher kein Technologiewechsel

Totaler Technologiewechsel 1995

- Altes System HADES auf IBM Mainframe Computern
  - EBCDIC Zeichensatz, Filesystem mit Recordstruktur
- Neues System MR-AFS auf UNIX (damals AIX, SUNOS und UNICOS)
  - ASCII Zeichensatz, Filesystem ohne Recordstruktur (bytestream)
- Benutzer mussten ihre Daten selbst transferieren!
- FTP-Interface von HADES ermöglichte Konversion “on the fly”
  - automatische Zeichensatz-Konversion für Text-Files
  - spezielle Konversion für W7AS-Schussfiles, möglich durch selbstbeschreibende Files mit Text, Integer- und Floatingpoint-Blöcken.

Warum AFS für Langzeitarchiverung

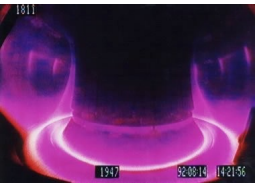


- AFS gibt es auf (fast) allen Plattformen (Linux, Unix, Windows, MacOs)
- Verteiltes weltweites Filesystem
- Ausreichende Sicherheit durch Kerberos Authentifizierung
- Granulare Vergabe von Rechten durch ACLs
- Volle Kontrolle über den Source-Code
  - OpenAFS ist open source
  - MR-AFS vom RZG gepflegt und weiterentwickelt
  - OpenAFS + Object Storage am RZG entwickelt
- Skaliert problemlos:
  - Hinzufügen von Servern im laufenden Betrieb
  - Umverteilung von Daten im laufenden Betrieb
- MR-AFS und AFS + Object Storage bieten HSM-Funktionalität
  - können beliebige HSM-Systeme benutzen
  - Transparenter Zugriff auf ausgelagerte Daten
- HSM-Technologiewechsel für Benutzer unsichtbar

Langzeitarchive für IPP Experimente

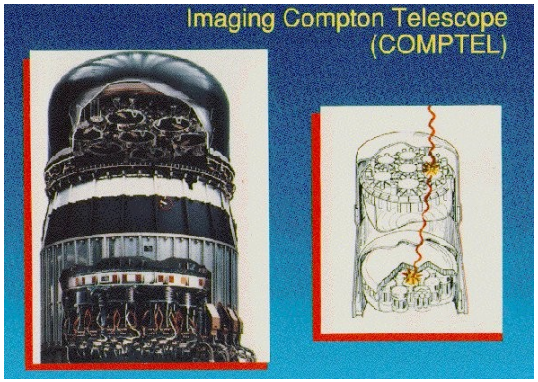
- Experiment Asdex des IPP (~1980-1990)
  - ca. 33500 Schussfiles kleiner 5 MB, insgesamt ~83 GB
    - von Bändern Ende der 90er in AFS kopiert

- Experiment W7AS des IPP (1988-2002)
  - ca. 60000 Schussfiles 1.5 bis 90 MB, insgesamt ~2.2 TB
    - 1988-1995 in VM/CMS auf IBM Mainframes erzeugt und in HADES gespeichert
    - 1995-2002 in VMS oder UNIX(AIX) erzeugt und in AFS gespeichert



- Experiment Asdex Upgrade am IPP (seit 1993)
  - ca. 20000 Schüsse mit jeweils vielen Diagnostikfiles, insges. ~100 TB
    - 1993-1995 in AMOS2 auf IBM Mainframe gespeichert
    - seit 1995 in AFS

Gamma-Ray-Astronomy MPE



- COMPTEL (1991-2000)
  - knapp 2 TB in container files von ~ 100 MB Größe
    - 1991-1995 in HADES gespeichert, dann -> AFS
    - 1995-2000 in AFS, jetzt alles auf Bändern
- EGRET Energetic Gamma Ray Experiment (1991-1996)
  - im AFS ~ 56 GB archiviert

- INTEGRAL (2001-2010 oder länger)
  - bisher ~8 TB in kleinen Files in AFS, alles auf Platten

Magic Projekt (MPP u.a.)

- Gamma-Strahlen-Teleskop auf Kanarischen Inseln
  - seit 2003 ~ 63 TB in AFS

Langzeitarchive geisteswissenschaftlicher Max-Planck-Institute

- Videos und Tondokumente des MPI für Psycholinguistik in Nimwegen
  - seit 2004, bisher ca. 7.8 TB
    - in Nimwegen direkt ins AFS kopiert
- Fototek der Biblioteca Hertziana in Rom
  - seit 2004, bisher ca. 2.5 TB
    - anfangs von verschickten USB-Platten am RZG ins AFS kopiert
    - jetzt in Rom direkt ins AFS
- Fototek des Kunsthistorischen Instituts Florenz
  - seit 2004, bisher ca. 4 TB
    - anfangs von verschicketn USB-Platten am RZG ins AFS kopiert
    - jetzt mit rsync von der GWDG direkt ins AFS

OpenAFS + Object Storage

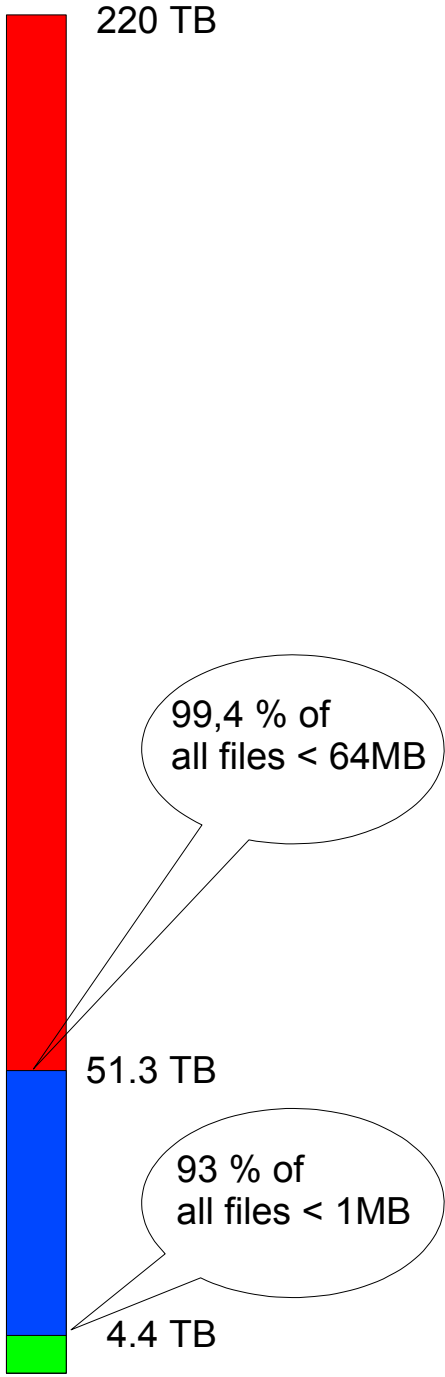
- Neuentwicklung als Ersatz für MR-AFS
  - global (weltweit sichtbar auf Unix, Linux, Windows u. MacOS)
  - verteilt (auf viele Server und OSDs)
  - hierarchisch (HSM)
  - sicher (Kerberos Authentifizierung)
  - skalierend (beliebig erweiterbar durch Hinzufügen neuer Server)
  - open source

- Benutzt für die Datamigration lokale HSM systeme
  - am RZG TSM-HSM

- Speziell für Langzeitarchivierung geeignet
  - speichert MD5 checksums für archivierte Files
  - Wechsel des unterliegenden HSM-Systems für Benutzer unsichtbar
  - mehrere Archivkopien möglich

Files im AFS des RZG

File Size Range	Files	%	run %	Data	%	run %
0 B - 4 KB	51797725	50.70	50.70	64.326 GB	0.03	0.03
4 KB - 8 KB	8166414	7.99	58.69	44.530 GB	0.02	0.05
8 KB - 16 KB	7353734	7.20	65.89	78.435 GB	0.03	0.08
16 KB - 32 KB	7794417	7.63	73.52	163.583 GB	0.07	0.16
32 KB - 64 KB	6593297	6.45	79.97	303.268 GB	0.13	0.29
64 KB - 128 KB	4546816	4.45	84.42	403.482 GB	0.18	0.47
128 KB - 256 KB	3316311	3.25	87.67	587.553 GB	0.26	0.73
256 KB - 512 KB	3468586	3.39	91.06	1.188 TB	0.54	1.27
512 KB - 1 MB	2460819	2.41	**93.47**	1.612 TB	0.73	**2.00**
1 MB - 2 MB	1631370	1.60	95.07	2.258 TB	1.03	3.03
2 MB - 4 MB	1535422	1.50	96.57	3.977 TB	1.81	4.84
4 MB - 8 MB	1328764	1.30	97.87	6.973 TB	3.17	8.01
8 MB - 16 MB	737305	0.72	98.59	7.914 TB	3.60	11.60
16 MB - 32 MB	508383	0.50	99.09	10.500 TB	4.77	16.37
32 MB - 64 MB	367269	0.36	**99.45**	15.246 TB	6.93	**23.30**
64 MB - 128 MB	241509	0.24	99.68	20.205 TB	9.18	32.48
128 MB - 256 MB	121515	0.12	99.80	20.996 TB	9.54	42.02
256 MB - 512 MB	77494	0.08	99.88	28.404 TB	12.91	54.93
512 MB - 1 GB	112096	0.11	99.99	76.463 TB	34.74	89.67
1 GB - 2 GB	8735	0.01	100.00	12.161 TB	5.53	95.19
2 GB - 4 GB	1705	0.00	100.00	4.399 TB	2.00	97.19
4 GB - 8 GB	374	0.00	100.00	2.297 TB	1.04	98.24
8 GB - 16 GB	121	0.00	100.00	1.236 TB	0.56	98.80
16 GB - 32 GB	57	0.00	100.00	1.183 TB	0.54	99.34
32 GB - 64 GB	27	0.00	100.00	1.159 TB	0.53	99.86
64 GB - 128 GB	4	0.00	100.00	308.307 GB	0.14	100.00
Totals:	102170269 Files			220.094 TB		



Probleme mit TSM-HSM und GPFS

- 08.02.2008: “reconcile”-Prozess beginnt Files aus TSM-Datenbank zu löschen.
- 12.02.2008: Reboot nach Software-Upgrade stoppt “reconcile”.
  - ca. 1.3 Mio Files sind bereits in der Datenbank gelöscht!
  - TSM hatte bereits 4 freigemachte Bänder mit neuen Files überschrieben!
- 14.02.2008: TSM-Datenbank wird auf Stand vom 8.2.2008 zurückgesetzt, nachdem alle seit dem 8.2. erzeugten Files wieder on-line gebracht wurden.
  - Die unter der anderen Datenbank on-line gebrachten neuen Files wurden dann allerdings von TSM kurzerhand wieder migriert ohne neue Bandkopie!
- Bis Mitte März: Aus restaurierter Datenbank vom 14.2. werden auf anderer Machine mit viel Handarbeit die zwischen dem 8. und 14. erzeugten Files restauriert und ins Originalsystem kopiert
- Bilanz:
  - 291 Files endgültig verloren.
  - Sehr viel zusätzliche Arbeit
  - Mehere Tage Betriebsunterbrechung
  - IBM kann Fehler nicht finden
  - Fehler wird im Zusammenspiel von GPFS und TSM-HSM vermutet

Empfehlung: REDUNDANZ

Je mehr Kopien eines File existieren, desto unwahrscheinlicher sein Totalverlust

- Am besten: möglichst mehrere unabhängige Kopien an verschiedenen Orten
  - Beispiel: KHI und MPL kopieren Files zur GWDG **und** zum RZG
- Am zweitbesten: Kopien innerhalb einer Organisation aber in unabhängigen Systemen
  - Nach dem TSM-HSM Problem Kopien in 2 unabhängigen TSM-HSM Systemen.
- Außerdem:
  - Nächtliche Replikation der AFS Volumes (Redundanz der Metadaten)
  - Dumps der Metadaten (damit man zur Not auch ohne AFS die Files im TSM-HSM wiederfinden könnte)

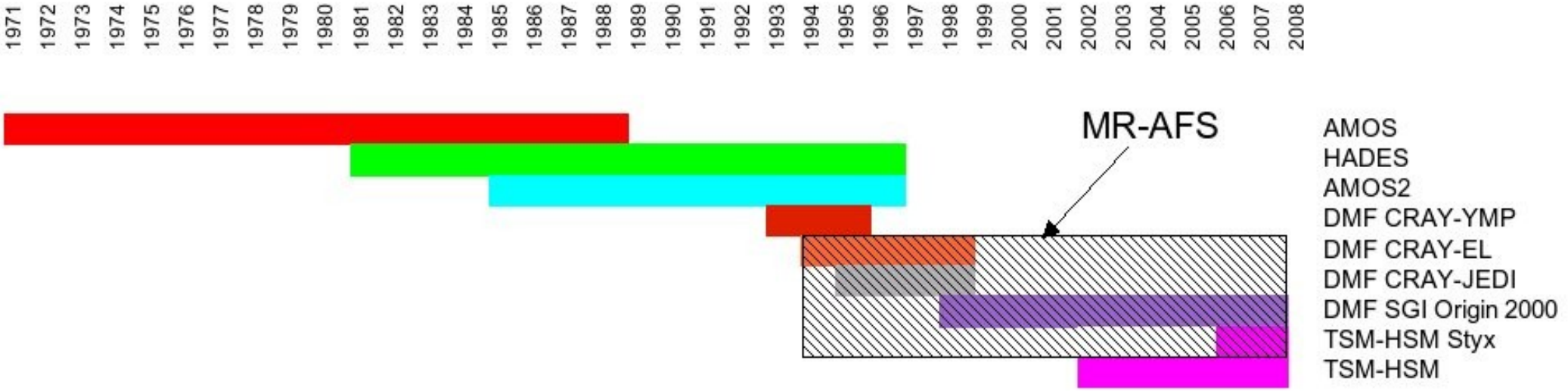
Empfehlungen (2)

- Kontrollsummen: (checksums) ermöglichen Verifizierung des Inhalts.
  - KHI liefert zu jedem File MD5-checksum in getrenntem File mit.
  - AFS + Object Storage speichert MD5 checksum in den Metadaten.

- Selbstbeschreibende Files: Wenn möglich,
  - innere Struktur der Files im File selbst beschreiben
  - dadurch auch automatische Versionskontrolle möglich.
  - Metadaten, die in der Datenbank stehen, auch im File selbst halten.

Danke für Ihre Aufmerksamkeit

HSM-Systeme am RZG



- AMOS
  - Eigenentwicklung des IPP: Betriebssystem für IBM Mainframe
  - Bänder und CDC–Massenspeicher (16 GB Gesamtkapazität.)
  - von 1971 – 1989 im Einsatz
- HADES
  - Weiterentwicklung von AMOS, OS unter VM oder Subsystem für MVS
  - Bandroboter, Netto–Datenvolumen ~ 1 TB
  - von 1981 – 1996 am RZG im Einsatz (In Heidelberg 20 Jahre)
- AMOS2
  - Neuentwicklung des IPP für Experiment ASDEX–Upgrade
  - von 1993 – 1995 im Einsatz
- MR–AFS (Multiple–Resident–AFS)
  - HSM–Erweiterung von AFS–Fileservern (nicht open–source)
  - seit 1994 am RZG im Einsatz
- OpenAFS+OSD
  - Erweiterung von OpenAFS, nutzt Object Storage Technik auch für HSM
  - Soll 2008 MR–AFS ersetzen