PAUL SCHERRER INSTITUT

PSI

Wir schaffen Wissen – heute für morgen

Paul Scherrer Institut
Edgar Barabas

SAM-FS/QFS Archiv am Paul Scherrer Institut

Paul Scherrer Institut

Das Paul Scherrer Institut PSI ist das grösste Forschungszentrum für Natur- und Ingenieurwissenschaften in der Schweiz.

Wir betreiben Spitzenforschung in den Bereichen Struktur der Materie, Mensch und Gesundheit sowie Energie und Umwelt.

Durch Grundlagen- und angewandte Forschung arbeiten wir an nachhaltigen Lösungen für zentrale Fragen aus Gesellschaft, Wirtschaft und Wissenschaft.

Übersicht

- Archiv Historie
- Archiv Migration 2010
- Auswahl neues System
- Migrations-Methode
- Anpassungen
- Archiv Hardware
- Zugriff
- Probleme
- Zukunft
- Statistik

Archiv Historie

1993	Einführung von UniTree V1.7
	Convex C210, Metrum RSS48 Library, 2 x VHS Tape Drives
1995	Migration auf Convex C220 und Unitree V2.0
	2 x ATL Produkts ACL 2640 Library, 6 x DLT 4000 Drives
1996	Migration auf Convex SPP 1600 und Unitree V3.0
	2 x ATL Produkts ACL 2640 Library, 6 x DLT 4000 Drives
1998	Umstellung auf DLT 7000 Drives
2002	Testinstallation Castor (CERN)
	Dell PE 2500, IBM 3584 Library mit 4 x LTO1 Drives
1/2003	Migration von UniTree auf DiskXtender/UNIX V2.5
	SUN Ultra Enterprise 405, 2 x ACL 2640 Library, 6 x DLT 7000 Drives
1/2003	Neuinstallation DiskXtender V2.5 auf Linux RH7.3
	Dell PE 2500, IBM 3584 Library mit 4 x LTO1 Drives
4/2003	Update auf DiskXtender V2.6 und IBM Library Erweiterung
	um einen weiteren Frame und 4 x LTO2 Drives
2004	Vorbereitung Daten-Migration DX/Sun nach DX/Linux
	SUN SPARC Big-Endian zu Intel X86 Little-Endian Format

Archiv Historie

10/2004 PSISPP Datenbank-Migration
in das neue DiskXtender Archivsystem

2/2005 Anbindung der DLT Libraries
an die neue Archiv Hardware (SCSI HVD)

2/2005 Start der DLT-Tape Datenmigration
mit Hlife der DXUL forcepack Funktion

2/2005 Upgrade von OS und DiskXtender
auf SL 304 und DXUL Version 2.9

1/2006 Hardware Upgrade
IBM eServer xSeries 345, LSI ProFibre 4000R FC 1,8 TB Cache, 6 x LTO2 Drives,
2 x Library Frame Expansion

9/2008 Hardware Upgrade
HP DL380G5, SUN STK 6140 FC 4,8 TB Cache, 4 x LTO2 + 4 x LTO3,
2 x Library Frame Expansion

Archiv Migration 2010

Es wird eine neue Archiv Lösung gebraucht

- Das EMC DiskXtender System entspricht nicht mehr den Erwartungen
- Die neuen High Resolution Detektoren am PSI stellen grössere Anforderungern.
  - Die >1 PB Grenze wurde Ende 2010 erreicht

Herausforderung:

- Migration von 780 TB alter LTO2/LTO3 Tapes auf neue LTO4 Tapes,
- minimale Beeinflussung des parallel laufenden produktiven Archiv-Betriebs

Archive Migration 2010: Auswahl des neuen Systems

EMC DiskXtender for Unix/Linux zu Oracle/Sun SAM-FS/QFS V4.6

Anforderungen

- gute Skalierbarkeit
- flexibler und sicherer Datenzugriff auf das Archiv File System
- support von ftps, nfs4, ...
- zumutbare Beeinflussung des Archiv-Services während der Migration der alten LTO2/3 DX Daten-Tapes auf LTO4

Auswahl: Oracle/Sun SAM-FS / QFS

- Gute Marktverbreitung (~3.000+ Installationen)
- Verfügbar für Solaris/x86 (nicht nur für SUN Sparc)
- Schreiben der Daten im 'tar'-Format, sichert langfristigen Datenzugriff, auch ohne SAMFS
- Anzahl der Files praktisch unbegrenzt
- Größe eines Files praktisch unbegrenzt (>> 1 TB)

Migrations Partner: *HMK Computer Technologies GmbH*, Kronberg - Germany

Langjährige Erfahrung, Toolsets für die Migration von DiskXtender zu SAM-FS / QFS

PSI Archive Migration: Method

SAM-HSM Migration Suite von HMK Computer Technologies

- Schnelle initiale Konvertierung der Meta-Daten
  - Production-ready des SAM-FS/QFS Filesystem nach einem Wochenende
- Alte, unmigrierte Tapes im nativen DX-Format können gelesen werden
  - Kein Parallelbetrieb von altem und neuem Archiv-System notwendig !
- Es stehen zwei Migrations-Methoden zur Auswahl
  - Fast *Bulk Migration* für ganze Tapes, als konstanter Hintergrundprozess
  - Direkter Zugriff mit anschliessender Migration von Files auf alten Tapes, getriggert durch einen Client Lesezugriff

Migration:

Start der Metadaten Migration	11. Jan. 2010
Start des produktiven Filesystems	13. Jan.
Start der Bulk Data Tape Migration	14. Jan. .
Ende der Tape Migration	09. Mai. 2010

→ Migration von bis zu 7TB DX-Tapedaten pro Tag

Anpassungen

Eine/zwei Kopien auf Tape

- Eine Kopie
  - gleichmässige Verteilung der Daten auf die Libraries West/Ost
  - Lösung:

    VSN Pool lto4_res mit reservierten Tapes , verteilt auf die Libraries West u. Ost
    Crontab Job, der den freien Speicher für Archive.set ‘all.1’ prüft. Wird 1 TB unterschritten, wird ein Tape in Lib. West und ein Tape in Lib. Ost freigegen

- Zwei Kopien
  - Erste Kopie in Library West
  - Zweite Kopie in Library Ost
  - Lösung

Archive Copy Definition über Pfad		/archiv/2tape/…
VSN Pools für Library West / Ost		lto4_w und lto4_e
VSN Association	1. Kopie	lto4_2cpy.1 li -pool lto4_w
	2. Kopie	lto4_2cpy.2 li -pool lto4_e

Anpassungen

VSN Pools

lto4_w	1. Kopie 2tape Pfad	Labelkreis 006000L4 bis 006999L4
lto4_e	2. Kopie 2tape Pfad	Labelkreis 007000L4 bis 007999L4
lto4_res	1. Kopie, verteilt West/Ost	Labelkreis 008000L4 bis 008999L4

Archive Hardware

Server SUN X4600M2 4*QuadCore CPU 2.7GHz, 16GB RAM
2x4Gb FC Dual Port HBA
OS: Solaris 10

Storage SUN STK6140 FC, 2xFC-Contr., 2x1GB Cache, 48x300GB FC-AL HDD 15k
8.7 TB Diskcache,
1.6 TB Diskarchiv für kleine Files

Library zwei IBM TS3500 Tape Libraries (Type 3584)

pro Library	1 x LTO2 Tape Drives
	2 x LTO3 Tape Drives
	5 x LTO4 Tape Drives
	2876 Slots

Logical Libraries

SAMFS	6 x LTO4, 4000 Slots
Networker	2 x LTO2, 4 x LTO4, 440 Slots
Test	4 x LTO3, 436 Slots

Zugriff

Zugriff von intern

FTP/FTPS (auch FTP Secure oder FTP-SSL)

NFS Vesion4 von dedizierten Servern

Zugriff von aussen

FTP/FTPS über QFS-Client in der DMZ

Standard FTP von ausserhalb des PSI war nicht mehr vorgesehen

FTP-Server

proFTP Version 1.3.3c

FTPS Clients

Unix/Linux lftp, gftp (GUI)

Windows FileZilla

Clients mit FTPS support:

http://www.ford-hutchinson.com/~fh-1-pfh/ftps-ext.html

Probleme

- Tape Library
  - Probleme nach S54 High Density Frame Erweiterung und Umstellung der Library auf ALMS (Advanced Library Management System)

    `'Exchange Media'` SCSI Kommando wird in Verbindung mit HD-Frames nicht mehr unterstützt
    - Die Nutzung dieses Kommandos in SAMFS ist default $\Rightarrow$ Nichts geht mehr!
    - Workaround implementiert durch HMK belebt das System wieder

    - *SUN case zum Beheben des Problems über Monate offen!*
    - *Case wird ohne Lösung seitens SUN geschlossen!*
    - *Ein Lösungsvorschlage wurde bereits bei Case Eröffnung eingereicht*

  - Seit der HD Frame-Erweiterung und Umbau der Gripper-Einheit sind bereits 4 mal Tapes auf den Boden gefallen

Zukunft

- Upgrade auf SAM-FS / QFS auf Version 5.2
- System Error Facility Reporting (SEF)
  - Tape / Drive - Probleme via Scripts auswerten / melden
- Statistik
  - Auswertungen automatisieren
- Oracle / Sun
  - Wartungsvertrag
  - Lizensierungsmodell für SAM-FS

Statistik

Class Report:

	ClassRange	Total Files	%	Total Size	%	Average Filesize	Total Online	%	Total Locked	%
(1):	0K - 4K	675288	13	860.40 M	0	1.30 K	15.12 G	0	0.00	0
(2):	4K - 8K	185792	4	1.01 G	0	5.71 K	2.25 G	0	0.00	0
(3):	8K - 16K	168109	3	1.82 G	0	11.35 K	2.51 G	0	0.00	0
(4):	16K - 32K	136094	3	2.97 G	0	22.90 K	2.93 G	0	0.00	0
(5):	32K - 64K	227747	5	10.50 G	0	48.35 K	3.34 G	0	0.00	0
(6):	64K - 128K	152656	3	12.61 G	0	86.60 K	2.61 G	0	0.00	0
(7):	128K - 256K	71346	1	12.74 G	0	187.21 K	3.28 G	0	0.00	0
(8):	256K - 512K	145707	3	48.54 G	0	349.29 K	7.15 G	0	0.00	0
(9):	512K - 1M	160618	3	110.89 G	0	723.92 K	8.34 G	0	0.00	0
(10):	1M - 2M	156299	3	224.16 G	0	1.47 M	5.90 G	0	0.00	0
(11):	2M - 4M	739856	15	2.17 T	0	3.07 M	1.09 G	0	0.00	0
(12):	4M - 8M	611225	12	3.08 T	0	5.29 M	8.16 G	0	0.00	0
(13):	8M - 16M	384313	8	3.95 T	0	10.78 M	132.83 G	1	0.00	0
(14):	16M - 32M	314659	6	6.78 T	1	22.58 M	891.92 G	10	0.00	0
(15):	32M - 64M	157469	3	6.60 T	1	43.96 M	272.44 G	3	0.00	0
(16):	64M - 128M	163394	3	13.82 T	2	88.67 M	2.19 T	25	0.00	0
(17):	128M - 256M	119824	2	20.51 T	2	179.49 M	505.06 G	6	0.00	0
(18):	256M - 512M	109724	2	37.16 T	4	355.10 M	735.81 G	8	0.00	0
(19):	512M - 1G	105682	2	79.29 T	10	786.69 M	1.24 T	14	0.00	0
(20):	1G - 2G	172333	3	277.03 T	33	1.65 G	878.66 G	10	0.00	0
(21):	2G - 4G	85111	2	213.06 T	26	2.56 G	317.17 G	4	0.00	0
(22):	4G - 8G	8504	0	44.56 T	5	5.37 G	72.32 G	1	0.00	0
(23):	8G - 16G	4486	0	47.85 T	6	10.92 G	43.07 G	0	0.00	0
(24):	16G - 32G	1279	0	25.13 T	3	20.12 G	64.00 G	1	64.00 G	1
(25):	32G - 64G	539	0	22.10 T	3	41.99 G	0.00	0	0.00	0
(26):	64G - 128G	179	0	14.29 T	2	81.76 G	71.83 G	1	0.00	0
(27):	128G - 256G	31	0	5.22 T	1	172.34 G	0.00	0	0.00	0
(28):	256G - 512G	10	0	3.42 T	0	350.08 G	0.00	0	0.00	0
(29):	512G - 1T	3	0	2.02 T	0	690.92 G	0.00	0	0.00	0
	Totals	5058277		828.45 T		171.74 M	7.38 T	85	64.00 G	1

Statistik

```
Total Files:          5058277
Total Zero Files:       25858
Total Dirs/Symlnk:     487331
Total size:            828.45 T
Average filesize:      171.74 M
Total size all copies:   1.11 P
Total online cap.:       8.72 T

Legend:

- Total Online is everything being online
- Total Locked is online, but cannot be released, because it is:
  + a directory, symbolic link
  + has the never archive or never release attribut set
  + belongs to a no_archive directive

The percentage for Online and Locked is relative to the available online capacity.
```

Timerange within	Creation time files	Creation time volume	Creation time %vol	Modify time files	Modify time volume	Modify time %vol	Acess time files	Acess time volume	Acess time %vol
1d	54	1.29T	0.2%	55	1.36T	0.2%	530	2.05T	0.2%
2d	189	5.54T	0.7%	188	5.48T	0.7%	529	6.14T	0.7%
1w	25762	1.90T	0.2%	25772	1.90T	0.2%	27788	5.42T	0.7%
1m	20718	3.46T	0.4%	20710	3.46T	0.4%	31368	20.36T	2.5%
2m	3611	5.52T	0.7%	3608	5.36T	0.6%	15551	15.28T	1.8%
3m	12044	3.83T	0.5%	11924	3.86T	0.5%	10419	5.53T	0.7%
1y	613103	189.65T	22.9%	519638	188.68T	22.8%	546114	242.62T	29.3%
2y	4382796	617.26T	74.5%	1060848	183.44T	22.1%	1153169	182.85T	22.1%
3y	0	0.00	0.0%	1032090	211.21T	25.5%	959012	132.77T	16.0%
older	0	0.00	0.0%	2383444	223.70T	27.0%	2313797	215.43T	26.0%
Total	5058277	828.45T	100.0%	5058277	828.45T	100.0%	5058277	828.45T	100.0%

PAUL SCHERRER INSTITUT

PSI

Danke für Ihre Aufmerksamkeit

Fragen?