Verteilte Dateisysteme - mit POSIX Semantik!

Andreas Hirczy

TU Graz
Institut für Theoretische Physik – Computational Physics

Grazer Linuxtage 2012
28. April. 2012

Vorstellung

Meine Installation umfasst das **Institut für Theoretische Physik – Computational Physics** an der TU Graz und einen **Computerlehr- und Arbeitsraum** für Mathematik- und Physik-Studenten.

Vorstellung

Meine Installation umfasst das **Institut für Theoretische Physik – Computational Physics** an der TU Graz und einen **Computerlehr- und Arbeitsraum** für Mathematik- und Physik-Studenten.

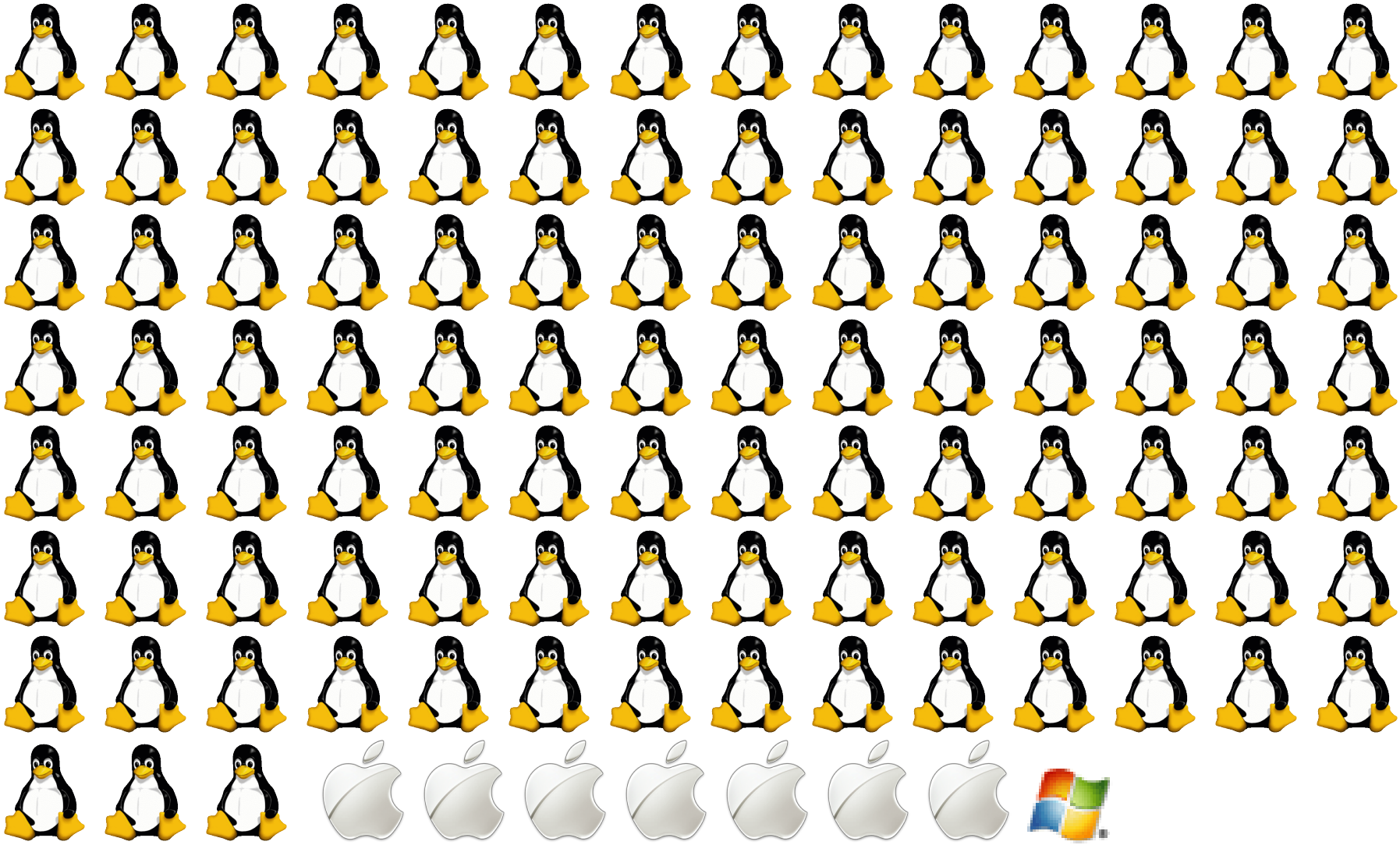
Der Arbeitsschwerpunkt liegt bei der mathematischen Behandlung von Fragestellungen vor allem im Bereich der **Vielteilchenphysik** und der **Plasmaphysik** – dabei kommen sowohl analytische als auch numerische Verfahren zum Einsatz.

Vorstellung

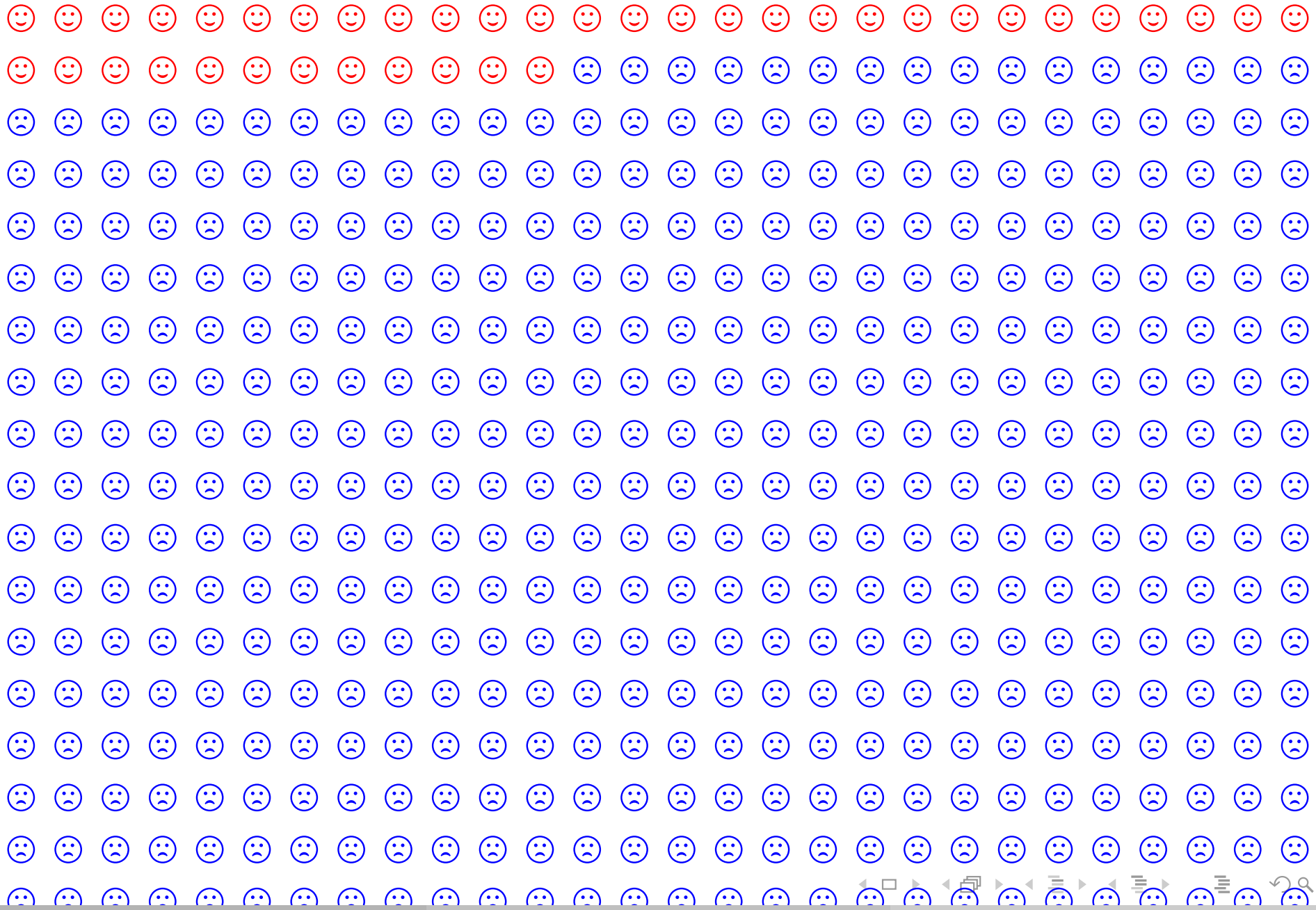
Meine Installation umfasst das **Institut für Theoretische Physik – Computational Physics** an der TU Graz und einen **Computerlehr- und Arbeitsraum** für Mathematik- und Physik-Studenten.

Der Arbeitsschwerpunkt liegt bei der mathematischen Behandlung von Fragestellungen vor allem im Bereich der **Vielteilchenphysik** und der **Plasmaphysik** – dabei kommen sowohl analytische als auch numerische Verfahren zum Einsatz.

Für Simulationsläufe verwenden wir das verteilte Batchsystem **Condor** - einige hundert gleichzeitig anlaufende Prozesse können die Fileserver ganz schön stressen.



≈ 600 Benutzer, davon im Schnitt etwa 40 gleichzeitig



Staff!

Just me! Hardware, purchase, operating systems, AFS, Kerberos, web services, mail, application software, hand-holding, programming ...

Netzwerkdateisysteme

- OpenAFS :-)
- NFS :-(

Warum was neues?

- **NFS!**

Warum was neues?

- **NFS!**
- Wir haben die Festplatten und sie werden nicht benutzt!

Warum was neues?

- **NFS!**
- Wir haben die Festplatten und sie werden nicht benutzt!
- **NFS** hat Durchsatzprobleme bei vielen Clients!

  Linux als Server – Solaris soll da besser sein

Warum was neues?

- **NFS!**
- Wir haben die Festplatten und sie werden nicht benutzt!
- **NFS** hat Durchsatzprobleme bei vielen Clients!

  Linux als Server – Solaris soll da besser sein
- **SPOF!**

Warum was neues?

- **NFS!**
- Wir haben die Festplatten und sie werden nicht benutzt!
- **NFS** hat Durchsatzprobleme bei vielen Clients!

  Linux als Server – Solaris soll da besser sein
- **SPOF!**
- Kein Bottleneck im Design!

Warum was neues?

- **NFS!**
- Wir haben die Festplatten und sie werden nicht benutzt!
- **NFS** hat Durchsatzprobleme bei vielen Clients!

  Linux als Server – Solaris soll da besser sein
- **SPOF!**
- Kein Bottleneck im Design!
- **NFS!**

Suche nach Alternativen - wir machen uns eine Daten-Cloud!

Was wollen wir eigentlich?

- POSIX - damit sich sie Anwender nicht umstellen müssen – Hadoop fällt raus!

Suche nach Alternativen - wir machen uns eine Daten-Cloud!

Was wollen wir eigentlich?

- POSIX - damit sich sie Anwender nicht umstellen müssen – Hadoop fällt raus!
- Diagnosemöglichkeiten

Suche nach Alternativen - wir machen uns eine Daten-Cloud!

Was wollen wir eigentlich?

- POSIX - damit sich sie Anwender nicht umstellen müssen – Hadoop fällt raus!
- Diagnosemöglichkeiten
- Selbstwartend - ich möchte auch Urlaub!

Suche nach Alternativen - wir machen uns eine Daten-Cloud!

Was wollen wir eigentlich?

- POSIX - damit sich sie Anwender nicht umstellen müssen – Hadoop fällt raus!
- Diagnosemöglichkeiten
- Selbstwartend - ich möchte auch Urlaub!
- einfach - AFS ist schon kompliziert genug

Suche nach Alternativen - wir machen uns eine Daten-Cloud!

Was wollen wir eigentlich?

- POSIX - damit sich sie Anwender nicht umstellen müssen – Hadoop fällt raus!
- Diagnosemöglichkeiten
- Selbstwartend - ich möchte auch Urlaub!
- einfach - AFS ist schon kompliziert genug
- **free software**

Suche nach Alternativen - wir machen uns eine Daten-Cloud!

Was wollen wir eigentlich?

- POSIX - damit sich sie Anwender nicht umstellen müssen – Hadoop fällt raus!
- Diagnosemöglichkeiten
- Selbstwartend - ich möchte auch Urlaub!
- einfach - AFS ist schon kompliziert genug
- **free software**
- vernünftige Datensicherheit: **Viele Server - viele Defekte!**

Suche nach Alternativen - wir machen uns eine Daten-Cloud!

Was wollen wir eigentlich?

- POSIX - damit sich sie Anwender nicht umstellen müssen – Hadoop fällt raus!
- Diagnosemöglichkeiten
- Selbstwartend - ich möchte auch Urlaub!
- einfach - AFS ist schon kompliziert genug
- **free software**
- vernünftige Datensicherheit: **Viele Server - viele Defekte!**
- nur lokaler Einsatz als *scratch*, muß **nicht** auf weltweiten Verteilung skalieren

Suche nach Alternativen - wer uns die Daten klaut!

Aus der Masse der vorhandenen Lösungen wurden 3 zur näheren Evaluierung ausgewählt:

- **GlusterFS**
- **Ceph**
- **MooseFS**

Suche nach Alternativen - wer uns die Daten klaut!

Aus der Masse der vorhandenen Lösungen wurden 3 zur näheren Evaluierung ausgewählt:

- **GlusterFS**
- **Ceph**
- **MooseFS**

Nach der Entscheidung für **MooseFS** bin ich noch auf **XtreemFS** gestoßen, das hätte meine Kriterien vermutlich auch erfüllt - wurde dann aber nicht mehr getestet, weil ...

Suche nach Alternativen - wer uns die Daten klaut!

Aus der Masse der vorhandenen Lösungen wurden 3 zur näheren Evaluierung ausgewählt:

- **GlusterFS**
- **Ceph**
- **MooseFS**

Nach der Entscheidung für **MooseFS** bin ich noch auf **XtreemFS** gestoßen, das hätte meine Kriterien vermutlich auch erfüllt - wurde dann aber nicht mehr getestet, weil ... MooseFS reicht!

`http://itp.tugraz.at/~ahi/admin/verteiltesDateisystem.html`

ERFAHRUNGEN?

GlusterFS

`http://www.gluster.org/`

$\approx$ 9 Monat im Probebetrieb (9 Serverknoten) – warum: es war so einfach zu installieren und ist sehr einfach durch durchschauen

- funktionierte gut im Regelbetrieb
- erzeugt gelegentlich etwas höhere Last auf den Speicherknoten
- **Aber:**

GlusterFS

`http://www.gluster.org/`

$\approx$ 9 Monat im Probebetrieb (9 Serverknoten) – warum: es war so einfach zu installieren und ist sehr einfach durch durchschauen

- funktionierte gut im Regelbetrieb
- erzeugt gelegentlich etwas höhere Last auf den Speicherknoten
- **Aber:** Redundanz wird manuell in die Konfiguration gebastelt. Wenn ein Knoten ausfällt, muß ich mich selbst darum kümmern.

http://ceph.newdream.net/

2 Wochen im Probebetrieb (6 Serverknoten)

- nicht so durchsichtig wie GlusterFS
- es gibt keinen *Bottleneck per Design* – die Clients rechnen sich via Hashfunktion aus, wo Dateien liegen!
- modularer Aufbau – Storage und POSIX-Layer getrennt – libvirt kann direkt aufs Storage zugreifen
- Snapshots!
- **Aber:**

http://ceph.newdream.net/

2 Wochen im Probebetrieb (6 Serverknoten)

- nicht so durchsichtig wie GlusterFS
- es gibt keinen *Bottleneck per Design* – die Clients rechnen sich via Hashfunktion aus, wo Dateien liegen!
- modularer Aufbau – Storage und POSIX-Layer getrennt – libvirt kann direkt aufs Storage zugreifen
- Snapshots!
- **Aber:** Man kann die Server (noch) sehr einfach gegen die Wand fahren lassen! Und die Doku war noch etwas dürftig.

Muß man sich merken!

`http://www.moosefs.org/roadmap.html`

Installation war einfach: Sourcen geholt – Debian-Pakete gebaut – installiert – konfiguriert

`http://www.moosefs.org/roadmap.html`

Installation war einfach: Sourcen geholt – Debian-Pakete gebaut – installiert – konfiguriert – Doku gelesen – konfiguriert – gelaufen

`http://www.moosefs.org/roadmap.html`

Installation war einfach: Sourcen geholt – Debian-Pakete gebaut – installiert – konfiguriert – Doku gelesen – konfiguriert – gelaufen

- nicht so durchsichtig wie GlusterFS
- Nachteil: MooseFS benötigt einen Metadaten-Server – der bildet zwar einen **SPOF** und einen **Flaschenhals**, ist aber bis jetzt nicht störend in Erscheinung getreten.
- Vorteil: Die Datenspeicher (*chunk servers*) balancieren sich selbst perfekt aus und sorgen automatisch für Redundanz, wenn ein Knoten ausfällt.

MooseFS!

Clients

Clients müssen FUSE unterstützen!

master server – Metadaten, Diagnose

MFS Moose File System

Info – | Servers – | Disks + | Exports + | Mounts + | Operations + | Master Charts + | Server Charts +

W3C css | W3C XHTML 1.0

Info

version	RAM used	total space	avail space	trash space	trash files	reserved space	reserved files	all fs objects	directories	files	chunks	all chunk copies	regular chunk copi
1.6.24	4.5 GiB	22 TiB	8.7 TiB	34 MiB	780	1.8 MiB	6	13381141	1856481	10641703	10174907	20410789	204107

All chunks state matrix (counts 'regular' hdd space and 'marked for removal' hdd space : switch to 'regular')

goal	valid copies											
	0	1	2	3	4	5	6	7	8	9	10+	all
0	3	297	-	-	-	-	-	-	-	-	-	300
1	-	-	-	-	-	-	-	-	-	-	-	0
2	-	-	10113329	-	-	-	-	-	-	-	-	10113329
3	-	-	-	61278	-	-	-	-	-	-	-	61278
4	-	-	-	-	-	-	-	-	-	-	-	0
5	-	-	-	-	-	-	-	-	-	-	-	0
6	-	-	-	-	-	-	-	-	-	-	-	0
7	-	-	-	-	-	-	-	-	-	-	-	0
8	-	-	-	-	-	-	-	-	-	-	-	0
9	-	-	-	-	-	-	-	-	-	-	-	0
10+	-	-	-	-	-	-	-	-	-	-	-	0
all 1+	0	0	10113329	61278	0	0	0	0	0	0	0	10174607

■ - missing (0) / ■ - endangered (0) / ■ - undergoal (0) / ■ - stable (10174607) / ■ - overgoal (0) / ■ - pending deletion (300)

Chunk operations info

loop time		deletions				replications	
start	end	invalid	unused	disk clean	over goal	under goal	rebalance
Sat Apr 28 12:56:28 2012	Sat Apr 28 13:01:31 2012	0/0	12/12	0/0	0/0	0/0	0

Filesystem check info

check loop start time	check loop end time	files	under-goal files	missing files	chunks	under-goal chunks	missing chunks
Sat Apr 28 09:01:44 2012	Sat Apr 28 13:02:21 2012	10641954	0	0	10174701	0	0

master server – Metadaten, Diagnose

Chunk Servers													
#	host	ip	port	version	'regular' hdd space				'marked for removal' hdd space				
					chunks	used	total	% used	chunks	used	total	% used	
1	faepop60.tu-graz.ac.at	129.27.161.5	9422	1.6.24	763120	495 GiB	829 GiB	59.71	0	0 B	0 B	-	
2	faepop65.tu-graz.ac.at	129.27.161.50	9422	1.6.24	745795	495 GiB	829 GiB	59.73	0	0 B	0 B	-	
3	faepop02.tu-graz.ac.at	129.27.161.91	9422	1.6.24	1530682	936 GiB	1.5 TiB	59.74	0	0 B	0 B	-	
4	faepop03.tu-graz.ac.at	129.27.161.96	9422	1.6.24	1469598	888 GiB	1.5 TiB	59.74	0	0 B	0 B	-	
5	faepop04.tu-graz.ac.at	129.27.161.97	9422	1.6.24	1203198	888 GiB	1.5 TiB	59.74	0	0 B	0 B	-	
6	faepop05.tu-graz.ac.at	129.27.161.98	9422	1.6.24	1431510	934 GiB	1.5 TiB	59.74	0	0 B	0 B	-	
7	faepop06.tu-graz.ac.at	129.27.161.99	9422	1.6.24	343829	213 GiB	357 GiB	59.72	0	0 B	0 B	-	
8	faepop07.tu-graz.ac.at	129.27.161.100	9422	1.6.24	317021	213 GiB	357 GiB	59.70	0	0 B	0 B	-	
9	faepop08.tu-graz.ac.at	129.27.161.101	9422	1.6.24	334186	213 GiB	357 GiB	59.72	0	0 B	0 B	-	
10	faepsv02.tu-graz.ac.at	129.27.161.107	9422	1.6.24	2604724	1.6 TiB	2.7 TiB	59.73	0	0 B	0 B	-	
11	faepop16.tu-graz.ac.at	129.27.161.125	9422	1.6.24	816653	520 GiB	870 GiB	59.73	0	0 B	0 B	-	
12	faepop18.tu-graz.ac.at	129.27.161.127	9422	1.6.24	1459236	949 GiB	1.6 TiB	59.73	0	0 B	0 B	-	
13	faepop20.tu-graz.ac.at	129.27.161.129	9422	1.6.24	1443491	949 GiB	1.6 TiB	59.73	0	0 B	0 B	-	
14	faepop35.tu-graz.ac.at	129.27.161.140	9422	1.6.24	332512	213 GiB	357 GiB	59.72	0	0 B	0 B	-	
15	faepop36.tu-graz.ac.at	129.27.161.141	9422	1.6.24	332264	213 GiB	356 GiB	59.73	0	0 B	0 B	-	
16	faepop37.tu-graz.ac.at	129.27.161.142	9422	1.6.24	298937	213 GiB	357 GiB	59.73	0	0 B	0 B	-	
17	faepop38.tu-graz.ac.at	129.27.161.143	9422	1.6.24	333634	213 GiB	357 GiB	59.74	0	0 B	0 B	-	
18	faepop39.tu-graz.ac.at	129.27.161.144	9422	1.6.24	333101	213 GiB	357 GiB	59.73	0	0 B	0 B	-	
19	faepop40.tu-graz.ac.at	129.27.161.146	9422	1.6.24	336221	213 GiB	357 GiB	59.74	0	0 B	0 B	-	
20	faepop41.tu-graz.ac.at	129.27.161.147	9422	1.6.24	333215	213 GiB	357 GiB	59.74	0	0 B	0 B	-	
21	faepop42.tu-graz.ac.at	129.27.161.148	9422	1.6.24	349674	213 GiB	357 GiB	59.73	0	0 B	0 B	-	
22	faepop43.tu-graz.ac.at	129.27.161.149	9422	1.6.24	336398	213 GiB	357 GiB	59.70	0	0 B	0 B	-	
23	faepop45.tu-graz.ac.at	129.27.161.151	9422	1.6.24	325419	213 GiB	357 GiB	59.72	0	0 B	0 B	-	
24	faepop46.tu-graz.ac.at	129.27.161.152	9422	1.6.24	333362	213 GiB	357 GiB	59.73	0	0 B	0 B	-	
25	faepop61.tu-graz.ac.at	129.27.161.169	9422	1.6.24	774668	495 GiB	829 GiB	59.74	0	0 B	0 B	-	
26	faepop62.tu-graz.ac.at	129.27.161.170	9422	1.6.24	773563	494 GiB	829 GiB	59.56	0	0 B	0 B	-	
27	faepop63.tu-graz.ac.at	129.27.161.171	9422	1.6.24	754778	495 GiB	829 GiB	59.74	0	0 B	0 B	-	

Metadata Backup Loggers			
#	host	ip	version
1	faepsv02.tu-graz.ac.at	129.27.161.107	1.6.24
2	faepsv05.tu-graz.ac.at	129.27.161.109	1.6.24
3	faepsv00.tu-graz.ac.at	129.27.161.138	1.6.24
4	faepsv09.tu-graz.ac.at	129.27.161.139	1.6.24

master server – Metadaten, Diagnose

Voraussetzungen:

- stabile Hardware – **SPOF**

master server – Metadaten, Diagnose

Voraussetzungen:

- stabile Hardware – **SPOF**
- ausreichend RAM – die Metadaten werden komplett im RAM gehalten. Bei $\approx 10 \times 10^6$ Dateien benötigt der Server bei uns dafür ≈ 4.5 GiByte RAM.

master server – Metadaten, Diagnose

Voraussetzungen:

- stabile Hardware – **SPOF**
- ausreichend RAM – die Metadaten werden komplett im RAM gehalten. Bei $\approx 10 \times 10^6$ Dateien benötigt der Server bei uns dafür ≈ 4.5 GiByte RAM.
- noch mehr RAM – wenn der *master server* seine Daten zu den *metaloggern* überträgt, wird der Bedarf kurzzeitig verdoppelt.

metalogger server

Die *metalogger server* erhalten Kopien der Metadaten - damit bei Ausfall des *masters* nichts verloren geht.

Empfehlenswert ist, daß diese Rechner die Bedingungen für einen *master server* erfüllen – dann können diese jederzeit die Rollen tauschen.

chunk server

Die *chunk server* sorgen für die Ablage der Daten auf ihren Festplatten. Außer ausreichend Kapazität und Bandbreite gibt es hier keine besonderen Anforderungen.

/etc/mfschunkserver.cfg

```
WORKING_USER = mfs
WORKING_GROUP = mfs
MASTER_HOST = fantasy
```

/etc/mfshdd.cfg

```
/srv/storage
*/srv/olddisk
*/media/disk_to_throw_away
```

Replikation

Das System sorgt selbständig für Replikation – dazu kann man ein *goal* einstellen, das angibt, wieviele Kopien eines Dateibaumes vorgehalten werden sollen.

```
$ mfsgetgoal -r /temp/ahi/
```

```
/temp/ahi/:
 files with goal        2 :     729485
 directories with goal  2 :     103867
```

Bei Ausfall eines Speicherknotens beginnt nach kurzer Zeit die Neuverteilung der vorhanden Kopien.

Snapshots

Snapshots können auch erzeugt werden, während ein Dateibaum in Verwendung ist (*copy on write*)

Konsole: mfsmakesnapshot source destination

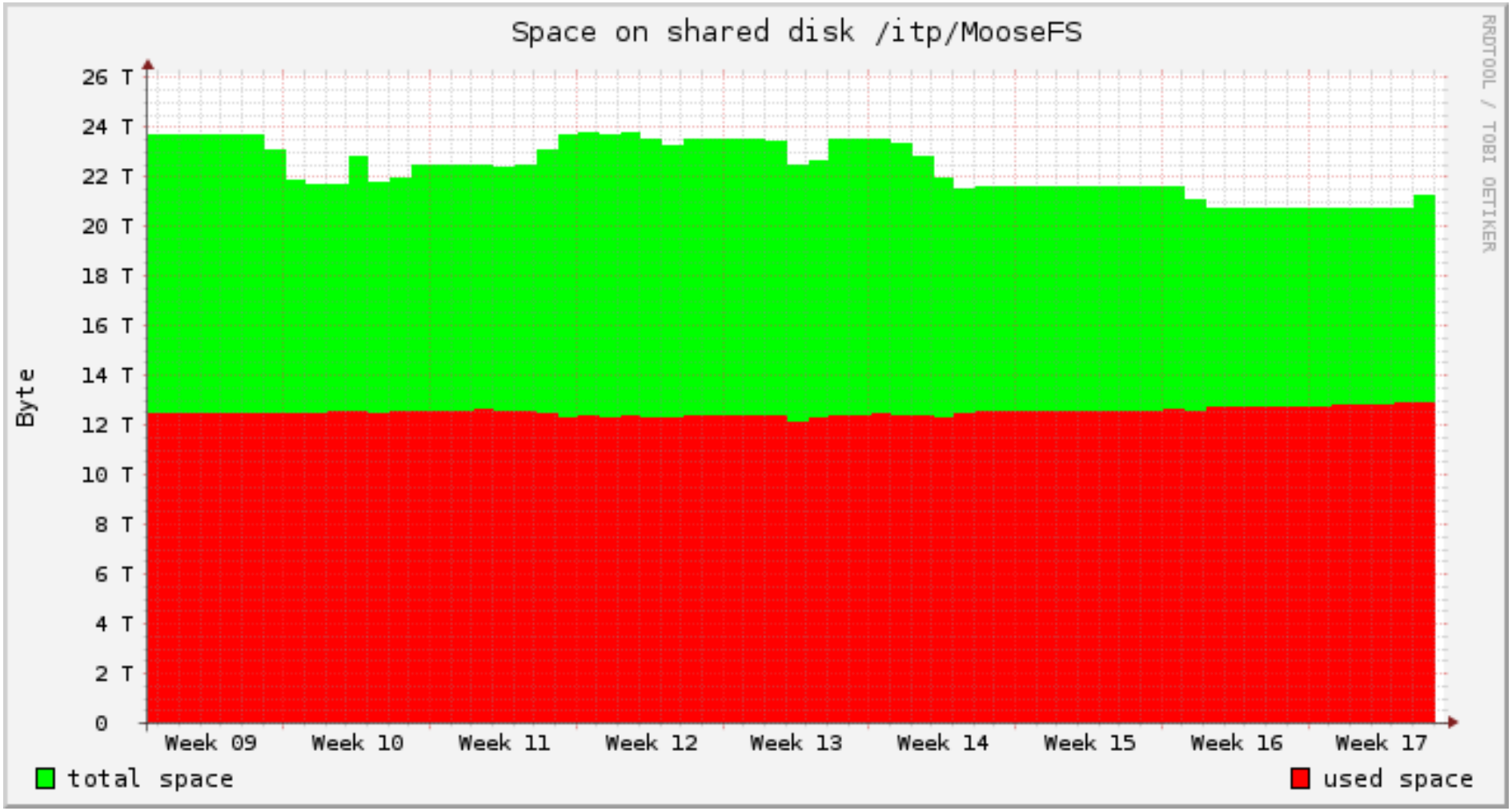
disk usage

Als schnelle Alternative zu *du* kann man *mfsdirinfo* verwenden:

$ mfsdirinfo -h /temp/ahi/

```
/temp/ahi/:
 inodes:         865Ki
  directories:   101Ki
  files:         712Ki
 chunks:         710Ki
 length:        873GiB
 size:          909GiB
 realsize:      1.8TiB
```

disk free


Space on shared disk /itp/MooseFS
RRDTOOL / TOBI OETIKER
Byte
26 T
24 T
22 T
20 T
18 T
16 T
14 T
12 T
10 T
8 T
6 T
4 T
2 T
0
Week 09
Week 10
Week 11
Week 12
Week 13
Week 14
Week 15
Week 16
Week 17
total space
used space

Wo gibst das?

Unterstützte Client-Plattformen sind wegen der Verfügbarkeit von FUSE:

- Linux
- FreeBSD
- OpenSolaris
- MacOS X

Wo gibst das?

Unterstützte Client-Plattformen sind wegen der Verfügbarkeit von FUSE:

- Linux
- FreeBSD
- OpenSolaris
- MacOS X

Als Server können zusätzlich noch

- Solaris
- MS Windows mit Cygwin

eingesetzt werden.

Debian

MooseFS ist noch nicht in Debian enthalten.

Die Sourcen enthalten aber ein Directory *debian/* mit allen notwendigen Daten um schnell selbst die Pakete zu bauen. Allerdings ist in der aktuellen Version 1.6.24 ein lästiger Fehler im init-Skript – funktioniert bei mir mit etwas cfengine-magic trotzdem.

Die Zukunft

Für die nächste Woche ist die Version 1.6.25 angekündigt, bei der das korrigierte Init-Skript und einige andere Verbesserungen einfließen sollen:

`http://www.moosefs.org/roadmap.html`

Danke für Ihre Aufmerksamkeit!

Andreas Hirczy
http://itp.tugraz.at/~ahi/
mailto:ahi@itp.tugraz.at

Diese Präsentationsunterlage finden Sie auf http://itp.tugraz.at/~ahi/VO/2012-04_GLT_mooseFS.pdf und demnächst auch auf http://linuxtage.at/.

Fragen?