science + computing

A Bull Group Company

RobinHood

Große Dateisysteme effizient verwalten

Daniel Kobras

science + computing ag

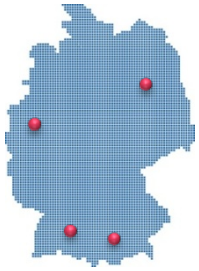
IT-Dienstleistungen und Software für anspruchsvolle Rechnernetze

Tübingen | München | Berlin | Düsseldorf

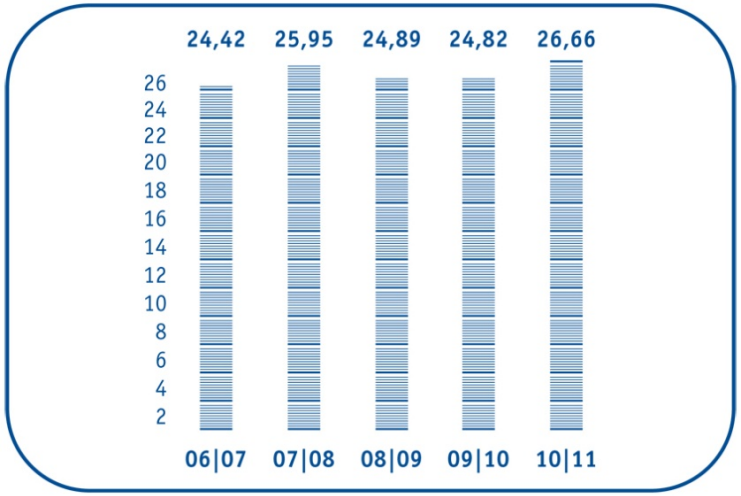
science+computing ag

science + computing
A Bull Group Company

Gegründet 1989
Standorte Tübingen
München
Berlin
Düsseldorf



Mitarbeiter 258
Anteilseigner Bull S.A. (100%)
Umsatz 10/11 26,66 Mio. Euro



Portfolio

IT Service für komplexe Berechnungsumgebungen
Umfassende Lösungen für Linux- und Windows-basiertes **HPC**

scVENUS Systemmanagement-Software zur effizienten Administration homogener und heterogener Netze

Übersicht

science + computing
A Bull Group Company

- Warum sollte ich mich für RobinHood interessieren?
- Wie kann ich RobinHood missbrauchen?
- Wie kann ich RobinHood gebrauchen?
- Was kann ich künftig von RobinHood erwarten?

science + computing

A Bull Group Company

Administration großer Dateisysteme

Dateisystem-Performance

science + computing
A Bull Group Company

- Optimiert für schnellen Datenzugriff

  open() → read()/write()

- Metadaten-Suche ineffizient

  readdir() → stat()

Dateisystem-Performance

- Optimiert für schnellen Datenzugriff

  open() → read()/write()

Anwender

- Metadaten-Suche ineffizient

  readdir() → stat()

  z.B. **`find`**, **`du -s`**, **`ls -lR`** über weite Teile eines Dateisystems

Administrator

Durchsatz vs. Latenz

science + computing
A Bull Group Company

- Metadaten-intensive Zugriffe führen zu **Random I/O**-Operationen
- Beschränkt durch **Latenz** des Speichersystems (IOPS)
- Ungleichgewicht Durchsatz vs. IOPS auf HDDs
  - typische SATA-Disk: 100MB/s, 100 IOPS
  - bei 4k-I/O-Größe:
    Random I/O ca. 250mal langsamer als Streaming I/O
- In großen Datenspeichern HDD nach wie vor Standard (Preis/Kapazität)
- Metadaten-Operationen dauern auf großen Dateisystemen im Verhältnis zur Gesamtgröße eines HDD-Datenträgers sehr lange

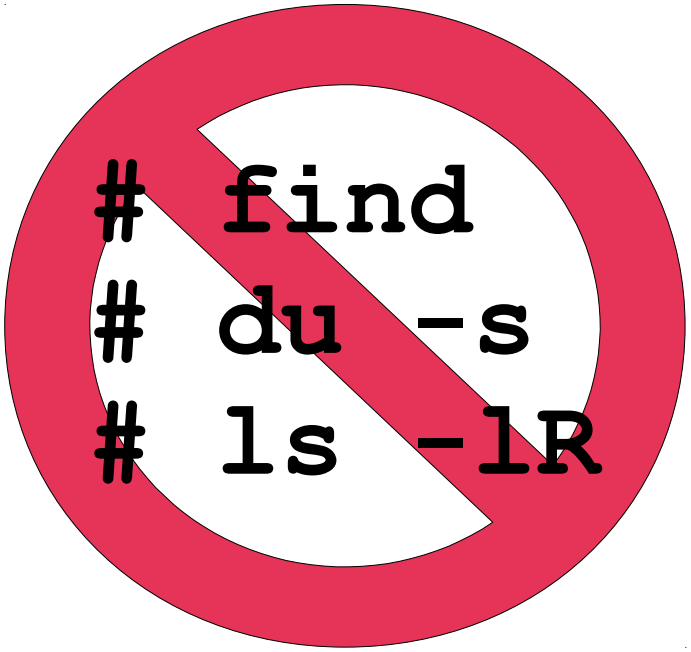
Schere im Kopf

science + computing
A Bull Group Company

```
# find
# du -s
# ls -lR
```

Schere im Kopf

science + computing
A Bull Group Company



Großes Dateisystem!

Schere im Kopf

science + computing
A Bull Group Company

- Administratoren haben es schwer, den Überblick über große Dateisysteme zu behalten
  - Wo liegen die größten Dateien?
  - Wie viele Einträge umfassen die größten Verzeichnisse?
  - Was ist durchschnittliche Dateigröße?
  - Welche Dateitypen werden am häufigsten verwendet?
  - Wie wird das Dateisystem genutzt?
- Informationen sind essentiell für Dateisystem-Anpassungen, Kapazitätsplanung, Accounting, Backup
- Informationsbeschaffung dauert zu lange
  - wird in nächtliche Skripte ausgelagert
  - wird unterlassen

SELECT statt find

science + computing
A Bull Group Company

- Idee: Stelle Metadaten eines Dateisystems als Datenbank zur Verfügung, die sich für beliebige Suchen optimieren lässt
- Ersetze Metadaten-Suchen im Dateisystem durch Datenbank-Abfragen

Datenbank-Betankung

science + computing
A Bull Group Company

- *Wunschtraum:* Datenbank ins Dateisystem integriert
  → schöne Idee, aber nicht der Status Quo

- *Auch gut:* Dateisystem bietet Schnittstelle, um alle Informationen über Änderungen an externe Programme weiterzureichen
  → ChangeLog-Feature in Lustre 2.x

- *Holzhammer:* Dateisystem regelmäßig scannen und dazwischen mit Inkonsistenzen leben
  → aufwändig, aber klappt immer

Kernfunktionalität von RobinHood

science + computing
A Bull Group Company

- *Wunschtraum*: Datenbank ins Dateisystem integriert
  → schöne Idee, aber nicht der Status Quo

- *Auch gut*: Dateisystem bietet Schnittstelle, um alle Informationen über Änderungen an externe Programme weiterzureichen
  → ChangeLog-Feature in Lustre 2.x

- *Holzhammer*: Dateisystem regelmäßig scannen und dazwischen mit Inkonsistenzen leben
  → aufwändig, aber klappt immer

- RobinHood beschafft über eine der beiden Methoden den Metadaten-Inhalt eines Dateisystems und repliziert ihn in eine Datenbank (derzeit: MySQL)

Vorteile

science + computing
A Bull Group Company

- Suche in Dateisystem-Metadaten dauert nur wenige **Sekunden statt Stunden** und Tagen
- Administrator kann Aufgaben **interaktiv** bearbeiten
- Datenbank-Abfragen befreien Dateisystem von **Metadaten-Last** durch Scan-Läufe (oder verlagern Scan-Läufe auf Bereiche außerhalb von Stoßzeiten)

RobinHood-Missbrauch

Arbeiten mit der RobinHood-DB

science + computing
A Bull Group Company

Alle Dateien eines Nutzers anzeigen

```
# find -uid <name>
```

wird zu

```
sql> SELECT fullpath FROM ENTRIES
     WHERE owner = <name>;
```

Arbeiten mit der RobinHood-DB

science + computing
A Bull Group Company

Alle Dateien anzeigen, die keinem aktuellen vorhandenen User-Account zugeordnet sind

```
sql> SELECT owner AS 'Owner',fullpath AS 'Path'
     FROM ENTRIES
     WHERE CONV(owner, 10, 10) = owner
     ORDER BY owner;
```

Arbeiten mit der RobinHood-DB

science + computing
A Bull Group Company

Schätze ab, wie viele Dateien im System bereits in einem komprimierten Format gespeichert sind

```
sql> SELECT COUNT(*) AS '#files',
     ROUND(SUM(size)/(1024*1024*1024),2)
        AS 'Size (GB)'
     FROM ENTRIES
     WHERE fullpath LIKE '%.gz' OR
           fullpath LIKE '%.bz2' OR
           fullpath LIKE '%.7z' OR
           fullpath LIKE '%.xz' OR
           fullpath LIKE '%.zip' OR
           fullpath LIKE '%.tgz';
```

Sonstige Ideen

science + computing
A Bull Group Company

- Projekt-Quotas „light“
- Nächtliche Anpassung von Dateisystem-Berechtigungen (Datenmanagement „light“)
- Überwachung der Dateisystembelegung (Kapazitätsmanagement „light“)
- Automatisches Komprimieren inaktiver Dateien
- Ersatz für beliebige Skripte vom Typ „find | dosomething“)
  → Einsatz dann sogar zum „Nulltarif“ (das heißt kein zusätzlicher Scan über das Dateisystem, sondern nur Ersatz eines bestehenden Scan-Skripts)

science + computing

A Bull Group Company

RobinHood-Gebrauch

RobinHood im Überblick

science + computing
A Bull Group Company

- Ursprünglich Software-Paket zur Verwaltung von temporären Daten in Dateisystemen (Tmpfs-Manager)
- Weitere Komponenten in Entwicklung:
  - effizientes Backup für Lustre 2.x
  - HSM-Policy-Engine für Lustre 2.y
- Entwicklung primär durch CEA
- Open-Source-Projekt (CeCILL-C-Lizenz, LGPL-Pendant)
- Projektseite: **`http://robinhood.sf.net/`**
- Binärpakete für RHEL5, RHEL6
- Git-Tree öffentlich zugänglich

Features Tmpfs-Manager

science + computing
A Bull Group Company

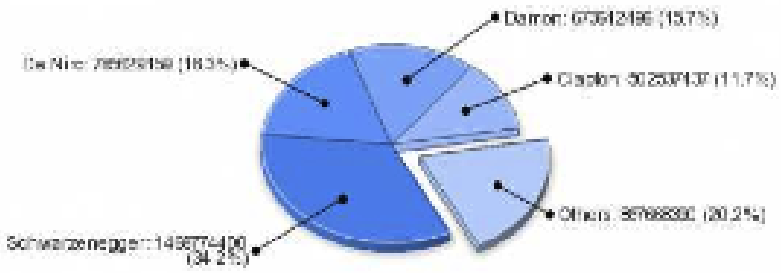
- Fokus: Verwaltung temporärer Daten auf großen Dateisystemen
- Policy-Engine zum automatisierten, regelbasierten Löschen von Daten
- Einfaches Klassifizieren von Daten
- E-Mail-Alerts
- Feste Kommandos für häufige Abfragen
- Web-GUI mit Dateisystem-Übersicht
- Effiziente Scan-Läufe auf Posix-Dateisystemen
- Nutzt ChangeLog-Feature auf Lustre-2.x-Dateisystemen
- Implementierung in C
- Keine Skripting-Möglichkeit für einfache Funktionserweiterungen

Web-Oberfläche

Users
Groups
Search

Volume | Count | List

Volume per user



User	Space used	Count
Schwarzenegger	1.37 GB	693 470
De Niro	749.23 MB	369 987
Damon	642.44 MB	318 368
Clapton	479.26 MB	156 565

View all users

Quelle: Thomas Leibovici/CEA

Kommando-Beispiele

```
# rbh-report --top-users

Rank:           1
User:           userX
Space used:     27.66 TB   (59400014112 blks)
Nb entries:     1382550
Size avg:       20.98 MB   (21996622bytes)

Rank:           2
User:           userY
Space used:     16.34 TB   (35089882808 blks)
Nb entries:     1673391
Size avg:       10.24 MB   (10736295bytes)
(...)
```

Kommando-Beispiele

science + computing
A Bull Group Company

```
# rbh-report --user-info
User:                kobras
    Type:            directory
    Count:           53
    Space used:      236.00 KB     (472 blks)

    Type:            file
    Count:           833
    Space used:      398.54 MB     (816200 blks)
    Size avg:        517.10 KB     (529512 bytes)

    Type:            symlink
    Count:           10
    Space used:      0             (0 blks)
    Size avg:        14            (14 bytes)
```

Kommando-Beispiele

```
# rbh-report --top-size

Rank:                1
Path:                /scr/foo/bar/xxx.big
Size:                476.11 GB (511223230464 bytes)
Last access:         2012/02/26 05:32:43
Last modification:   2010/11/09 09:18:43
Owner/Group:         myuser/mygroup
Purge class:         back_server2
Stripe count:        1
Stripe size:         1.00 MB (1048576 bytes)
Pool:                production
Storage units:       OST #21
(...)
```

Kommando-Beispiele

science + computing
A Bull Group Company

```
# rbh-report --class-info

Class:          back_server1
Nb entries:     1063201
Space used:     21.82 TB   (46863965240 blks)
Size min:       0          (0 bytes)
Size max:       327.25 GB  (351382807603 bytes)
Size avg:       21.52 MB   (22566027 bytes)

Class:          back_server2
Nb entries:     2039364
Space used:     25.71 TB   (55219685744 blks)
Size min:       0          (0 bytes)
Size max:       476.11 GB  (511223230464 bytes)
Size avg:       13.22 MB   (13864848 bytes)
```

Beispiele für E-Mail-Alerts

science + computing
A Bull Group Company

Warnung (an Admins) bei neuen Dateien >100GByte

/etc/robinhood.d/tmpfs/myfs.conf:

```
(...)
Alert recent_large_file {
   type == file
   and size > 100GB
   and last_mod < 1d
}
(...)
```

Beispiele für E-Mail-Alerts

Warnung (an Admins), falls eine Gruppe gr_diskhog zuviel Platz belegt („Quota light“)

`/etc/robinhood.d/tmpfs/myfs.conf`:

```
(...)
trigger_on = group_usage(gr_diskhog);
high_threshold_vol = 15TB;
notify = TRUE;
(...)
```

Auch möglich:
„Hardcore-Quota“: Lösche Daten der Gruppe, falls sie zuviel Platz belegt (gemäß vorgegebener Purge-Policy)

science + computing
A Bull Group Company

RobinHood in Zukunft

Geplante Weiterentwicklungen

science + computing
A Bull Group Company

- Kommandos **`rbh-find`**, **`rbh-du`** als effizienter Ersatz für **`find`** und **`du`**
- E-Mail-Warnungen direkt zum Anwender verschicken
- Automatisches Ausbalancieren von Lustre-OSTs
- Ausführen beliebiger, benutzerdefinierter Aktionen anhand von Policies (Komprimieren, Verschieben, ...)
- Höhere Skalierbarkeit durch NOSQL-Datenbank

Zusammenfassung

science + computing
A Bull Group Company

- RobinHood spiegelt Metadaten eines Dateisystems in SQL-Datenbank
- Ersetzt langwierige Kommandos durch effiziente Datenbank-Abfragen
- Entlastet Dateisysteme
- Ermöglicht schnelleres Arbeiten für Administratoren
- **Simple Idee bringt spürbare Erleichterungen im Administrationsalltag**

science + computing

A Bull Group Company

Vielen Dank für Ihre Aufmerksamkeit.

Daniel Kobras
science + computing ag
www.science-computing.de

Telefon 07071 9457-0
info@science-computing.de