# Keep inventory of your data on different file systems with ISH

Ramses van Zon

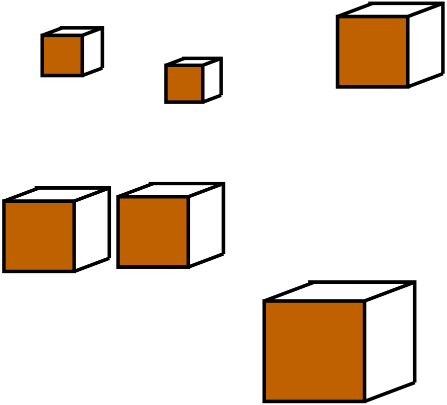
SciNet HPC Consortium

April 9, 2014

# Do you know where your data is?

- SciNet $HOME file system
- SciNet $SCRATCH file system
- SciNet HPSS system
- Your lab computer
- Your laptop
- . . .

# Inventory Hell



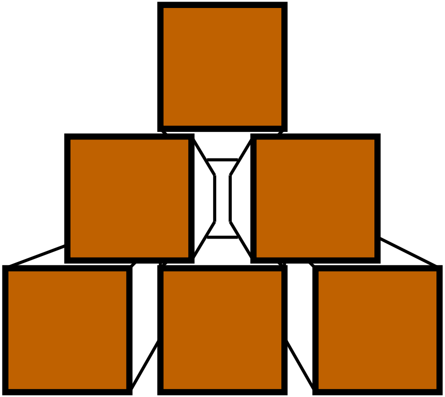
- Files and data stored all over the place
- In tar balls, directories, . . . ?
- Want to know what's where, how big it is, when it was changed, . . . , without having to log in?
- Can get cumbersome:

```
$ ls -R > list1.txt

$ tar -ztvf tarball.tgz > list2.txt

$ grep helloworld.c list1.txt list2.txt
```

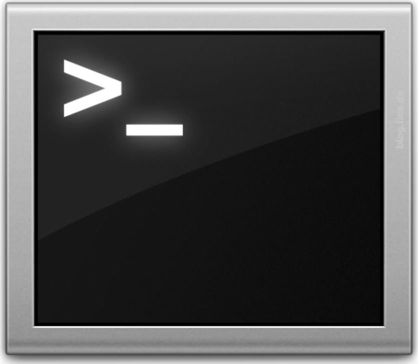
# Inventory SHell



- **ish:** A little shell that allows you to browse through the content of a tar file.
- Saves ‘table of content’ of a tar into an **index file**.
- It can do the same with file metadata in directories (filenames, sizes, . . . ).
- If you copy the index files, you can browse this tar or directory anywhere with ish.

# Why a Shell?

- ish = Inventory **shell**
- Presents a prompt to the user to browse data
- Use unix-like commands:
  - ls
  - cd
  - pwd
  - find
  - du
- Nice for tar balls, for which there is no ‘shell’ environment.
- Text-based interface means you can script it, too.



Caveats:

- Not a full-fledged linux shell.
- Must run under bash (Linux, Mac, Cygwin)

# Getting ISH

**On your machine:**

```
$ git clone git://github.com/vanzonr/ish
$ chmod +x ish/ish
$ ish/ish
[ish]>
```

**On SciNet:**

```
$ module load extras
$ ish
[ish]>
```

We'll denote the bash prompt with $ and the ish prompt with `[ish]>` for the rest of this talk.

# Usage in a Nutshell

- Find the directory or tarball to index.

  ```
  $ cd THIS-DIRECTORY
  ```

- Index it with ish.

  ```
  $ ish index data.tgz
  ```

- Copy the index file over to anywhere you'd want to browse it from.

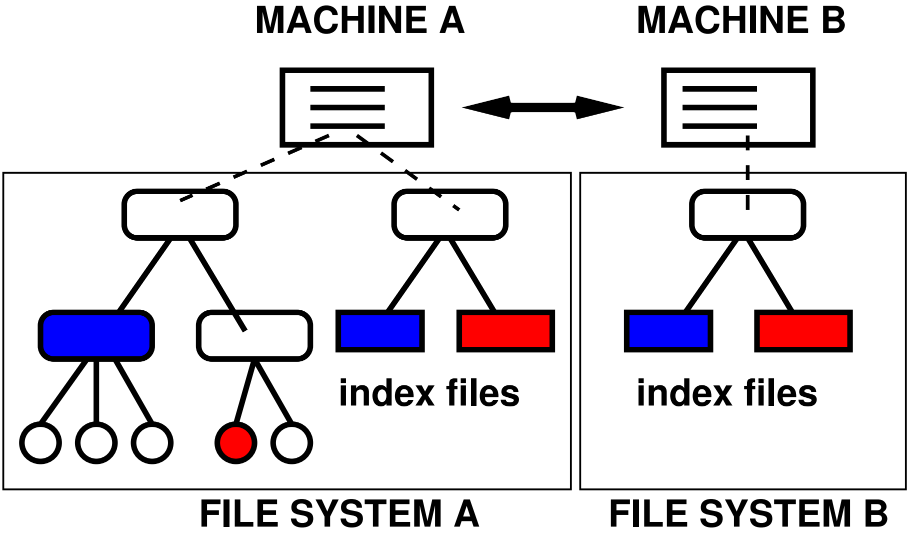
  ```
  $ cd ~/.ish_register
  $ scp data.tgz.igz otherhost:
  ```

- Browse it with ish on otherhost.

# Browse it with ish

```
$ ssh otherhost
otherhost$ ish data.tgz.igz
[ish]> ls -l
-rw-r--r-- rzon/users 1865 2014-04-08 13:03 boxes.fig.bak
-rw-r--r-- rzon/users 1871 2014-04-08 13:22 boxes.fig
-rw-r--r-- rzon/users 2803 2014-04-08 13:22 boxes.pdf
-rw-r--r-- rzon/users 2495 2014-04-08 13:03 boxes.png
-rw-r--r-- rzon/users 2319 2014-04-08 13:20 boxmess.fig.bak
-rw-r--r-- rzon/users 2175 2014-04-08 13:21 boxmess.fig
-rw-r--r-- rzon/users 2938 2014-04-08 13:21 boxmess.pdf
[ish]> exit
otherhost$
```

# Where's everything?



# DETAILS

- Commands
- Common use cases

# Indexing

```
[ish]> index DIRECTORY
[ish]> index TARBALL
```

- Creates index of the DIRECTORY or TARBALL (.tar,.tgz,. . . ). containing filenames, dates, sizes, ownership, permissions.
- The index is put in a file in the directory `$HOME/.ish_register`. (can be changed by setting `$ISHREGISTER`)
- Index files have the extension `.igz`
- Index name is TARBALL.igz or ABSPATH.igz, where ABSPATH is the absolute path to DIRECTORY, with slashes replaced by underscores.
- This new index becomes the 'active one'.

# Listing

```
[ish]> use INDEXFILE
[ish]> ls
[ish]> ls -l
[ish]> ls -lr
[ish]> cd DIR
```

- First commands selects index file from `$HOME/.ish_register` (Only one index file can be browsed at the same time.)
- Second command lists the content in the root directory within the INDEXFILE.
- `-l` : long listing (dates, sizes, owner, etc)
- `-r` : recursive listing (all files in all subdirectories)
- Last command changes directory in the INDEXFILE.

# Finding and accounting

```
[ish]> ls ? *.tex
[ish]> find ? *.tex
[ish]> du
[ish]> du -r
```

- List all files with single-character names and all files ending in .tex in this directory
- Finds all files with single-character names and all files ending in .tex in this directory and its subdirectories
- `du`: Count number of files and kilobytes in current directory
- `du -r`: Count number of files and kilobytes in current directory and subdirectories

# More commands

| | | |
|---|---|---|
| avail | [-a] | list (all) available index files |
| colour | 1\|0 | set colour usage |
| help | [COMMAND] | show help on (all) commands |
| register | [DIR] | set new index file location |
| use | [INDEX] | use INDEX or list available ones |
| unuse | | use previous index file again |
| info | | show properties of index file |
| pwd | | show current directory |
| settings | | show settings (colour, etc.) |
| tar -[z]cf | TARFILE DIR[/FILES] | tar and make index |
| check [-n] | [COMMENT] | exit ish if error in command |
| !COMMAND | [ARGS] | local commands (ls, cd, pwd) |

# Single command mode

- You can give the index list to use as an argument.
  E.g.

  ```
  $ ish data.tgz.igz
  [ish]data.tgz.igz>
  ```

- You can additionally give a single command as an argument.
  Ish will run the command and exit.
  E.g.

  ```
  $ ish data.tgz.igz find mylonglosttar.*
  datadir/mylonglosttar.tgz
  $
  ```

# HPSS application

- Every group at SciNet can have up to 2TB on the High Performance Storage System (HPSS).
- For HPSS, even file listings have to be obtained through hsi.
- ish interfaces with hsi and htar and can make indices:

```
[ish]> hindex
[ish]> hindex DIRECTORY
[ish]> hindex TARBALL
[ish]> htar zcf TARBALL *.nc
```

- These commands take $ARCHIVE as the root of relative HPSS directories.

# HPSS application: indexing

```
[ish]> hindex
[ish]> hindex DIRECTORY
[ish]> hindex TARBALL
```

- Note: This will only work in an hpss session (needs hsi/htar):

  ```
  $ qsub -q archive -I
  ```

  or

  ```
  $ qsub gethindex.pbs
  ```

- The first form indexes your $ARCHIVE into the index file to ‘hpss.igz’
- When starting ‘ish’ without parameters, it will load ‘hpss.igz’ by default.

```
#!/bin/bash
#PBS -l walltime=1:00:00
#PBS -q archive
#This is gethindex.pbs
module load extras
ish hindex
```

# HPSS application: tar-ing

```
[ish]> htar zvf TARBALL tarthis
```

is equivalent to

```
$ htar zvf TARBALL tarthis
$ ish hindex TARBALL
```

- This too will only work in an hpss session (needs htar):
  `$ qsub -q archive -I`
  or
  `$ qsub dohtar.pbs`

- TARBALL will live in your $ARCHIVE on hpps
- tarthis lives on $HOME or $SCRATCH
- Creates index `TARBALL.igz`

```
#!/bin/bash
#PBS -l walltime=1:00:00
#PBS -q archive
#This is dohtar.pbs
module load extras
ish htar TARBALL tarthis
```

# Scratch purging application

- As a special SciNet feature, one can ask for the index of the monthly scratch purging list.

```
$ ish
[ish]> pindex
[ish]> cd /scratch/s/scinet/rzon
[ish]> ls
```

- Must of course be on SciNet, and must have something to purge.
- If you have a lot of files, the index building can be very slow.
- The 'root' of the index is /, so you'll have to `cd` a bit to get to your files.

# Thank you for your attention!

**Links:**

- Source code: https://github.com/vanzonr/ish
- Documentation: http://wiki.scinethpc.ca/wiki/index.php/ISH (or use ish's help system).
- On HPSS: http://wiki.scinethpc.ca/wiki/index.php/HPSS

**Questions?**

**Feature requests?**