

Hierarchical Storage Interface (HSI)

<http://www.sdsc.edu/Storage/hsi>

<http://www.sdsc.edu/Storage/hsi/Install>

Michael Gleicher
Gleicher Enterprises, LLC



HSI - Topics

- Version History
- Current Version (2.8)
 - New commands/options
 - New extension APIs
- HSIRC Overview
- Auto-COS Overview
- Autoscheduling Overview



Special Thanks to

- ORNL (Funding) PROBE/
SciDAC/SDM
 - Randy Burris
 - Dan Million
 - Nina Hathaway
 - Stan White
- NERSC – PROBE
 - Nancy Meyer
 - Damian Hazen
- Indiana University
 - Haichuan Yang
- LLNL
 - George Richmond
 - Vicky Renbarger
 - Jerry Shoopman
 - Neil O’Neill
 - Jeff Long
- SDSC
 - Kate Ericson

HSI Recent Versions – 2.6.X

- multi-HPSS

- *Logical drive* command syntax

- ```
get local : [drive:]HPSS
```

- I/O improvements – buffer pools, extended I/O calls (hpss\_ReadX, hpss\_WriteX)

- hsirc enhancements



# HSI Recent Versions – 2.7.x

## - **Commands**

lsjunctions - list HPSS junctions  
lsfilesets - list filesets

## - **Command line changes**

*<comment>* - removed (rarely, if ever used)  
*# comment* - added

"here document" syntax (implemented in "chcos", "get" and "stage" commands; extend to remaining cmds in future)

## - **Features**

- Partial file xfers
- *<reget/reput>* (-t) for *get* / *put* commands
- Use server-side copy for *cp* command if same subsystem
- Use HPSS.conf for network options, kerberos realm->cell mapping
- Auto-scheduling

## - **Misc**

- directory list speedup, NEC SX5 support, no-stage/cache purge options for *get*, other



# HSI 2.7 (cont.)

- Support removed for pre-HPSS 4.1.1.4 (mover protocol address changed)



# HSI 2.7 (cont.)

## New NDAPI library functions (HPSS Client API)

- `hpss_CopyFile`
- `hpss_StageCallBack`, `hpss_GetAsyncStatus`

## Fileset calls

(`hpss_FilesetCreate/Delete/GetAttributes/ListAll/SetAttributes`)

## Acct calls

(`hpss_AcctCodeToName`, `hpss_AcctNameToCode`)

# HSI 2.7 (cont.) – Extension Library

- HPSS.conf parser

  - APIs

    - `hpss_Cfg_Parse, hpss_Cfg_Free, hpss_Cfg_FindKey, hpss_Cfg_Free`

- Internally change NDAPI/Extension library to use config parser, HPSS.conf network options

- Globus GSI authentication method

- Autoscheduling APIs



# ndapi library

## multi-hpss-version support

- “minor” api structure or definition changes = big translation effort
- 4.0 – 4.1.1.3
- 4.1.1.4
- 4.2
- 4.3
- 4.5

Example:

```
ns_Attr_t
ns_DirEntry_t
hpss_fileattr_t,. hpss_xfileattr_t
```

```
hpss_xfileattr_t
bf_xattrib_t
 bf_sc_attrib_t
 bf_vv_attrib_t
```

# HSIRC

Purpose – Configuration options

Residence:

- HPSS (Global)
- Host Global
- User private .hsirc



# HSIRC – configuration file

- 1.x – 2.5 – few options, mostly compile-time flags, some environment variables

Format: keyword = value

or

# comment

- 2.6 and beyond
  - Most compile time options moved to hsirc
  - Multiple stanzas
    - Global Section
    - Site Stanza

# HSIRC – cont.

```

This file is composed of "sections" in the following manner:

global: (reserved section name)
 global option 1
 global option 2...
site1: (HPSS sitename section)
 site1 option 1
 site1 option 2...
site2: (HPSS sitename section)
 site2 option 1
 site2 option 2...
```

# HSIRC – site stanzas (cont)

- Use with HSI:

```
hsi -s site_stanza_name <- command line
```


```
open -s site_stanza_name <- multi-hpss
```

- Default site can be specified in *global* section
- Command line options can override settings

```
hsi -s sdsc.prod -h hpss24i -p 1223
```

# HSIRC "global" section

auto\_backup = [on|off] enable/disable rename of existing file on get/put  
auto\_schedule = [on|off] enables or disable auto scheduling for <get>, etc.  
authmethod = [kerberos|dce|keytab|gsi]  
columns = [display columns]  
copies = [copies - ???]  
drive = [default "drive" letter if no site specified]  
enable\_ipi3 = [on|off]  
enable\_san3p = [on|off]  
enable\_sharedmem = [on|off]  
helpfile = [path to local helpfile?]  
keytab = [path to keytab file if "authmethod" = "keytab"]  
lines = [display lines]  
netopt\_path = [path to hpss\_netopt.conf]  
nologin\_path = [path to nologin file; doesn't seem to work]  
nwif\_path = [path of "network interfaces" file; e.g., pftp.config]  
promptlen = [max length of entire prompt string]  
promptdirlen = [max length of directory part of prompt string]  
PS1 = [main prompt string]  
PS2 = [continuation prompt string]  
principal = [principal name[,authmethods]]  
pwwfile = [password file; not sure how used?]  
site = [default sitename section to use]



# HSIRC – site stanzas

**site\_name:**

**hpss\_id = *ID*** [site ID (any – normally from list of DCE cell IDs for site)]

Purpose: allows multiple stanzas to reference same site, with some duplicate params (e.g. drive ID), if multiple stanzas needed for multiple servers

**default\_authmethod = *method*** [default authmethod to use for site]

**host = *server\_host*[,*server\_host*,...]** [Names or IP addresses of hosts running the NDAPI daemon]

**port = *port\_number*** [NDAPI daemon port number]

**authmethod = *kerberos|dce|keytab|gsi***

- one entry for each possible auth method to use for site

- followed immediately by auth-method-specific settings, e.g

**keytab = *path*** (for keytab auth method)

**credfile = *path*** (path to Kerberos credentials cache for site)

**pwfile = *path*** (for DCE combo method)

**drive = *drive\_letter*:** [logical drive letter to use for the site]

**principal = *principal\_name*[,*auth\_method*]** [principal or template name (%U)]



# HSI Auto-COS Selection

- Purpose: Augment HPSS COS-selection Hints
  - Number of Copies
  - Project-specific resources
    - Restrict by UID, Group ID or Account ID
  - Group sets of COS-s by Name  
(Named COSLISTs)
- Provide Framework for Site Customization



# Auto-COS Selection – tools

- `make_cos.pl` – generate initial template file

```
[/opt/hpss/local/hsi/2.8/hsi.2.8/datafiles] ls
HPSS.conf.template README account
cos cos.ORNL hsi.help.data
hsi.wrapper* hsirc.sample make_cos.pl
```

# Auto-COS Selection (NCDC)

```
rime:/hpss/mgleiche ->hsi
[connecting to yoda_s/1220]
Username: mgleiche_UID: 2359 CC: 2359 Copies: 1 [hsi.2.7 Fri Dec 6 14:31:47 EST 2002]
[HSI]/hpss/mgleiche->lscos
```

57 HPSS Classes of Service defined

| COS ID | Name                       | Exclusion Flags | Copies | Min File Size | Max Size + 1   |
|--------|----------------------------|-----------------|--------|---------------|----------------|
| 1      | 3280                       |                 | 1      | 0             | 10,485,760     |
| 2      | nmc                        |                 | 1      | 0             | 33,554,432,000 |
| 3      | smos                       |                 | 1      | 0             | 33,554,432,000 |
| 4      | nexrad                     |                 | 1      | 0             | 33,554,432,000 |
| 5      | common                     |                 | 1      | 0             | 33,554,432,000 |
| 6      | srrs                       |                 | 1      | 0             | 33,554,432,000 |
| 8      | common > 10M               |                 | 1      | 10,485,760    | 33,554,432,000 |
| 9      | 3280 > 10M                 |                 | 1      | 10,485,760    | 33,554,432,000 |
| 10     | level1b                    |                 | 1      | 0             | 33,554,432,000 |
| 12     | backup 3280                |                 | 1      | 0             | 10,485,760     |
| 14     | backup nmc                 |                 | 1      | 0             | 33,554,432,000 |
| 16     | backup srrs                |                 | 1      | 0             | 33,554,432,000 |
| 18     | backup 3280 > 10M          |                 | 1      | 10,485,760    | 33,554,432,000 |
| 19     | backup level1b             |                 | 1      | 0             | 33,554,432,000 |
| 22     | nexrad level 2             |                 | 1      | 0             | 33,554,432,000 |
| 23     | backup nexrad level 2      |                 | 1      | 0             | 33,554,432,000 |
| 24     | Fleet Fields               |                 | 1      | 0             | 33,554,432,000 |
| 26     | Level1b Historical         |                 | 1      | 0             | 33,554,432,000 |
| 27     | ssmi_3660                  |                 | 1      | 0             | 33,554,432,000 |
| 28     | backup Level1B Historical  |                 | 1      | 0             | 33,554,432,000 |
| 29     | backup ssmi 3660           |                 | 1      | 0             | 33,554,432,000 |
| 30     | common dsk-to-tape datsav3 |                 | 1      | 0             | 33,554,432,000 |
| 104    | New Nexrad                 |                 | 1      | 0             | 33,554,432,000 |
| 106    | New SRRS                   |                 | 1      | 0             | 33,554,432,000 |
| 108    | New Common > 10M           |                 | 1      | 0             | 33,554,432,000 |
| 109    | New 3280 > 10M             |                 | 1      | 0             | 33,554,432,000 |
| 110    | New Level1B                |                 | 1      | 0             | 33,554,432,000 |
| 112    | New Backup 3280            |                 | 1      | 0             | 33,554,432,000 |
| 114    | New Backup NMC             |                 | 1      | 0             | 33,554,432,000 |

# NCDC (cont.)

|      |                          |   |             |                |
|------|--------------------------|---|-------------|----------------|
| 116  | New Backup SRRS          | 1 | 0           | 33,554,432,000 |
| 118  | New Backup 3280 > 10M    | 1 | 0           | 33,554,432,000 |
| 119  | New Backup Level1B       | 1 | 0           | 33,554,432,000 |
| 124  | New Fleet Fields         | 1 | 0           | 33,554,432,000 |
| 126  | New Level1B Hist.        | 1 | 0           | 33,554,432,000 |
| 127  | New SSMI_3660            | 1 | 0           | 33,554,432,000 |
| 128  | New Backup Level1B Hist. | 1 | 0           | 33,554,432,000 |
| 129  | New Backup SSMI_3660     | 1 | 0           | 33,554,432,000 |
| 200  | HPSS Config              | 1 | 0           | 33,554,432,000 |
| 1001 | R2D2 Small               | 1 | 0           | 4,096,000      |
| 1002 | R2D2 Medium              | 1 | 4,096,001   | 32,768,000     |
| 1003 | R2D2 Large               | 1 | 32,768,001  | 33,554,432,000 |
| 1004 | R2D2 Huge                | 1 | 131,072,001 | 33,554,432,000 |
| 1101 | Nexrad Level 2           | 1 | 0           | 33,554,432,000 |
| 1102 | Backup Nexrad Level 2    | 1 | 0           | 33,554,432,000 |
| 1201 | Backup Small             | 1 | 0           | 4,096,000      |
| 1202 | Backup Medium            | 1 | 4,096,001   | 32,768,000     |
| 1203 | Backup Large             | 1 | 32,768,001  | 33,554,432,000 |
| 1204 | Backup Huge              | 1 | 131,072,001 | 33,554,432,000 |
| 1301 | IG88 Small               | 1 | 0           | 4,096,000      |
| 1302 | IG88 Medium              | 1 | 4,096,001   | 32,768,000     |
| 1303 | IG88 Large               | 1 | 32,768,001  | 33,554,432,000 |
| 1304 | IG88 Huge                | 1 | 131,072,001 | 33,554,432,000 |
| 1401 | GOES Small               | 1 | 0           | 16,777,216     |
| 1403 | GOES Large               | 1 | 16,777,216  | 33,554,432,000 |
| 1405 | Satellite_LTO Small      | 1 | 0           | 16,777,216     |
| 1407 | Satellite_LTO Large      | 1 | 16,777,216  | 33,554,432,000 |
| 8888 | LTO GEN2 Test            | 1 | 0           | 33,554,432,000 |

Flags: U/G/A - unavailable to current uid/gid/account N-no auto assignment

# Auto-COS Selection – File Format

## Line Format:

# Comment

text # trailing comment

text [\] <- for continuation lines

## Stanza Format:

name: type = *type* [,uidlist,gidlist,acctlist,coslist,cos]

entry [format depends o stanza type]

entry

...

name: type = *type* [cos,uidlist,gidlist,acctlist]

entry entry

...



# COS File – UIDLIST/GIDLIST/ACCTID List

UIDLIST/GIDLIST/ACCTLIST:

[-] ID[ ID...] [numeric IDs]

[-] ID[ ID ...]

- prefix excludes UID/GID/ACCTID
- multiple IDs (uid, gid, account ID) per line, space-separated
- multiple lines

# COS FILE – COS Stanza Format

*name:* type = cos

noauto [user must specifically ask for this COS]

comment = “comment string”

default\_auto [select this COS if filesize is unknown]

id = *hpss\_COSID*

cosname = “*class of service name*” [primarily for display]

copies = *n* [based on hierarchy]

uid = [-]*numeric\_uid* [uid to grant or deny access to for this COS]

uidlist = *uidlist* [previously defined uidlist stanza name]

gid = [-]*numeric\_gid* [gid to grant or deny access to for this COS]

gidlist = *gidlist* [previously defined gidlist stanza name]

acct = [-]*acct ID* [account ID to grant or deny access to for this COS]

acctlist = *acctlist* [previously defined acctlist stanza name]




# COS FILE – COS Stanza (cont)

media\_type = memory|disk|tape|rw\_optical|worm\_optical  
media\_subtype = *subtype name* [see template COS file]

Options from HPSS HINTS Structure (\*\* parsed, but not currently used)

\*\* writeops = write|rewrite|append|random|parallel  
\*\* readops = read|random|parallel  
\*\* frequency = hourly|daily|weekly|monthly|archive  
transfer\_rate = *n* [bps|kbs|mbs|gbs] (case-insensitive)  
min\_size = *n* [kb|mb|gb]  
max\_size = *n* [kb|mb|gb|tb|pb]  
\*\* block\_size = *n* [kb|mb]  
\*\* access\_size = *n* [kb|mb|blk]  
\*\* stripe\_width = *n* [default = 1]  
\*\* stripe\_size = *n* [kb|mb|blk]  
hierarchy = "*hierarchy name*" [for display only]  
latency = *n* [display only]  
stage\_code = "*stage code, e.g. On Open*" [display only]



# COS File – “Named COS” list

*name:* type = coslist

*cosID* [... *cosID*]

*cosID* [... *cosID*]

...

- cosIDs refer to previously defined COS stanzas, based on the “cosID” param in the stanza
- selection algorithm is restricted to COS stanzas from the list
- HSI usage:

```
set coslist = name
```


```
put (cp) coslist=name ...
```

# AutoCOS Selection

- COS selection is performed on server for non-DCE HSI (code is enabled when compiling DCE version of HSI)
  - Multi-HPSS considerations:
    - different COS list for each site
    - COS selection performed on server where file will be created



# AutoCOS Selection –Algorithm (1)

1. If filesize known, set `min_file/max_file` size hints, `HIGHLY_DESIRED_PRIORITY`
  2. Choose a set of eligible COS entries, based upon filesize, number of copies, [DCE] UID, GID, Account ID
  - 3. If no entries, use system default, as configured in HPSS
  - 4. Eliminate entries that are not auto-assignable (noauto option)
  - 5. If filesize is unknown, return first entry marked as “default\_auto” if there is one
  - 6. If no default\_auto, keep track of COS that can contain the largest file
  - 7. Find fastest COS for each media type of the entries that have `media_type` defined (memory, disk, tape, rw\_optical, worm)
  - 8. If no auto-assignment entries, and file size unknown, return entry that can contain largest file
  - 9. Return COS with fastest device based on type – memory, disk, tape, optical, worm
- 

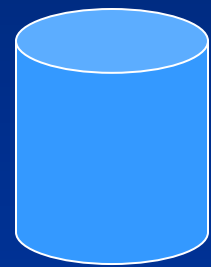
# AutoCOS Selection – APIs

- hpssex\_SelectCOS (select best COS)
- hpssex\_GetCOSList (list of defined COS stanzas)
- hpssex\_GetNamedCOSLists (list of Named COSs)
- hpssex\_GetNamedCOSInfo (list of COS IDs associated with Named COS)

# HSI Auto-Scheduling

- What is Auto-Scheduling?

HSI get file1 file2 file3 file4 file5 file6 file7 file8 file9 file10



file1  
file10

file5, pos=10  
file3, pos=25

file7, pos=100  
file8, pos=12100

file6, pos=12100  
file9, pos=10001

file2, pos=100  
file4, pos=12100



# HSI Auto-Scheduling (2)

Determining where File(s) live

## HSI ls -U – list HPSS-specific info

me/gleicher/pub/tools:

```
----- 1 gleicher hpss 6001 10190 TAPE 80947 Dec 17 2000slk.tar.g
```

COS Acct Residence



# HSI Auto-Scheduling (3)

– ls –V – list 1<sup>st</sup> tape level

```
/home/gleicher/pub/tools:
```

```
-rw----- 1 gleicher hpss 6001 10190 TAPE
 80947 Dec 17 2000 lslk.tar.gz
```

```
Storage VV Stripe
```

```
Level Count Width Bytes at Level
```

```

1 (tape) 1 1 80947
```

```
VV[0]: Object ID: 22ffbeae-bdae-11d5-a117-10005afa75bf
 ServerDep: 37f35956-fdd2-11d0-93cb-10005afa75bf
```

```
Pos:116871 PV List: X6100000
```

# HSI Auto-Scheduling (4)

– ls -X – list all levels

```
-rw----- 1 gleicher hpss 6001 10190 TAPE 756320 Dec 17 2000
 lsof.tar.gz
```

Storage VV Stripe

```
Level Count Width Bytes at Level
```

-----  
0 (disk) 0 1 (no data at this level)

1 (tape) 1 1 756320

VV[ 0]: Object ID: 22ffbeae-bdae-11d5-a117-10005afa75bf  
 ServerDep: 37f35956-fdd2-11d0-93cb-10005afa75bf  
 Pos:116872 PV List: X6100000

2 (tape) 1 1 756320

VV[ 0]: Object ID: 869bc31e-bddf-11d5-a117-10005afa75bf  
 ServerDep: 37f35956-fdd2-11d0-93cb-10005afa75bf  
 Pos:116870 PV List: X6105000

3 (tape) 0 0 (no data at this level)

4 (tape) 0 0 (no data at this level)



# HSI AutoScheduling (5)

– `ls -P` – list “-V” info on one line

```
FILE /home/gleicher/pub/tools/lslk.tar.gz 80947 80947
116871 X6100000
```

```
FILE/home/gleicher/pub/tools/lsof.tar.gz 756320 756320
116872 X6100000
```

– simple script to sort/create HSI “IN” file

*op xxx* [chcos, stage, get, ...]

*op yyy*



# HSI AutoScheduling (background)

- ORNL (Jae Kerr)
  - Unitree, early NDAPI library contribution
    - Implemented to meet needs of ARM project
    - Polling
    - Non-DCE library
    - Conditional code

```
hpss_QueueInit
```

```
hpss_QueueFile
```

```
hpss_CheckFile
```

# HSI AutoScheduling - Goals

- Optimize tape mounts
- General purpose mechanism for any file-oriented HSI command
- Usable by other Client API applications (e.g., htar)



# HSI AutoScheduling

- General Purpose Scheduling APIs
  - usable in DCE and non-DCE HSI
  - create sorted lists (tape, pos. within tape)
  - background staging with notification from BFS (no polling)
  - foreground retrieval of files from disk cache
  - detect BFS restart or inability for BFS to notify
  - throttling on in-progress staging



# HSI AutoScheduling - Overview

1. initialize new queue
2. add files to the queue
3. finalize queue – sort by on-disk, position within each tape VV
4. chcos / stage : step through sorted list, perform function
5. get:
  - initiate background staging
  - loop until all files transferred, waiting for next <ready> file (on disk, or on-tape)

# HSI AutoScheduling - APIs

## Main API(s)

`hpssex_SchedInitQueue` – prep. for new scheduling session

`hpssex_FreeQueue` – cleanup scheduling queue

`hpssex_SchedAddFile` – add file to queue

- determines disk/tape residence
- adds to list of files for disk or tape VV

`hpssex_SchedSortQueue`

- sorts files by position within each tape VV

`hpssex_SchedGetQueueEntry`

- walks through list of sorted entries

`hpssex_SchedQueryFileStatus`

- lookup status of a queued entry (by handle)

`hpssex_SchedInitBGStage`

- Issue background stage requests
- receive notifications from BFS
- detect/retry (BFS restarts, inability of BFS to notify)

# HSI AutoScheduling

- Interesting problem:
  - What happens if HSI is stopped/restarted while background stages are active?



# HSI AutoScheduling

## HSI-specific

- New hsirc global setting
  - auto\_scheduling = [on/off]
- Command line option to disable (-N) or enable (-A)
- Automatically disabled for HPSS < 4.3
- “here document” syntax (chcos,get,stage)

– interactive mode or IN file

```
get [options] [local : hpss ...] <<EOF
```

```
local : hpss [local : hpss] ...
```

```
local : hpss [local : hpss] ...
```

```
EOF
```



# HSI – Conclusion

Thank you for your attention.

Please visit the HSI web site at:

[www.sdsc.edu/Storage/hsi](http://www.sdsc.edu/Storage/hsi)

Upcoming attractions:

Configuring and Tuning HSI – Q & A session  
with Cygwin Example (if Mike's PC is working)

