Crimson STA Streaming Operators Guide

Dorota Genser and
Harry L. Melanson

August 14, 1995

ABSTRACT

This document describes the day-to-day procedures for running the STA streaming on the Crimsons. It is applicable to the "dpm" system, installed in July, 1995.

1. Introduction

STA streaming is performed by DØ on an SGI Crimson, consisting of two nodes, *fnstfy* and *fnstfz*. Each node processes data independently. The streaming operation is performed in five logical steps: input spooling, streaming, merging, output spooling and rcp-file copying. The input spoolers copy data from tape onto disk. The streaming is done by the program *omni_filter*. Merging takes various streams and merges them together, for output to tape by the output spoolers. The rcp-file copying process communicates with the FATMEN and production data base utilities on *DØFS* by copying rcp files. At this time, there are 2 input spoolers, 1 *omni_filter*, 2 mergers and 4 output spoolers running on each node. All processing is carried out under the account *dzero*.

2. System Startup

In order to start up all streaming related processes, say after system maintainence, you enter the commands:

```
% dpm_inp submit bg nostop all_hosts all_jobs
% dpm_omf submit bg nostop all_hosts all_jobs
% dpm_mrg submit bg nostop all_hosts all_jobs
```

```
% dpm_out submit bg nostop all_hosts all_jobs
% dpm_rcp submit bg nostop all_hosts all_jobs
```

These commands need to be entered on each machine separately (eg. both *fnstfy* and *fnstfz*). Note that if you only want to start streaming on one node, you should enter these commands on that node, without the "all_hosts" option.

The commands start the input spoolers (inp), *omni_filter*(s) (omf), mergers (mrg), output spoolers (out) and rcp-file copiers (rcp).

3. Daily Operations

Running the system requires daily monitoring (actually you should check things a few times per day). In general, one needs to:

1. Make sure all processes are running properly (eg. inp, omf, mrg, out and rcp).
2. Make sure the disks have not filled up (indicating that some of the processes are not running properly).
3. Make sure the tape drives are working OK.
4. Make sure the input spoolers have a list of tapes to process.
5. Make sure the output spoolers have a list of blank tapes to use.

The following sections describe how to perform these tasks.

3.1. Checking Processes

In order to check the general status of processes, execute the command:

```
% job_stat
```

It will list the latest status of each process on each machine, along with the date and time of the last report for that process. The first sign of problems is when this date/time is not relatively close to the current date/time. Hung processes will keep the same date/time for ever. Occasionally, you will see "No match" listed – this should eventually go away as the system launches a new version of that process. To confirm a process is hung, list all current *dzero* processes by using the command

```
% sys_stat
```

If your suspect process has a "38" in the first column of the report (actually only the "8" is important), then that process cannot be killed (it is uninterruptable) except by rebooting.

If you identify a hung process, it is probably one or more input or output spoolers. These hung processes can occur when a SCSI reset happens. The only known solution is to have that machine rebooted. You will have to follow the commands listed in the System Shutdown section. It is likely that the SCSI resets were caused by a bad tape drive, so you should follow the instructions in Checking Tape Drives as well. To get the node rebooted, send a message to *farm-admin@fnsg01*. Note that the way the disks are organized, you can continue to run streaming on *fnstfz* while *fnstfy* is being rebooted, but not the opposite (the usr disk is mounted on *fnstfz*).

3.2. Checking Disks

You should monitor the disks, to make sure they are not full. Use the command:

```
% disk_stat
```

Disks that are near 90% full are considered full by the streaming system, and are indicative of a problem. Check the processes and tape drives carefully.

3.3. Checking Tape Drives

One symptom of a bad tape drive is it generates errors. Use the following command to list the last 10 tape errors detected by the system:

```
% tape_errors
```

Tape drives which have multiple errors should be considered bad, and flagged to the system as broken using the command

```
% ocs_broken -h host -t drive
```

where *host* is the host name (*fnstfy* or *fnstfz*) and *drive* is the drive name (eg. `sts66`). To list all currently known broken drives on the system, use the command

```
% ocs_tape -h any -s broken
```

When this list gets long (say 4 drives), you should arrange for the drives to be replaced by sending a message to *farm-admin@fnsg01*. This is typically done on Tuesday afternoons and you will have to stop the system that morning by following the instructions in the System Shutdown section.

Another symptom of a bad drive is a SCSI reset (which can also cause a lot of other problems!). To get a list of recent SCSI resets, use the command:

```
% scsi_resets
```

You can correlate the SCSI resets with the tape errors, to help confirm that a drive is broken.

Occasionally the tape drives need to be cleaned. To make a list of drives that should be cleaned, enter the command

```
% ocs_clean_list -c on
```

To request that these drives be cleaned, use the command

```
% ocs_clean_list -c on -o short | ocs_clean_it -A -q 3600 &
```

The operators will then clean the drives when they become available (if they become available in "3600" seconds).

3.4. Check Input Tape List

To check what is currently in the "to-do" list for the input spoolers, issue the command:

```
% intape_stat
```

When there are fewer than 200 partitions to process, you need to add more to the list. You will have to have some idea of what has not yet been processed, although the system does keep track and will not merge/output already processed files. Typically you will add a group of partitions based on Reco version and run range. The following example is for Reco 12.20 and runs in the range 92000-92999.

The first step is to make the list of partitions using FATMEN. Log onto *D0FS* (you can use `% rlogin d0fsa` if you are logged in as *dzero* on *fnstfy* or *fnstfz*) and do the following:

1. `$ libtest fatmen`
2. `$ d0setup fatmen`
3. `$ fm_list` *gen_list out_list*

where *gen_list* is the generic file name, eg.

```
//fnal/d0/clda/93p1e18/sta01reu1220/allall00nonex00/09*/r092*
```

and *out_list* is the name of the output list to create, eg.

```
crimson_092_1220.list
```

When this is done, go to *fnstfy* or *fnstfz*, and ftp the above output list to the ~/`dpm` area. Then enter the command

```
% dpm_inp add crimson_092_1220.list STREAM 50
```

When this is done, it is useful to rename this list to indicate that you have added it to the input spoolers to-do list, eg.

```
% mv crimson_092_1220.list crimson_092_1220.list_added_950809
```

where "950809" is the date.

3.5. Checking Output Tape List

To check the list of blank tapes used by the output spoolers, enter the command:

```
% outtape_stat
```

When the number of blank tapes becomes less than 20, you should add to the list by issuing the command:

```
% dpm_out add ~/dpm/OUTSPOOL_tapes YYxxxx YYzzzz
```

where "YYxxxx" and "YYzzzz" are the starting and ending YY tapes to add. Typically one adds 100 tapes to the list each time.

3.6. Checking Everything

To facilitate checking everything, the command

```
% check_all
```

can be used. It runs each of the above procedures.

3.7. Other Utilities

To get a summary of the number of events streamed on a particular day, use

% omf_day_sum *date*

where *date* is, eg. 950813.

To get a summary of all events streamed (since the new system was installed in July, 1995), use

```
% omf_sum
```

This can take awhile.

To get more detailed summaries of each of the streaming processes, use the commands:

```
% check_inp_log
% check_omf_log
% check_mrg_log
% check_out_log
```

These commands show the last 20 lines of the log files for the input spoolers, *omni_filter*'s, mergers and output spoolers respectively.

4. System Shutdown

In order to stop all of the streaming related processes, say for system maintainence, enter the following commands:

```
% dpm_inp stop all_hosts all_jobs
% dpm_omf stop all_hosts all_jobs
% dpm_mrg stop all_hosts all_jobs
% dpm_out stop all_hosts all_jobs
% dpm_rcp stop all_hosts all_jobs
```

These commands need only be entered on one of the nodes. Note that if you only want to stop streaming on one node, you can enter these commands on that node, without the "`all_hosts`" option.

In principle this stops all processes. However, there have been problems with "hung" processes which cannot be stopped without rebooting the system (see the Daily Running section for more info on identifying such processes). In order to gracefully handle this situation, you must kill the corresponding "perl script" for these hung processes by hand. After issuing the above `stop` commands, wait awhile for things to come to a stop (say 10 minutes). You may then identify the hung processes using the utility `sys_stat`. Locate the process with the `.scr` name corresponding to the particular hung job and kill it with the `kill -9` command. The corresponding `exe` process is the actual hung job, which can only be removed by rebooting. By killing the `.scr` process, the streaming system will clean up old files and be in a more well-defined state for rebooting.

5. Further Reading

For more information, see the following:

1. Fermilab GA0015, *OCS Overview* (available as a postscript file in $OCS_DIR/doc/overview.ps.

2. Fermilab GA0012, *OCS Reference Guide* (available as a postscript file in $OCS_DIR/doc/reference.ps.