The Data Services Archive

Rena A. Haynes, Wilbur R. Johnson
Sandia National Laboratories
rahayne@sandia.gov, wrjohns@sandia.gov

NASA IEEE MSST 2004
12th NASA Goddard/21st IEEE Conference on
Mass Storage Systems & Technologies
The Inn and Conference Center
University of Maryland University College
Adelphi MD USA
April 13-16, 2004

Sandia
National
Laboratories

ADVANCED
SIMULATION &
COMPUTING™

Archival Storage Configuration

High Performance Storage System (HPSS)

- StorageTek and IBM tape libraries
- Parallel tape
    - 4-way, 2-way, 1-way tape striping
- Minimal HPSS managed disk storage
- Access through FTP and PFTP

The Data Services Archive capability is designed to simplify and optimize the process of archiving large data sets.

Massive Data and File Movement Issues

- User interface
  - File specification
  - Determining Progress
  - Error Detection
  - Recovery Procedures

- Resource requirements
  - Disk
  - Network
  - Archival

Movement across Distance Issues

- Data integrity and robustness
  End-to-end integrity
  Error recovery can exacerbate problems

- Archival device and media latencies can be magnified

- Networks and protocols not tuned for large archival data transfers

On-Demand Access to Parallel Archive Issues

- Inefficient use of resources

  Large transfer followed by small transfer followed by large transfer

  Multiple large transfers can block smaller transfers

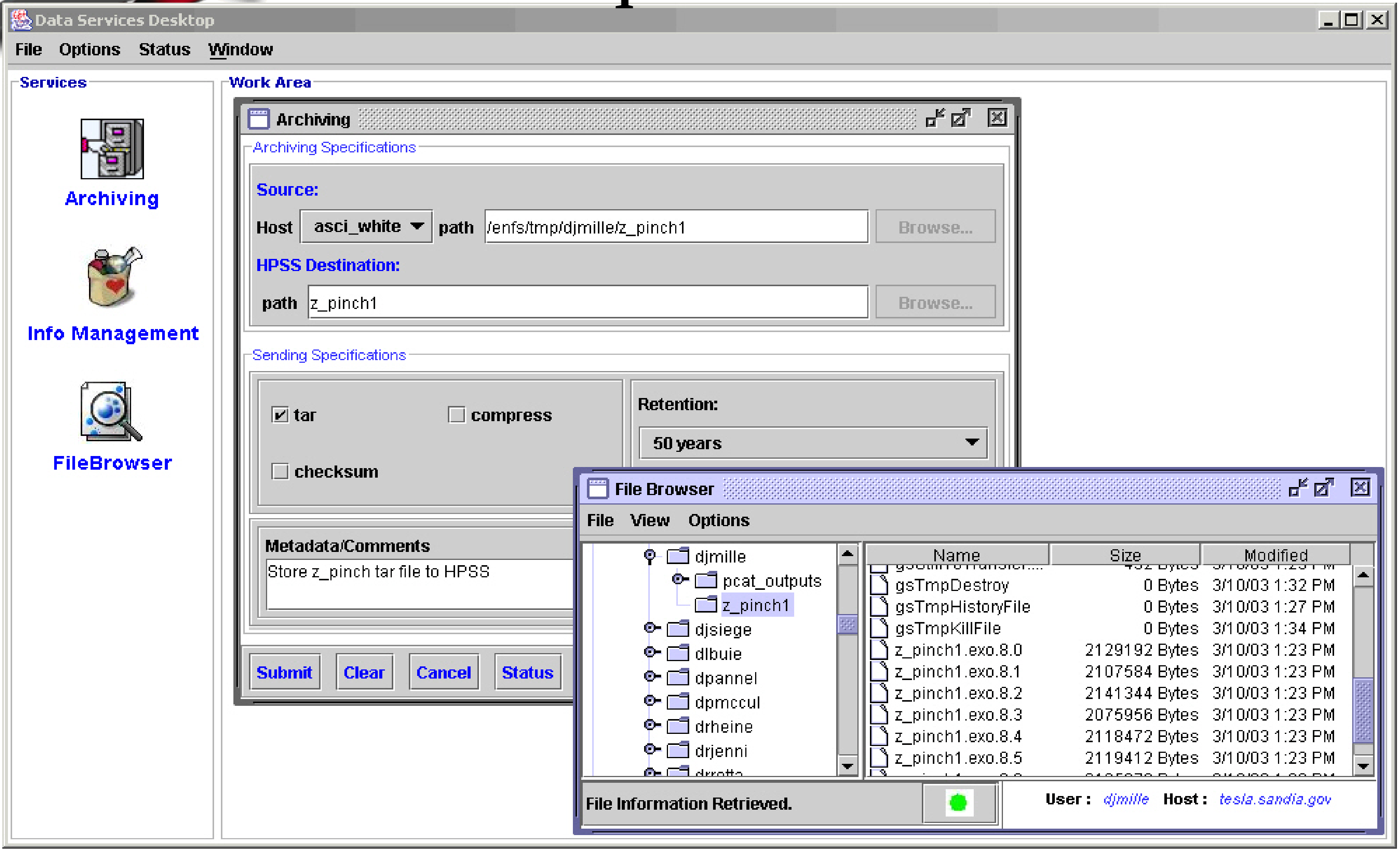
  Use of parallel tapes to store small files

- Over-subscription of resources

  Can cause failures due to credential timeout and ultimately denial of service

  Cascades into additional resource requirements, e.g., disk

DSA Request Interface



Archiving Specifications

Source:

Host: asci_white

path: /enfs/tmp/djmille/z_pinch1

HPSS Destination:

path: z_pinch1

Sending Specifications

☑ tar
□ compress
□ checksum

Retention: 50 years

Metadata/Comments: Store z_pinch tar file to HPSS

- djmille
  - pcat_outputs
  - z_pinch1
- djsiege
- dlbuie
- dpannel
- dpmccul
- drheine
- drjenni

Name	Size	Modified
gsTmpDestroy	0 Bytes	3/10/03 1:32 PM
gsTmpHistoryFile	0 Bytes	3/10/03 1:27 PM
gsTmpKillFile	0 Bytes	3/10/03 1:34 PM
z_pinch1.exo.8.0	2129192 Bytes	3/10/03 1:23 PM
z_pinch1.exo.8.1	2107584 Bytes	3/10/03 1:23 PM
z_pinch1.exo.8.2	2141344 Bytes	3/10/03 1:23 PM
z_pinch1.exo.8.3	2075956 Bytes	3/10/03 1:23 PM
z_pinch1.exo.8.4	2118472 Bytes	3/10/03 1:23 PM
z_pinch1.exo.8.5	2119412 Bytes	3/10/03 1:23 PM

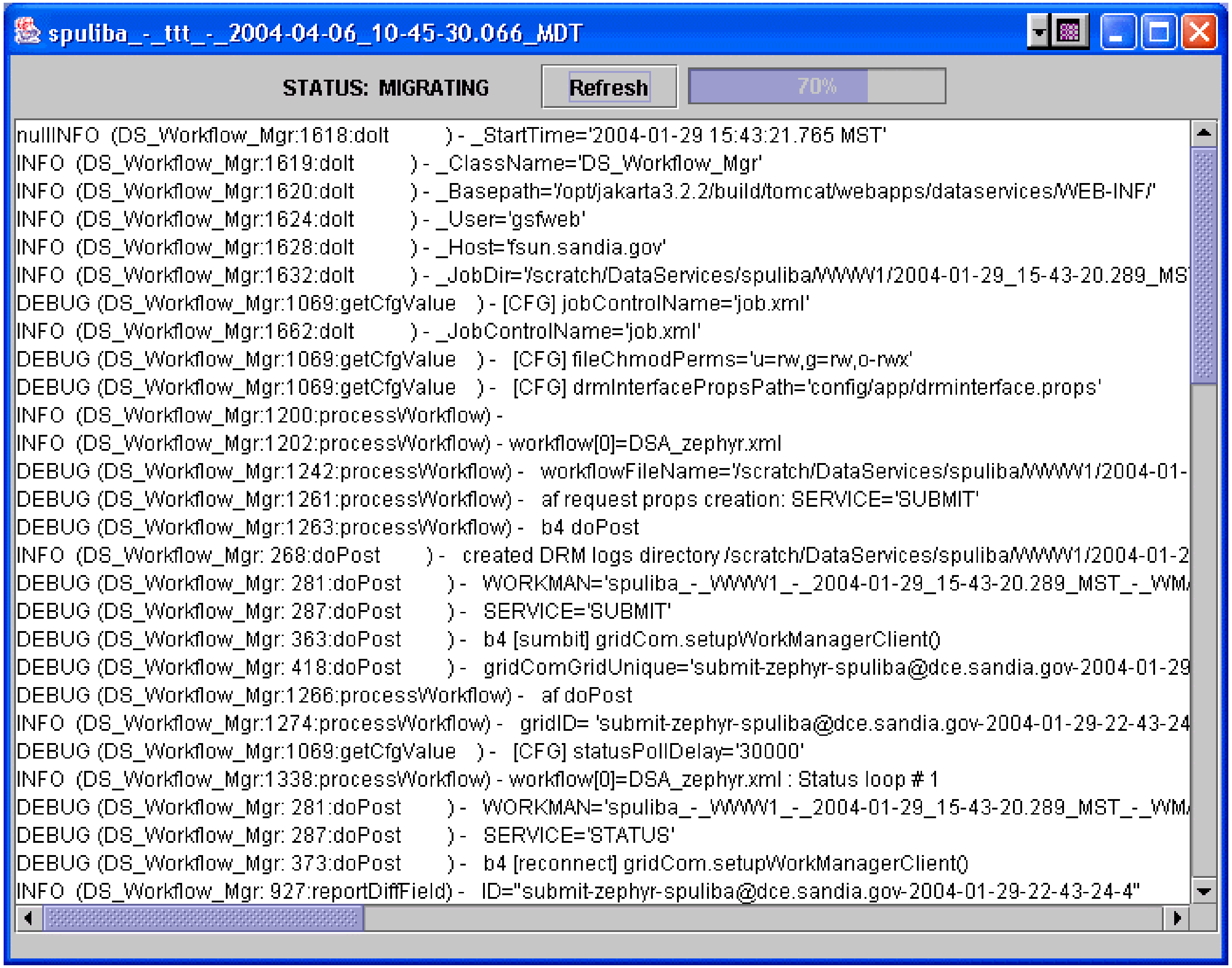
File Information Retrieved. ●

User : *djmille* Host : *tesla.sandia.gov*

Sandia National Laboratories

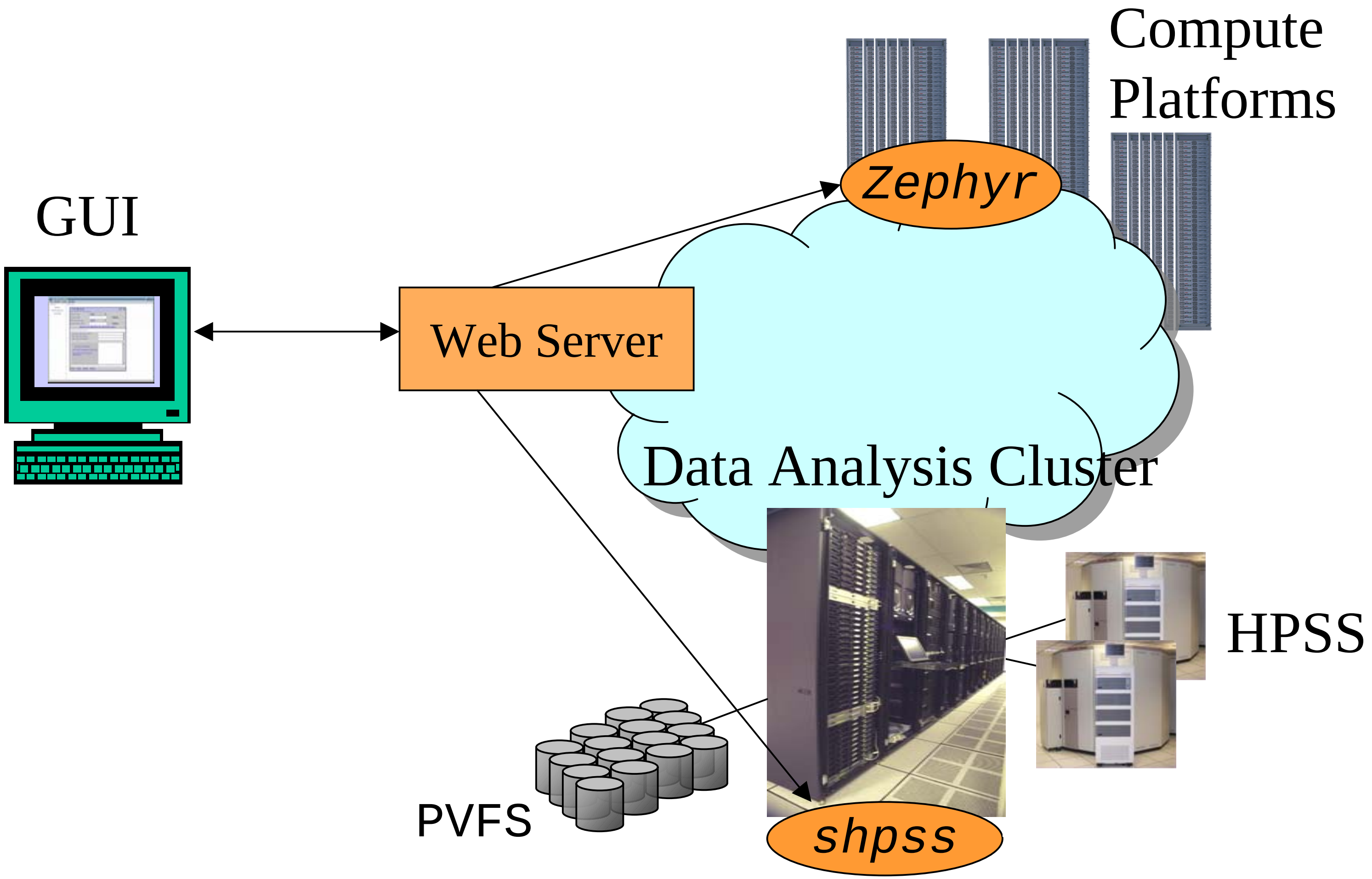
Advanced Simulation & Computing™

Monitoring DSA Request



```
nullINFO  (DS_Workflow_Mgr:1618:doIt         ) - _StartTime='2004-01-29 15:43:21.765 MST'
INFO  (DS_Workflow_Mgr:1619:doIt         ) - _ClassName='DS_Workflow_Mgr'
INFO  (DS_Workflow_Mgr:1620:doIt         ) - _Basepath='/opt/jakarta3.2.2/build/tomcat/webapps/dataservices/WEB-INF/'
INFO  (DS_Workflow_Mgr:1624:doIt         ) - _User='gsfweb'
INFO  (DS_Workflow_Mgr:1628:doIt         ) - _Host='fsun.sandia.gov'
INFO  (DS_Workflow_Mgr:1632:doIt         ) - _JobDir='/scratch/DataServices/spuliba/WWW1/2004-01-29_15-43-20.289_MS
DEBUG (DS_Workflow_Mgr:1069:getCfgValue   ) - [CFG] jobControlName='job.xml'
INFO  (DS_Workflow_Mgr:1662:doIt         ) - _JobControlName='job.xml'
DEBUG (DS_Workflow_Mgr:1069:getCfgValue   ) -   [CFG] fileChmodPerms='u=rw,g=rw,o-rwx'
DEBUG (DS_Workflow_Mgr:1069:getCfgValue   ) -   [CFG] drmInterfacePropsPath='config/app/drminterface.props'
INFO  (DS_Workflow_Mgr:1200:processWorkflow) -
INFO  (DS_Workflow_Mgr:1202:processWorkflow) - workflow[0]=DSA_zephyr.xml
DEBUG (DS_Workflow_Mgr:1242:processWorkflow) -   workflowFileName='/scratch/DataServices/spuliba/WWW1/2004-01-
DEBUG (DS_Workflow_Mgr:1261:processWorkflow) -   af request props creation: SERVICE='SUBMIT'
DEBUG (DS_Workflow_Mgr:1263:processWorkflow) -   b4 doPost
INFO  (DS_Workflow_Mgr: 268:doPost        ) -   created DRM logs directory /scratch/DataServices/spuliba/WWW1/2004-01-2
DEBUG (DS_Workflow_Mgr: 281:doPost        ) -   WORKMAN='spuliba_-_WWW1_-_2004-01-29_15-43-20.289_MST_-_WM
DEBUG (DS_Workflow_Mgr: 287:doPost        ) -   SERVICE='SUBMIT'
DEBUG (DS_Workflow_Mgr: 363:doPost        ) -   b4 [sumbit] gridCom.setupWorkManagerClient()
DEBUG (DS_Workflow_Mgr: 418:doPost        ) -   gridComGridUnique='submit-zephyr-spuliba@dce.sandia.gov-2004-01-29
DEBUG (DS_Workflow_Mgr:1266:processWorkflow) -   af doPost
INFO  (DS_Workflow_Mgr:1274:processWorkflow) -   gridID= 'submit-zephyr-spuliba@dce.sandia.gov-2004-01-29-22-43-24
DEBUG (DS_Workflow_Mgr:1069:getCfgValue   ) -   [CFG] statusPollDelay='30000'
INFO  (DS_Workflow_Mgr:1338:processWorkflow) - workflow[0]=DSA_zephyr.xml : Status loop # 1
DEBUG (DS_Workflow_Mgr: 281:doPost        ) -   WORKMAN='spuliba_-_WWW1_-_2004-01-29_15-43-20.289_MST_-_WM
DEBUG (DS_Workflow_Mgr: 287:doPost        ) -   SERVICE='STATUS'
DEBUG (DS_Workflow_Mgr: 373:doPost        ) -   b4 [reconnect] gridCom.setupWorkManagerClient()
INFO  (DS_Workflow_Mgr: 927:reportDiffField) -   ID="submit-zephyr-spuliba@dce.sandia.gov-2004-01-29-22-43-24-4"
```

DSA is a Distributed Application



DSA use of Grid Technologies

- Grid workflow processor to sequence data transfers to intermediate disks then to tape
- Globus toolkit used for submitting partitions to queues
- Globus toolkit enhancements
  - Immediate feedback from components
  - Ability to request specific resources

Resource Contention

- Staging files to intermediate disk cache
  - Reduces potential for retries/retransmissions caused by tape latencies
  - Allows grouping, or partitioning, of data to optimize use of archive resources
  - Simplifies interface to parallel archive
- Intermediate disk cache directly accessible by HPSS movers reduces external network traffic
- Resource management to minimize contention

Dataset Partitioning

- Place dataset files into transfer groups based on file characteristics (currently, file size)
- Number of files in a partition are limited to facilitate recovery and prevent large transfers from starving smaller transfers
- Partitions are scheduled based on parallel tape resource requirements

Scheduling Techniques

HPSS state and resources are modeled as nodes in a compute cluster

- Holds ftp sessions that cannot be satisfied
- Increase HPSS utilization
- Requires knowledge of HPSS configuration
- Requires knowledge of HPSS striping policies

Scheduling Techniques

DSA queuing system scheduler is configured to use backfill

- Allows pending transfers which can complete with available resources in the time to start the transfer at the head of the queue
- Requires estimating time to transfer

$$T_{\text{job}} = T_{\text{login}} + \sum_{i=1}^{i=N} \left[T_{\text{startup}} + \left\lceil \left(X_i / R_{\text{rate}} \right) \right\rceil + \left(S_{\text{width}} \times T_{\text{load}} \right) \right]$$

Performance Observations

- Staging
  - Allows for local file puts (up to 30MB/s per tape stripe)
  - Relieves source space more efficiently
- Partitioning
  - A tunable scheduling parameter for optimizing archive flow into HPSS
- Scheduling with backfill
  - Increases resource utilization and archive request throughput

Conclusions

- Reducing resource contention, managing data partitions, and scheduling transfers have increased overall performance of the HPSS parallel tape archival system.
- More performance data should be collected and analyzed to adjust transfer time estimations and determine additional tuning parameters.
- Partitioning scheme should be optimized based on partition data size as well as file count.