VINCI SAS/Grid Overview

Mark Ezzo

VINCI SAS Administrator

December 1, 2011

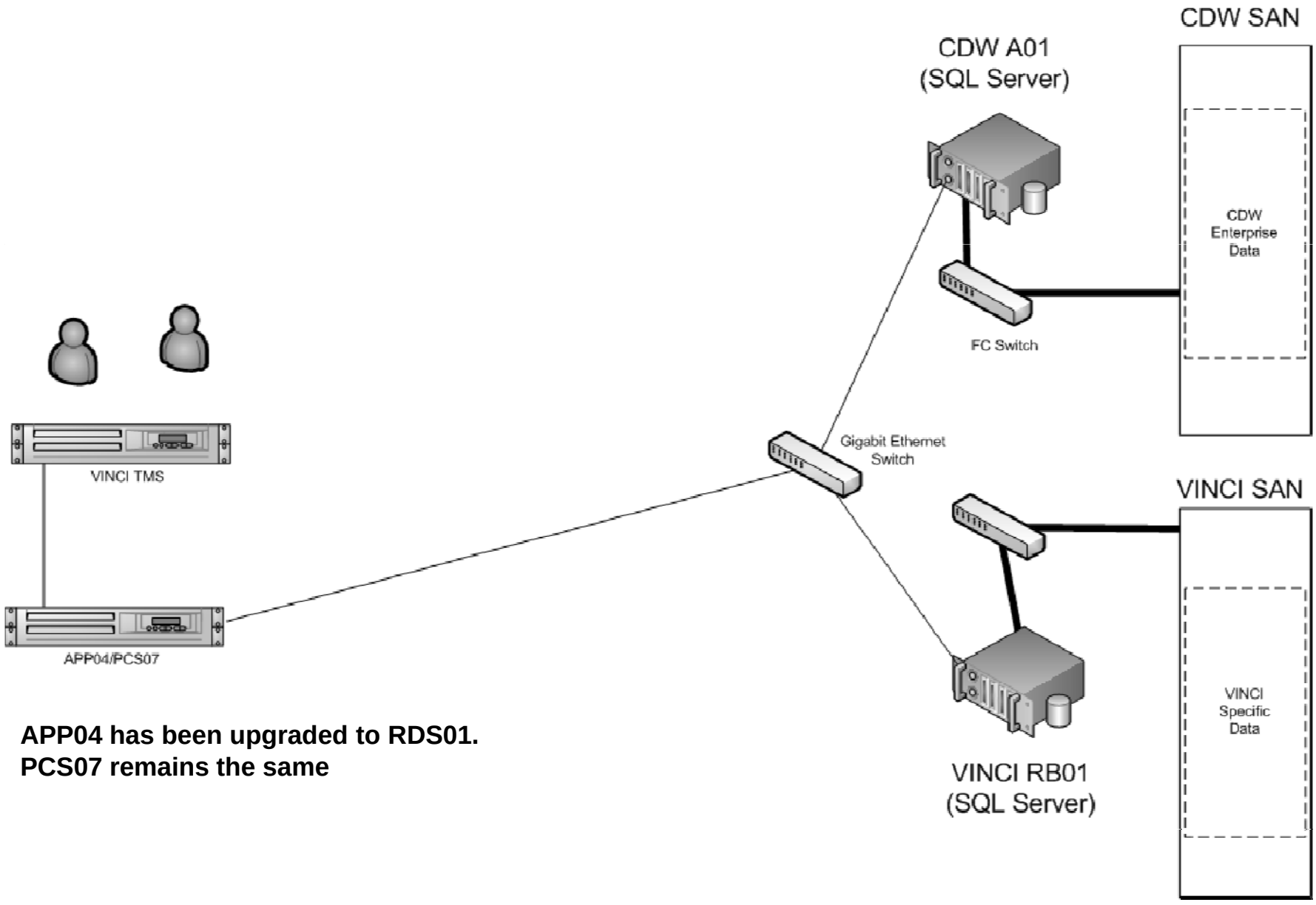
VINCI SAS/Grid Agenda

- **SAS Grid Introduction**
  - Overview
  - User Interfaces
  - Grid Enabling Existing SAS
- **Examples**
  - Base SAS
  - Enterprise Miner
  - Batch Submit
- **Summary**

VA Informatics and Computing Infrastructure (VINCI) Foundations

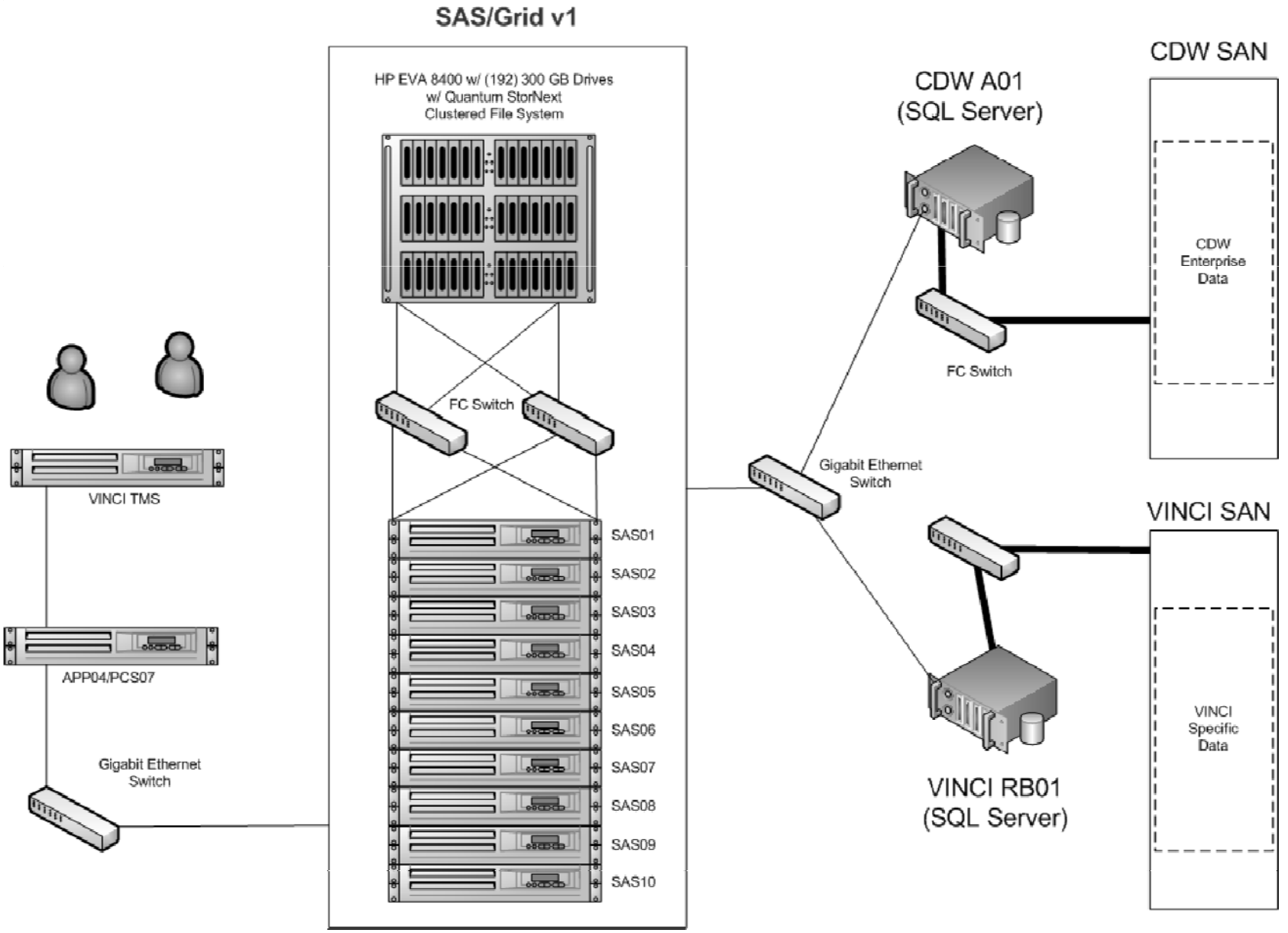
- **VINCI Research Infrastructure**
  - -- A Leading – Edge Research Venue utilizing SAS and SAS/Grid
- **New Data Privacy & Security Capabilities**
  - – New and improved methods to protect the privacy and security of data used in the research environment.
- **Remote Analysis Capability**
  - – New methods for researchers to perform research remotely without removing data from the VINCI system.
- **Prepare for Future Data Capabilities**
  - – Develop and implement methods to extract and make useable unstructured data such as text, genomic data, and images.

SAS/Grid v1.0: Pre Grid

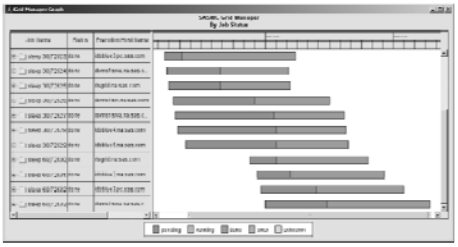


APP04 has been upgraded to RDS01.
PCS07 remains the same

SAS/Grid Windows: Architecture

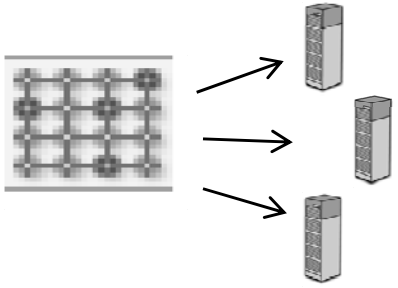


SAS/Grid Value



1. Approved Resource Control

- Resource Allocation, High Availability, Failover
- User Prioritization, Queuing, Monitoring
- Job Management, Scheduling, Suspension



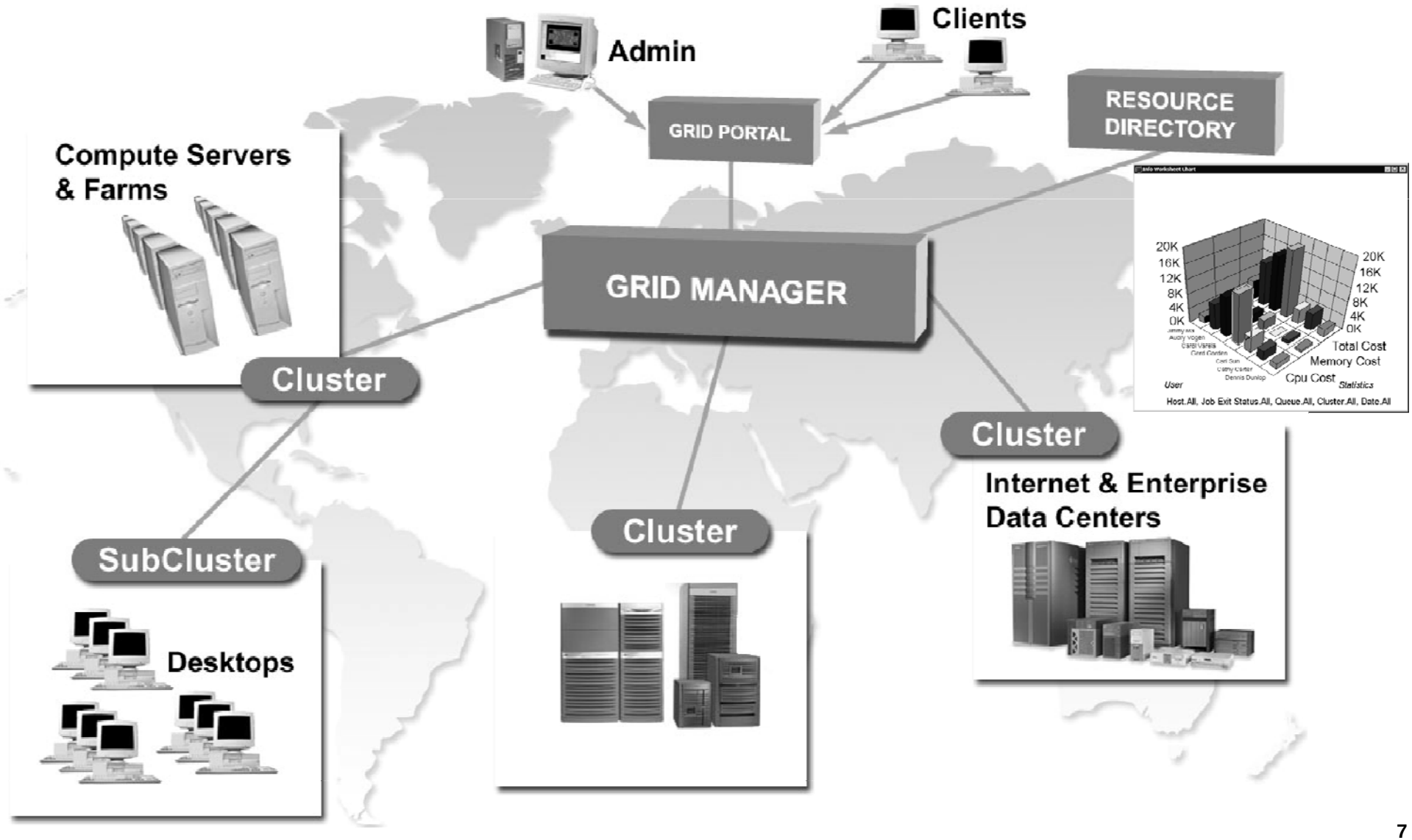
2. Increases Performance

- Workload Balance Users Across Available Resources
- Parse Jobs to Simultaneously Process on Multiple CPUs
- Improve long running jobs as well as Ad-Hoc Queries

3. Batch Processing

- User can submit and forget (no need to remain connected)
- User can view job output while job is running
- Allows for SAS checkpoint/restart capability
- Uses SAS Grid Manager metadata for centralized control

Grids are built, not bought

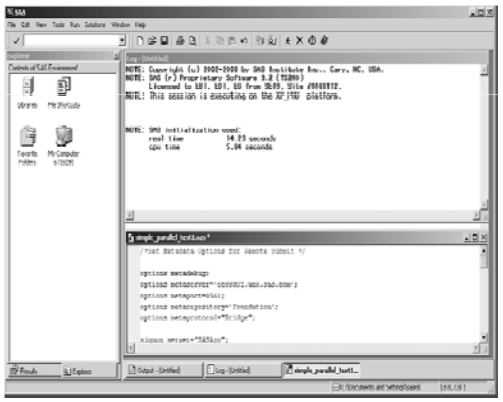


SAS/Grid Road Map

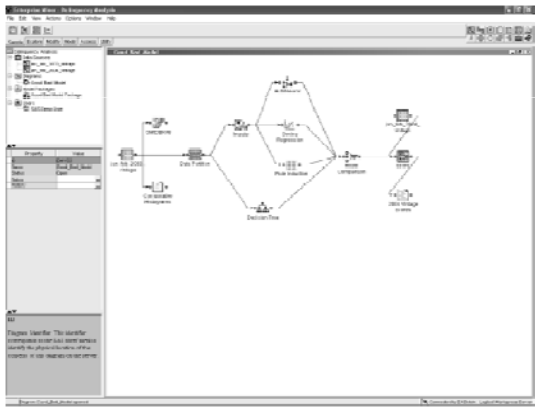
- **SAS/Grid v1.0** **Now: (10) Node Windows 2008 R2 Cluster**
- **SAS/Grid v2.0** **Plus (6) Node RHEL Linux Cluster w/ modified clustered file system architecture**

Run Windows and Linux in parallel for a period of time then decide on one platform going forward

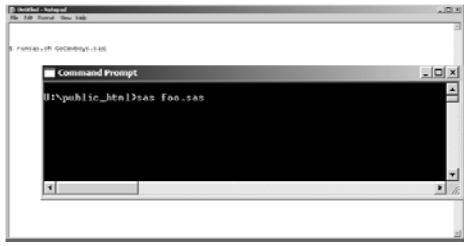
SAS User Interfaces



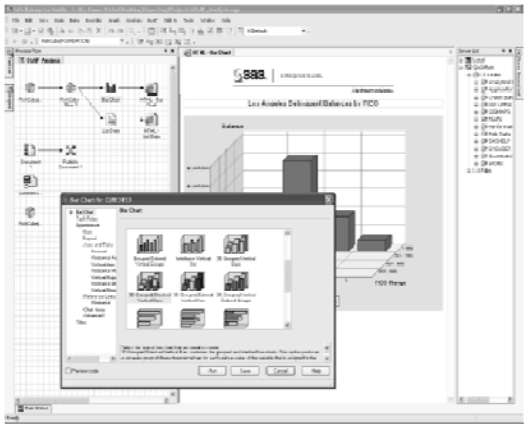
Base SAS ®



SAS® Enterprise Miner



Batch Submit



SAS® Enterprise Guide (TBD)

New Additions

- **The SAS Add-In for Microsoft Office is a Component Object Model (COM) add-in that extends Microsoft Office by enabling you to use the power of SAS analytics and access data directly from**

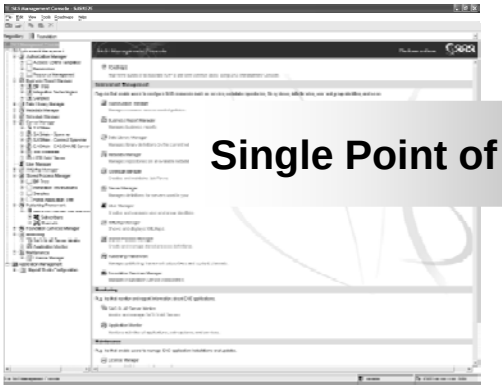
  Microsoft Excel, Word, PowerPoint, and Outlook.

- **SAS 9.3 Stored Processes introduces stored process reports and the STP procedure. A stored process report is a new object type that contains stored process output that is cached. The output can be viewed without re-executing the stored process. PROC STP enables users to execute a stored process from a SAS program. PROC STP can be executed in an interactive, batch, or server SAS session and can even be executed by another stored process. Essentially, anyone with a Web Viewer can execute and view the results, without using SAS itself.**

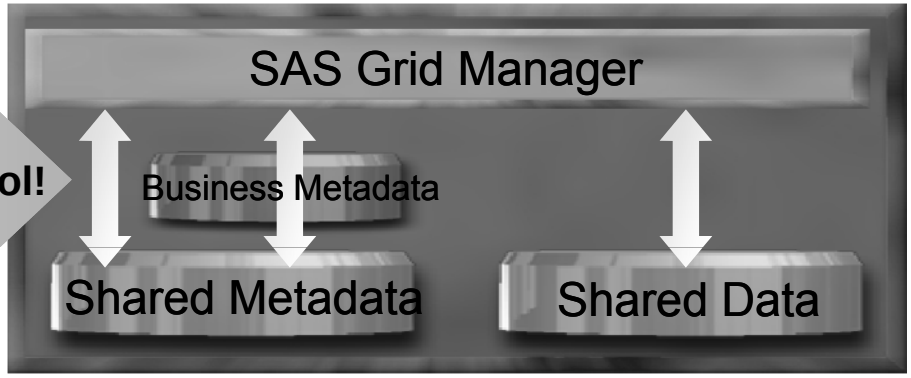
SAS Administration

Scheduling

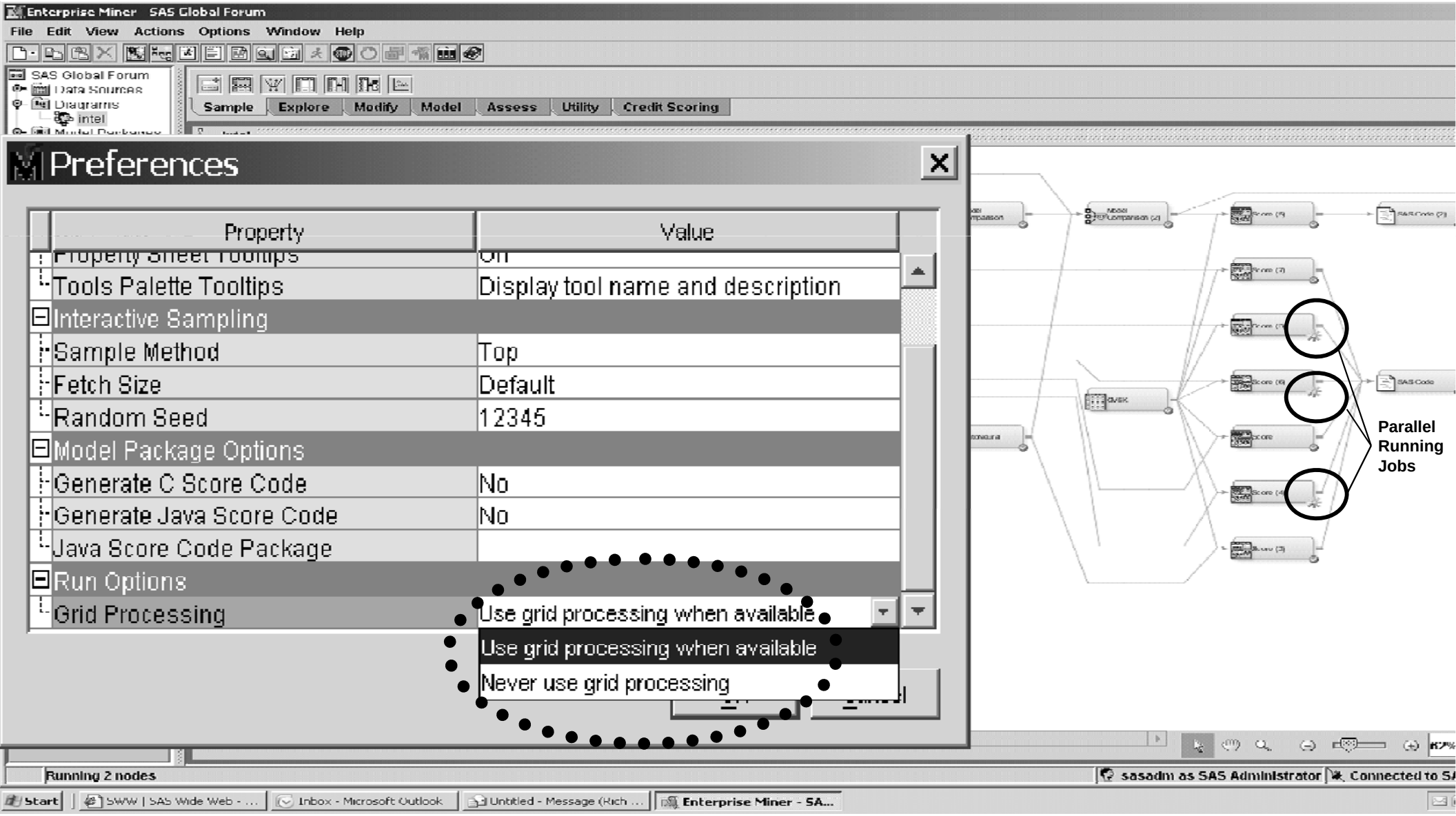
RTM (Report , Track, Monitor)



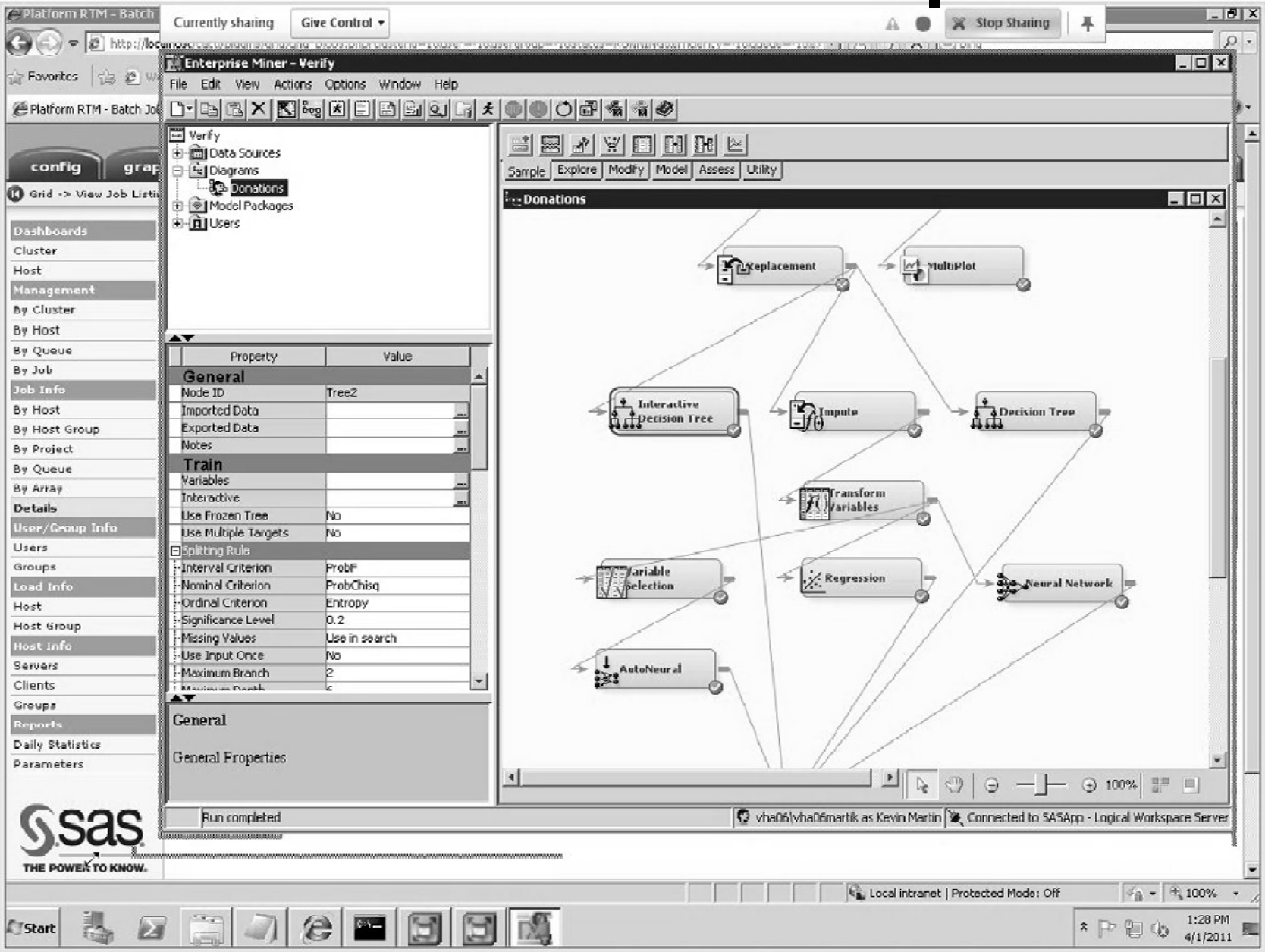
SAS® Management Console



SAS Enterprise Miner (User Interface for SAS Data Mining and Modeling) – Parallelized Workload Balancing



VINCI SAS Grid Example



SAS/Grid Next Steps

- **Arrange training sessions**
- **Establish participatory user groups**
- **Assist in optimization of project techniques**

VINCI SAS/Grid Summary

- **Load Balancing & Parallel Processing**
  - Support for large SAS Community
- **Reuse of existing SAS Programs**
  - Small learning curve for immediate value or submit legacy code immediately to the grid in SAS 9.3 via the GSUB Server
- **High Availability/Automatic Failover**
  - Less downtime = more research
- **Easy Administration**
  - Quickly provide help and support to users

Grid Enabling

- **SAS Display Manager and the SAS Grid Parallel Execution (Examples)**

Submit to Grid from SAS Display Manager (The "Five Lines of Code")

- options metaserver='dnnnn'; options metaport=8561; %let rc=%sysfunc(grdsvc_enable(grid, server=SASApp)); signon grid; rsubmit wait=no persist=no;
- /* insert SAS program here */

Key Definition for Grid Submission

- Save following to external file (c:\gpre.sas for ex.)

```
options noconnectpersist;
options noconnectwait;
options metaserver='dnnnn';
options metaport=8561;
%let rc=%sysfunc(grdsvc_enable(grid, server=SASApp));
signon grid;
rsubmit;
```

- Create new DMS key definition (F12 for ex.)

```
gsubmit "%include „c:\gpre.sas‟;"; rsubmit;
```

Syntax to Enable Grid

Grid function calls

- grdsvc_enable
- grdsvc_nnodes
- grdsvc_getname
- grdsvc_getaddr
- grdsvc_getinfo

SAS Grid Computing for 9.2 doc for complete syntax (available on the SAS Sharepoint site http://vaww.vinci.med.va.gov/vincicentral/projectsites/SASGrid/default.aspx)

Tell SAS to Use the Grid

```
options metaserver=„metadata.sas.com';
options metaport=8561;
%let rc = %sysfunc(grdsvc_enable (_all_,
  server=“explicit project server”));
```

SAS Grid Parallel Execution

Parallelized Workload Balancing

```
%let rc=%sysfunc(grdsvc_enable(_all_, resource=SASMain));
signon task1;
rsubmit task1 wait=no;
        /* code to be remote submitted */
endrsubmit;
signon task2;
rsubmit task2 wait=no;
        /* code to be remote submitted */
endrsubmit;
waitfor _all_ task1 task2;
        /* continue local execution */
```

Grid Enabling Existing SAS Program – Original Serial Program

```
libname a "/u/users/sales";
data a.one;
    do x = 1 to 1000000;
    output; end; run;

libname b "/u/users/sales";
data b.two;
    do y = 1 to 100000;
    output; end; run;
```

Grid Enabling Existing SAS Program – Load Balance Parallel Work Units

```
options metaserver='xxx.yyy.zzz.com‘;
        options metaport=8561;
        %let rc=%sysfunc(grdsvc_enable(_all_, resource=SASMain));
signon host1;
rsubmit wait=no;
         libname a "/u/users/sales";
         data a.one;
           do x = 1 to 1000000;
           output; end; run;
endrsubmit;

signon host2;
rsubmit wait=no;
         libname b "/u/users/sales";
         data b.two
           do y = 1 to 1000000;
           output; end; run;
endrsubmit;
signoff _all_;
```

The Problem

- Lots of existing SAS programs
  - long and complex
  - Written a long time ago
  - Original author no longer available
- Advances in computing architectures and SAS products
  - multi-core and distributed systems
  - SAS metadata, SAS Enterprise Miner/Guide
- Existing programs could be more efficient
  - manual improvements time consuming and costly

The Solution

- SAS code analyzer
  - SAS procedure new in 9.2
  - executes an existing SAS program
  - analyzes job steps, input/output data and dependencies
  - records information used to enhance efficiency and manageability of the program

A Sample Invocation

```
proc scaproc; record 'out.txt';

data a;
     input x y z @@; cards;
1 2 3 4 5 6 7 8 9
run;

proc summary data=a;
     var x;
     output out=new1 mean=mx;
     run;
proc summary data=a;
     var y;
     output out=new2 mean=my;
     run;
proc summary data=a;
     var z;
     output out=new3 mean=mz;
     run;

proc scaproc; write; run;
```

Sample Output

```
/* JOBSPLIT: DATASET OUTPUT SEQ WORK.A.DATA */
/* JOBSPLIT: ELAPSED 62  */
/* JOBSPLIT: PROCNAME DATASTEP */
/* JOBSPLIT: STEP SOURCE FOLLOWS */

data a;
     input x y z @@; cards;
1 2 3 4 5 6 7 8 9
run;
```

```
/* JOBSPLIT: DATASET INPUT SEQ WORK.A.DATA */
/* JOBSPLIT: DATASET OUTPUT SEQ WORK.NEW1.DATA */
/* JOBSPLIT: SYMBOL GET SYSSUMTRACE */
/* JOBSPLIT: ELAPSED 46  */
/* JOBSPLIT: PROCNAME SUMMARY */
/* JOBSPLIT: STEP SOURCE FOLLOWS */
proc summary data=a;
     var x;
     output out=new1 mean=mx;
     run;
```

Benefits

- Parallel execution results in accelerated run-times
- Easy to use tool
  - eliminates costly manual analysis of program flow
  - inserts necessary syntax to eliminate programming errors
  - accelerates learning curve for creating parallel programs

SAS Command-Line Grid Submission Utility

- Standalone utility that will allow user to
  - submit SAS program to grid for processing
  - display status of user‟s jobs on the grid
  - retrieve output from user‟s jobs to local directory
  - kill jobs

Advantages

- User can submit and forget (Batch Jobs)
  - no need to remain connected to process job
- User can view job output while job is running
- Allows for SAS checkpoint/restart capability
- Uses SAS Grid Manager metadata for centralized control

- NOTE - requires shared file system between client and grid

Submitting a Job

- sasgsub –gridsubmitpgm <sas_pgm>
  - other parameters stored in configuration file
    - -GRIDWORK <shared_file_dir>
    - <metadata_connection_parameters>
    - -GRIDAPPSERVER <app_server_name>
    - [-GRIDLICENSEFILE <license_file_pathname>]
    - [-GRIDFILESIN <file_list>]
    - [-GRIDJOBNAME <job_name>]
    - [-GRIDJOBOPTS <job_options>]
    - [-GRIDRESTARTOK]
    - [-GRIDSASOPTS <sas_options_for_job>]
    - [-GRIDWORKLOAD <workload_values>]
    - [-GRIDWORKREM <remote_shared_file_dir>]

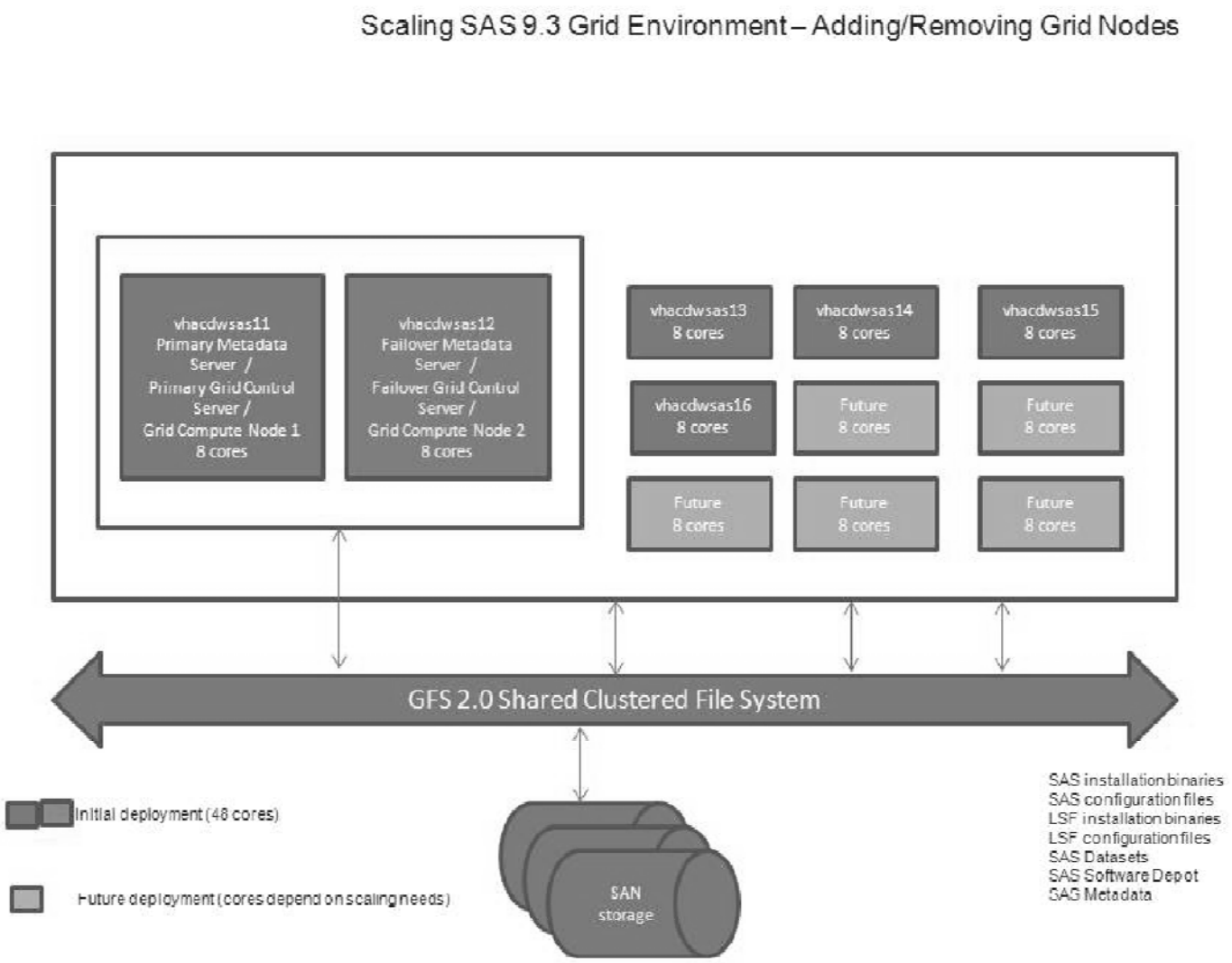
Example Output

Job ID: 6772

Job directory: "/CNT/sasgsub/gridwork/sascnn1/SASGSUB-2009-03-17_14.09.52.847_testPgm"

Job log file: "/CNT/sasgsub/gridwork/sascnn1/SASGSUB-2009-03-17_14.09.52.847_testPgm/testPgm.log“

Currently Installing Linux Grid



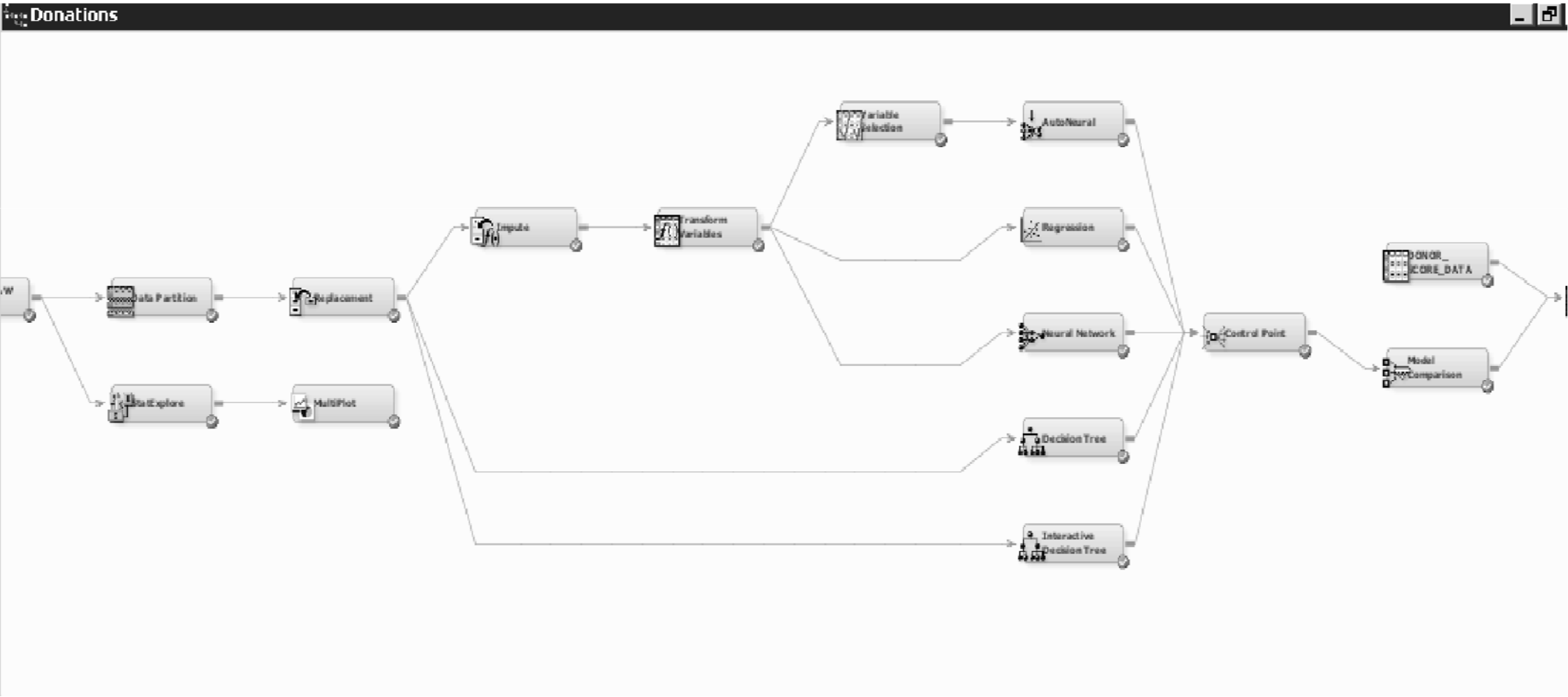
Thank you for attending.

Please contact Mark Ezzo VINCI SAS Administrator:

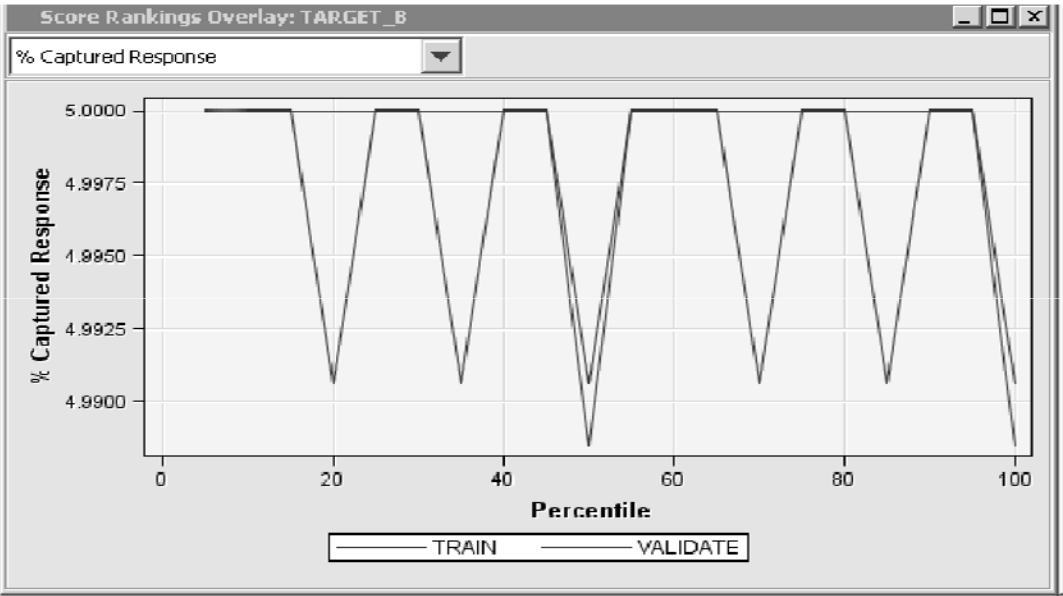
Mark.Ezzo@va.gov

with any questions or comments.

Enterprise Miner Interface

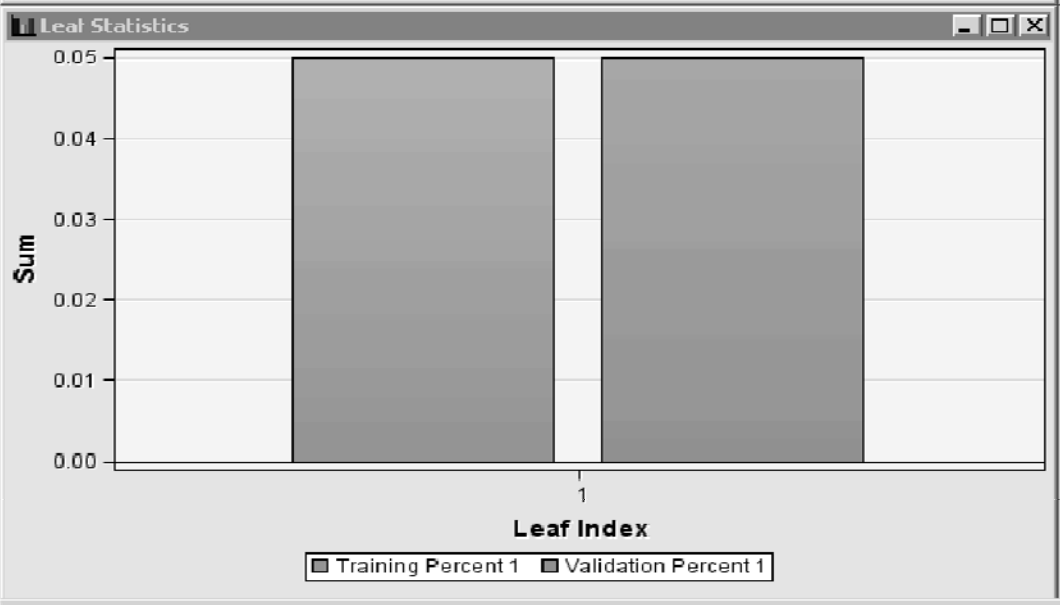


Enterprise Miner Sample Results



Fit Statistics

Target	Fit Statistics	Statistics Label	Train	Validation	Test
TARGET_B	_NOBS_	Sum of Fre...	10654	8718	.
TARGET_B	_SUMW_	Sum of Cas...	21308	17436	.
TARGET_B	_MISC_	Misclassific...	0.250047	0.249943	.
TARGET_B	_MAX_	Maximum A...	0.749953	0.749953	.
TARGET_B	_SSE_	Sum of Squ...	3995.75	3268.75	.
TARGET_B	_ASE_	Average Sq...	0.187523	0.187471	.
TARGET_B	_RASE_	Root Avera...	0.43304	0.43298	.
TARGET_B	_DIV_	Divisor for A...	21308	17436	.
TARGET_B	_DFT_	Total Degre...	10654	.	.
TARGET_B	_APROF_	Average Pr...	1.2	1.2	.
TARGET_B	_PROF_	Total Profit f...	12784.8	10461.6	.
TARGET_B	_PASE_	Average Sq...	0.166869	0.1668	.
TARGET_B	_PMISC_	Misclassific...	0.05	0.05	.



Output

```
*-------------------------------------------------------------*
User:                   vha06martik
Date:                   April 01, 2011
Time:                   13:26:31
*-------------------------------------------------------------*
^ Training Output
*-------------------------------------------------------------*



Variable Summary

                    Measurement     Frequency
Role                   Level          Count

INPUT               BINARY               4
INPUT               INTERVAL            31
INPUT               NOMINAL             12
REJECTED            INTERVAL             1
```