Benchmarks Online

Volume 4 - Number 12 * December 2001

Columns

Other Resources

Feature Articles

Campus Computing News

If $150,000 in funding and funding for summer faculty release time sounds like a good thing to you, you won't want to miss reading all about the 2002-2003 Teaching with Technology Grants in this article.

Winter Break Hours

Read this article to find out what is open and what isn't over the Winter Break.

Forget about Santa Claus; 'GAUSS' is coming to town!

During the winter break, a new server will be brought online which promises to significantly contribute to statistical research here at UNT. Read all about it (picture included)!

EagleMail News

Lots of good things are going on with EagleMail and this article catches you up on all of them. And like it says in the article, having an Eagle Mail account is not just a good idea, it is a requirement for all students.

Lab-of-the-Month

Click on the title above and find out what lucky General Access Lab was profiled this month.

IDE RAID Technology

About Benchmarks *Online*

Subscribe to Benchmarks *Online*

Join Duane Gustavus as he takes a look at a new disk storage array strategy that offers an attractive price/storage point as compared to current commercial SAN or NAS products.

Today's Cartoon

Click on the title above for an information age laugh.

Don't forget to check out our monthly columns. This month's topics:

- **RSS Matters** -- "Dealing with Outliers in Bivariate Data: Robust Correlation." Last time we examined the smoothed bootstrap, this month we demonstrate the calculation of a robust correlation measure.
- **SAS Corner** -- "Data Capability: Transporting data from the Mainframe." Learn how to move SAS data to local platforms from the academic mainframe.
- **The Network Connection** -- "Back to the Basics: Domain Names." Over the last five years, the term "dot-com" has entered into the common vocabulary. Just what does that mean, though?
- **Link of the Month** -- "SmartForce." Got some time on your hands over the Winter Break? Why not do some online learning?
- **WWW@UNT.EDU** -- "How much do you care about security?" Are you concerned about your privacy? This article gives you plenty of reasons to be.
- **Short Courses** -- The Academic Computing Services (ACS) short courses are over for the fall, but that doesn't mean there aren't new items on this page . . .
- **IRC News** -- Minutes of the Information Resources Council are printed here when they are available. The November minutes are contained in this issue.
- **Staff Activities** -- New employees, people

who are no longer employed at the Computing Center, and employee awards and recognitions are included in this article. Then there is the special treat at the end . . .

Research and Statistical Support
University of North Texas

RSS Matters

Dealing with Outliers in Bivariate Data: Robust Correlation

By Dr. Rich Herrington, Research and Statistical Support Consultant

Last time we examined the smoothed bootstrap, this month we demonstrate the calculation of a robust correlation measure. The GNU S language, "R" is used to implement this procedure. R is a statistical programming environment that is a clone of the S and S-Plus language developed at Lucent Technologies. In the following document we illustrate the use of a GNU Web interface to the R engine on the "rss" server, http://rss.acs.unt.edu/cgi-bin/R/Rprog. This GNU Web interface is a derivative of the "Rcgi" Perl scripts available for download from the CRAN Website, http://www.cran.r-project.org (the main "R" Website). Scripts can be submitted interactively, edited, and be re-submitted with changed parameters by selecting the hypertext link buttons that appear below the figures. For example, clicking the "Run Program" button below creates 100 random numbers from a bivariate normal distribution; uses linear regression to fit a least squares line to two of the columns of data; and calculates person's product moment correlation. To view any text output, scroll to the bottom of the browser window. To view the scatterplot, select the "Display Graphic" link. The script can be edited and resubmitted by changing the script in the form window and then selecting "Run the R Program". Selecting the browser "back page" button will return the reader to this document.

```
# n=sample size; p=number of columns
# u=mean of columns; s=standard deviation
# of columns; S=correlation matrix

mvrnorm <- function(n, p, u, s, S) {
Z <- matrix(rnorm(n*p), p, n)
t(u + s*t(chol(S))%*% Z)
}

# Sample from Bivariate Normal

z<-mvrnorm(n=100, p=2, u=c(100,100), s=c(15,15), S=matrix(c(1, .4, .4, 1), ncol=2,
nrow=2,byrow=T))

# Fit least squares regression

z.fit<-lm(z[,1]~z[,2])
z.fit

# Plot scatterplot
plot(z[,1], z[,2])

# Plot best fit line
abline(z.fit)

# Calculate Pearson's Correlation
cor(z[,1], z[,2])
```

Run Program

Pearson's Product Moment Correlation

Let $(X_1,Y_1),\ldots(X_n,Y_n)$ be a random sample from a bivariate distribution. The sample estimate of the population correlation ρ is given by:

$$r=\frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum(X_i-\bar{X})^2\sum(Y_i-\bar{Y})^2}}$$

If X and Y are independent, $\rho=0$. A test of $H_0:\rho=0$ is based on the test statistic:

$$T=r\sqrt{\frac{n-2}{1-r^2}}$$

where n is the number of X,Y pairs of scores. If H_0 is true, T has a Student's t distribution with df = n-2 degrees of freedom if at least one of the marginal distributions is normal. Reject $H_0:\rho=0$ if $|T|>t_{1-\alpha/2}$, for the $1-\alpha/2$ quantile of Student's t distribution with n-2

degrees of freedom. When the null hypothesis is true, the distribution of sample correlation coefficients tends to be normally distributed for increasing sample size. The standard error of this distribution of sample correlations is approximately:

$$\sigma_r = \frac{1}{\sqrt{N}}$$

When the sample size is reasonably large ($N \geq 50$), then it is possible to test the significance of the sample correlation coefficient by forming the usual z statistic and referring it to the normal distribution. In situations where one wishes to test for a non-zero value of the sample correlation coefficient, many sources have recommended Fisher's r-toZ transformation when computing confidence intervals, but it is not asymptotically correct when sampling from nonnormal distributions. Moreover, simulation studies do not support Fisher's r-to-Z transformation for small sample sizes. The S script below samples from a bivariate normal distribution whose population correlation is .40. A scatterplot is plotted with a best fit regression line added to the scatterplot. A classical t-test of the correlation coefficient is performed, and the standard error under the null hypothesis is calculated. Change the script below to have different sample sizes and see the effect of sample size on the standard error; also see how the sample correlation fluctuates from sample to sample around the true population correlation value. With smaller sample sizes, the sample correlation coefficient can be quite different from the population value.

```
mvrnorm <- function(n, p, u, s, S) {
Z <- matrix(rnorm(n * p), p, n)
t(u + s*t(chol(S))%*% Z)
}

z<-mvrnorm(n=100, p=2, u=c(100,100), s=c(15,15), S=matrix(c(1,.4, .4, 1), ncol=2,
nrow=2, byrow=T))

z.fit<-lm(z[,1]~z[,2])

plot(z[,1], z[,2])
abline(z.fit)

z.cor<-cor.test(z[,1], z[,2])
stderr<-1/(length(z[,1]))^.5
stderr
z.cor
```

Run Program

Robust Correlation: The Biweight Midcorrelation

For an introduction to robust statistics and robust measures of location, see the July issue of Benchmarks. Let $(X_1, Y_1), \ldots (X_n, Y_n)$ be a random sample from a bivariate distribution. Following Wilcox (1997, page 197), let:

$$U_i = \frac{X_i - M_x}{9 \cdot MAD_x}$$

where M_x is the median, and MAD_x is the median absolute deviation. A sample estimate of the median absolute deviation is given by:

$$MAD_x = MEDIAN\left\{\left|X_1 - M_x\right|, \ldots\ldots \left|X_i - M_x\right|\right\}$$

where M_x is the sample median. MAD has a finite sample breakdown point of approximately .5. Let V_i be a function of the Y_i scores as is U_i is for the X_i scores. The sample biweight midcovariance between X and Y is given by:

$$s_{bxy} = \frac{n\sum a_i (X_i - M_x)(1-U_i^2)^2 b_i (Y_i - M_y)(1-V_i^2)^2}{(\sum a_i (1-U_i^2)(1-5U_i^2))(\sum b_i (1-V_i^2)(1-5V_i^2))}$$

where,

$$a_i = 1 \text{ if } -1 \le U_i \le 1, \text{ otherwise } a_i = 0$$

and

$$b_i = 1 \text{ if } -1 \le V_i \le 1, \text{ otherwise } b_i = 0.$$

An estimate of the biweight midcorrelation between X and Y is given by:

$$r_b = \frac{s_{bxy}}{\sqrt{s_{bxx} s_{byy}}}$$

where s_{bxx} and s_{byy} are the biweight midvariance for the X and Y scores. The S script below calculates both the Pearson product-moment correlation and the biweight midcorrelation for a bivariate normal distribution whose population correlation is .40. An outlier is added to the sample, and then the two correlation coefficients are recalculated and compared again. A scatterplot with a best fit line is plotted after the outlier is added. Try changing various parameters and resubmitting the S script.

```
#### Calculate biweight midcovariance
####   from Wilcox (1997).

bicov<-function(x, y){
mx <- median(x)
my <-median(y)
ux <- abs((x - mx)/(9 * qnorm(0.75) * mad(x)))
uy <- abs((y - my)/(9 * qnorm(0.75) * mad(y)))
aval <- ifelse(ux <= 1, 1, 0)
bval <- ifelse(uy <= 1, 1, 0)
top <- sum(aval * (x - mx) * (1 - ux^2)^2 * bval * (y - my) * (1 - uy^2)^2)
top <- length(x) * top
botx <- sum(aval * (1 - ux^2) * (1 - 5 * ux^2))
boty <- sum(bval * (1 - uy^2) * (1 - 5 * uy^2))
bi <- top/(botx * boty)
bi
}

#### Calculate biweight midcorrelation.

bicor<-function(x,y){
x <-as.numeric(x)
y<-as.numeric(y)
bicov(x,y)/(sqrt(bicov(x,x)*bicov(y,y)))
}

#### Sample from multivariate normal

mvrnorm <- function(n, p, u, s, S) {
Z <- matrix(rnorm(n * p), p, n)
t(u + s*t(chol(S))%*% Z)
}

z <- mvrnorm(n=100, p=2, u=c(100,100), s=c(15,15), S=matrix(c(1, .4, .4, 1), ncol=2, nrow=2, byrow=T))
```

```
#### Calculate biweight midcovariance
####   from Wilcox (1997).

bicov<-function(x, y){
mx <- median(x)
my <-median(y)
ux <- abs((x - mx)/(9 * qnorm(0.75) * mad(x)))
uy <- abs((y - my)/(9 * qnorm(0.75) * mad(y)))
aval <- ifelse(ux <= 1, 1, 0)
bval <- ifelse(uy <= 1, 1, 0)
top <- sum(aval * (x - mx) * (1 - ux^2)^2 * bval * (y - my) * (1 - uy^2)^2)
top <- length(x) * top
botx <- sum(aval * (1 - ux^2) * (1 - 5 * ux^2))
boty <- sum(bval * (1 - uy^2) * (1 - 5 * uy^2))
bi <- top/(botx * boty)
bi
}

#### Calculate biweight midcorrelation.

bicor<-function(x,y){
x <-as.numeric(x)
y<-as.numeric(y)
bicov(x,y)/(sqrt(bicov(x,x)*bicov(y,y)))
}

#### Sample from multivariate normal

mvrnorm <- function(n, p, u, s, S) {
Z <- matrix(rnorm(n * p), p, n)
t(u + s*t(chol(S))%*% Z)
}

z <- mvrnorm(n=100, p=2, u=c(100,100), s=c(15,15), S=matrix(c(1, .4, .4, 1), ncol=2,
nrow=2, byrow=T))

# Pearson Correlation and Robust Correlation
# before outlier added

cor(z[,1], z[,2])
bicor(z[,1], z[,2])

# Add outlier to Y vector for smallest X value

z[z[,1]==min(z[,1]),2]<-z[z[,1]==min(z[,1]),2]*2.5
z.fit<-lm(z[,1]~z[,2])
plot(z[,1], z[,2])
abline(z.fit)

# Pearson Correlation and Robust Correlation
# after outlier added

cor(z[,1], z[,2])
bicor(z[,1], z[,2])
```

```
#### Calculate biweight midcovariance
####   from Wilcox (1997).
```

Run Program

Calculating Empirical P-values, Empirical Power, and Confidence Intervals for a Robust Correlation Coefficient

The S script below calculates observed p-values, power, and confidence intervals for the biweight midcorrelation coefficient. A comparison is made between the Pearson correlation coefficient and the biweight midcorrelation before and after an outlier is added. Confidence intervals are calculated by simulating the null sampling distribution, and also by sampling from the alternate sampling distribution (Hall & Wilson, 1991). For details on calculating the percentile bootstrap, and calculating power using the percentile bootstrap, see the September issue of Benchmarks. For details on the percentile bootstrap see the August issue of Benchmarks. After running the following S script, try changing various parameters: sample size, population correlation, correlation coefficient (assign "est" to have value "bicor" or "cor") and the size of the outlier (change the multiplying factor 2.5 to a smaller value). Changing "est" to have value "cor" will allow one to calculate the p-value and power of the observed sample for the Pearson product-moment correlation.

```
# Function for extracting M-Estimator from the MASS Library
# Calculating Empirical P-values,
# Empirical Power, and Confidence
# Intervals for a Robust Correlation
# Coefficient Population parameters

pop.cor<-.40
pop1.mean<-100
pop1.stdev<-15
pop2.mean<-100
pop2.stdev<-15
sample.size<-40

# Sample from bivariate normal distribution
# with given parameters

# Set random number generator seed
set.seed(12)

z<-mvrnorm(n=sample.size, p=2, u=c(pop1.mean,pop2.mean), s=c(pop1.stdev,pop2.stdev), S=matrix(c(1, pop.cor, pop.cor, 1), ncol=2, nrow=2, byrow=T))

# Correlation before addition of outlier

cor.nolie<-cor(z[,1], z[,2])
cor.nolie.test<-cor.test(z[,1], z[,2])
bicor.nolie<-bicor(z[,1], z[,2])

# Add outlier to Y vector for smallest X value

z[z[,1]==min(z[,1]),2]<-z[z[,1]==min(z[,1]),2]*2.5

#######################################################
# Assign data vectors

z.x<-z[,1]
z.y<-z[,2]

### Set number of bootstrap samples
nboot<-1000

### Set Estimation Method
est<-bicor

# Select "cor" for Pearson r
# Select "bicor" for
# Robust correlation based on bi-weight covariance

# Calculate observed effect

cor.empirical<-cor(z.x, z.y)
diff.empirical<-est(z.x, z.y)

#################################
# Sampling from H0: sample columns independently
# to simulate correlation coefficient under H0

h0datax.list<-vector("list", nboot)

for (i in 1:nboot){
h0datax.list[[i]]<-sample(z.x, size = length(z.x), replace=T)
}

h0datay.list<-vector("list", nboot)

for (i in 1:nboot){
h0datay.list[[i]]<-sample(z.y, size = length(z.y), replace=T)
}

# Calculate correlation for each bootstrap sample
# under H0 sampling distribution

h0bvec<-unlist(lapply(1:nboot, function(i, x, y)est(x[[i]], y[[i]]), x = h0datax.list, y = h0datay.list))
mean.h0bvec<-mean(h0bvec)
stdev.h0bvec<-var(h0bvec)^.5

# Calculate Critical cutoffs on H0 Sampling distribution

critlow<-quantile(h0bvec,.025)
critup<-quantile(h0bvec,.975)

# Calculate Confidence Intervals from H0;
# shift null sampling distribution by observed effect

h0.ci <- 0
h0.ci[1] <- critlow + diff.empirical
h0.ci[2] <- critup + diff.empirical

#### Calculate P-value for H0
count<-length(h0bvec[abs(h0bvec)>=abs(diff.empirical)])
pvalue.empirical<-count/nboot

#################################
# Sample from H1: Sample with rows
# intact; bootstrap sampling of row indices

data.index <- matrix(sample(length(z.x), size = length(z.x) * nboot, replace = T), nrow = nboot)

# Use bootstrap sampled row indices # to extract the same row from both X
```

```
# Function for extracting M-Estimator from the MASS Library
# Calculating Empirical P-values,
# Empirical Power, and Confidence
# Intervals for a Robust Correlation
# Coefficient Population parameters

pop.cor<-.40
pop1.mean<-100
pop1.stdev<-15
pop2.mean<-100
pop2.stdev<-15
sample.size<-40

# Sample from bivariate normal distribution
# with given parameters

# Set random number generator seed
set.seed(12)

z<-mvrnorm(n=sample.size, p=2, u=c(pop1.mean,pop2.mean), s=c(pop1.stdev,pop2.stdev), S=matrix(c(1, pop.cor, pop.cor, 1), ncol=2, nrow=2, byrow=T))

# Correlation before addition of outlier

cor.nolie<-cor(z[,1], z[,2])
cor.nolie.test<-cor.test(z[,1], z[,2])
bicor.nolie<-bicor(z[,1], z[,2])

# Add outlier to Y vector for smallest X value

z[z[,1]==min(z[,1]),2]<-z[z[,1]==min(z[,1]),2]*2.5

#######################################################
# Assign data vectors

z.x<-z[,1]
z.y<-z[,2]

### Set number of bootstrap samples
nboot<-1000

### Set Estimation Method
est<-bicor

# Select "cor" for Pearson r
# Select "bicor" for
# Robust correlation based on bi-weight covariance

# Calculate observed effect

cor.empirical<-cor(z.x, z.y)
diff.empirical<-est(z.x, z.y)

#################################
# Sampling from H0: sample columns independently
# to simulate correlation coefficient under H0

h0datax.list<-vector("list", nboot)

for (i in 1:nboot){
h0datax.list[[i]]<-sample(z.x, size = length(z.x), replace=T)
}

h0datay.list<-vector("list", nboot)

for (i in 1:nboot){
h0datay.list[[i]]<-sample(z.y, size = length(z.y), replace=T)
}

# Calculate correlation for each bootstrap sample
# under H0 sampling distribution

h0bvec<-unlist(lapply(1:nboot, function(i, x, y)est(x[[i]], y[[i]]), x = h0datax.list, y = h0datay.list))
mean.h0bvec<-mean(h0bvec)
stdev.h0bvec<-var(h0bvec)^.5

# Calculate Critical cutoffs on H0 Sampling distribution

critlow<-quantile(h0bvec,.025)
critup<-quantile(h0bvec,.975)

# Calculate Confidence Intervals from H0;
# shift null sampling distribution by observed effect

h0.ci <- 0
h0.ci[1] <- critlow + diff.empirical
h0.ci[2] <- critup + diff.empirical

#### Calculate P-value for H0
count<-length(h0bvec[abs(h0bvec)>=abs(diff.empirical)])
pvalue.empirical<-count/nboot

#################################
# Sample from H1: Sample with rows
# intact; bootstrap sampling of row indices

data.index <- matrix(sample(length(z.x), size = length(z.x) * nboot, replace = T), nrow = nboot)

# Use bootstrap sampled row indices # to extract the same row from both X
# and Y vector

datax <- matrix(z.x[data.index], nrow = nboot, ncol=length(z.x))
datay <- matrix(z.y[data.index], nrow = nboot, ncol=length(z.y))
h1datax.list<-vector("list", nboot)

for (i in 1:nboot){h1datax.list[[i]]<-datax[i,]}

h1datay.list<-vector("list", nboot)

for (i in 1:nboot){h1datay.list[[i]]<-datay[i,]}

# Calculate the correlation under H1 sampling
# distribution

h1bvec<-unlist(lapply(1:nboot, function(i, x, y) est(x[[i]], y[[i]]), x = h1datax.list, y = h1datay.list))
mean.h1bvec<-mean(h1bvec)
stdev.h1bvec<-var(h1bvec)^.5

# Calculate confidence Intervals from H1

effectlow<-quantile(h1bvec,.025)
effectup<-quantile(h1bvec,.975)
h1.ci <- 0
h1.ci[1] <- effectlow
h1.ci[2] <- effectup

# Calculate Bootstrap Empirical Power
countup<-length(h1bvec[h1bvec>=critup])
countlow<-length(h1bvec[h1bvec<=critlow])
power.twotail<-(countup+countlow)/nboot

##### Report Results
# Pearson Correlation and bi-weight midcorrelation
# before outlier is added

cor.nolie
cor.nolie.test
bicor.nolie

# Results after adding outlier
# Mean and Standard Deviation (standard error)
# of Bootstrap Samples for Robust Correlation under H0

mean.h0bvec
stdev.h0bvec

# Normal theory standard error

1/(length(z.x)^.5)

# Mean and Standard Deviation (standard error)
# of Bootstrap Samples for Robust Correlation under H1

mean.h1bvec
stdev.h1bvec
mean.h1bvec-pop.cor

### Pearson r and Bias

cor.empirical
cor.empirical-pop.cor

### Robust Correlation and Bias
```

```
# Function for extracting M-Estimator from the MASS Library
# Calculating Empirical P-values,
# Empirical Power, and Confidence
# Intervals for a Robust Correlation
# Coefficient Population parameters

pop.cor<-.40
pop1.mean<-100
pop1.stdev<-15
pop2.mean<-100
pop2.stdev<-15
sample.size<-40

# Sample from bivariate normal distribution
# with given parameters

# Set random number generator seed
set.seed(12)

z<-mvrnorm(n=sample.size, p=2, u=c(pop1.mean,pop2.mean), s=c(pop1.stdev,pop2.stdev), S=matrix(c(1, pop.cor, pop.cor, 1), ncol=2, nrow=2, byrow=T))

# Correlation before addition of outlier

cor.nolie<-cor(z[,1], z[,2])
cor.nolie.test<-cor.test(z[,1], z[,2])
bicor.nolie<-bicor(z[,1], z[,2])

# Add outlier to Y vector for smallest X value

z[z[,1]==min(z[,1]),2]<-z[z[,1]==min(z[,1]),2]*2.5

#######################################################
# Assign data vectors

z.x<-z[,1]
z.y<-z[,2]

### Set number of bootstrap samples
nboot<-1000

### Set Estimation Method
est<-bicor

# Select "cor" for Pearson r
# Select "bicor" for
# Robust correlation based on bi-weight covariance

# Calculate observed effect

cor.empirical<-cor(z.x, z.y)
diff.empirical<-est(z.x, z.y)

#################################
# Sampling from H0: sample columns independently
# to simulate correlation coefficient under H0

h0datax.list<-vector("list", nboot)

for (i in 1:nboot){
h0datax.list[[i]]<-sample(z.x, size = length(z.x), replace=T)
}

h0datay.list<-vector("list", nboot)

for (i in 1:nboot){
h0datay.list[[i]]<-sample(z.y, size = length(z.y), replace=T)
}

# Calculate correlation for each bootstrap sample
# under H0 sampling distribution

h0bvec<-unlist(lapply(1:nboot, function(i, x, y)est(x[[i]], y[[i]]), x = h0datax.list, y = h0datay.list))
mean.h0bvec<-mean(h0bvec)
stdev.h0bvec<-var(h0bvec)^.5

# Calculate Critical cutoffs on H0 Sampling distribution

critlow<-quantile(h0bvec,.025)
critup<-quantile(h0bvec,.975)

# Calculate Confidence Intervals from H0;
# shift null sampling distribution by observed effect

h0.ci <- 0
h0.ci[1] <- critlow + diff.empirical
h0.ci[2] <- critup + diff.empirical

#### Calculate P-value for H0
count<-length(h0bvec[abs(h0bvec)>=abs(diff.empirical)])
pvalue.empirical<-count/nboot

#################################
# Sample from H1: Sample with rows
# intact; bootstrap sampling of row indices

data.index <- matrix(sample(length(z.x), size = length(z.x) * nboot, replace = T), nrow = nboot)

# Use bootstrap sampled row indices # to extract the same row from both X
# and Y vector

datax <- matrix(z.x[data.index], nrow = nboot, ncol=length(z.x))
datay <- matrix(z.y[data.index], nrow = nboot, ncol=length(z.y))
h1datax.list<-vector("list", nboot)

for (i in 1:nboot){h1datax.list[[i]]<-datax[i,]}

h1datay.list<-vector("list", nboot)

for (i in 1:nboot){h1datay.list[[i]]<-datay[i,]}

# Calculate the correlation under H1 sampling
# distribution

h1bvec<-unlist(lapply(1:nboot, function(i, x, y) est(x[[i]], y[[i]]), x = h1datax.list, y = h1datay.list))
mean.h1bvec<-mean(h1bvec)
stdev.h1bvec<-var(h1bvec)^.5

# Calculate confidence Intervals from H1

effectlow<-quantile(h1bvec,.025)
effectup<-quantile(h1bvec,.975)
h1.ci <- 0
h1.ci[1] <- effectlow
h1.ci[2] <- effectup

# Calculate Bootstrap Empirical Power
countup<-length(h1bvec[h1bvec>=critup])
countlow<-length(h1bvec[h1bvec<=critlow])
power.twotail<-(countup+countlow)/nboot

##### Report Results
# Pearson Correlation and bi-weight midcorrelation
# before outlier is added

cor.nolie
cor.nolie.test
bicor.nolie

# Results after adding outlier
# Mean and Standard Deviation (standard error)
# of Bootstrap Samples for Robust Correlation under H0

mean.h0bvec
stdev.h0bvec

# Normal theory standard error

1/(length(z.x)^.5)

# Mean and Standard Deviation (standard error)
# of Bootstrap Samples for Robust Correlation under H1

mean.h1bvec
stdev.h1bvec
mean.h1bvec-pop.cor

### Pearson r and Bias

cor.empirical
cor.empirical-pop.cor

### Robust Correlation and Bias
```

```
# Function for extracting M-Estimator from the MASS Library
# Calculating Empirical P-values,
# Empirical Power, and Confidence
# Intervals for a Robust Correlation
# Coefficient Population parameters

pop.cor<-.40
pop1.mean<-100
pop1.stdev<-15
pop2.mean<-100
pop2.stdev<-15
sample.size<-40

# Sample from bivariate normal distribution
# with given parameters

# Set random number generator seed
set.seed(12)

z<-mvrnorm(n=sample.size, p=2, u=c(pop1.mean,pop2.mean), s=c(pop1.stdev,pop2.stdev), S=matrix(c(1, pop.cor, pop.cor, 1), ncol=2, nrow=2, byrow=T))

# Correlation before addition of outlier

cor.nolie<-cor(z[,1], z[,2])
cor.nolie.test<-cor.test(z[,1], z[,2])
bicor.nolie<-bicor(z[,1], z[,2])

# Add outlier to Y vector for smallest X value

z[z[,1]==min(z[,1]),2]<-z[z[,1]==min(z[,1]),2]*2.5

#######################################################
# Assign data vectors

z.x<-z[,1]
z.y<-z[,2]

### Set number of bootstrap samples
nboot<-1000

### Set Estimation Method
est<-bicor

# Select "cor" for Pearson r
# Select "bicor" for
# Robust correlation based on bi-weight covariance

# Calculate observed effect

cor.empirical<-cor(z.x, z.y)
diff.empirical<-est(z.x, z.y)

#################################
# Sampling from H0: sample columns independently
# to simulate correlation coefficient under H0

h0datax.list<-vector("list", nboot)

for (i in 1:nboot){
h0datax.list[[i]]<-sample(z.x, size = length(z.x), replace=T)
}

h0datay.list<-vector("list", nboot)

for (i in 1:nboot){
h0datay.list[[i]]<-sample(z.y, size = length(z.y), replace=T)
}

# Calculate correlation for each bootstrap sample
# under H0 sampling distribution

h0bvec<-unlist(lapply(1:nboot, function(i, x, y)est(x[[i]], y[[i]]), x = h0datax.list, y = h0datay.list))
mean.h0bvec<-mean(h0bvec)
stdev.h0bvec<-var(h0bvec)^.5

# Calculate Critical cutoffs on H0 Sampling distribution

critlow<-quantile(h0bvec,.025)
critup<-quantile(h0bvec,.975)

# Calculate Confidence Intervals from H0;
# shift null sampling distribution by observed effect

h0.ci <- 0
h0.ci[1] <- critlow + diff.empirical
h0.ci[2] <- critup + diff.empirical

#### Calculate P-value for H0
count<-length(h0bvec[abs(h0bvec)>=abs(diff.empirical)])
pvalue.empirical<-count/nboot

#################################
# Sample from H1: Sample with rows
# intact; bootstrap sampling of row indices

data.index <- matrix(sample(length(z.x), size = length(z.x) * nboot, replace = T), nrow = nboot)

# Use bootstrap sampled row indices # to extract the same row from both X
# and Y vector

datax <- matrix(z.x[data.index], nrow = nboot, ncol=length(z.x))
datay <- matrix(z.y[data.index], nrow = nboot, ncol=length(z.y))
h1datax.list<-vector("list", nboot)

for (i in 1:nboot){h1datax.list[[i]]<-datax[i,]}

h1datay.list<-vector("list", nboot)

for (i in 1:nboot){h1datay.list[[i]]<-datay[i,]}

# Calculate the correlation under H1 sampling
# distribution

h1bvec<-unlist(lapply(1:nboot, function(i, x, y) est(x[[i]], y[[i]]), x = h1datax.list, y = h1datay.list))
mean.h1bvec<-mean(h1bvec)
stdev.h1bvec<-var(h1bvec)^.5

# Calculate confidence Intervals from H1

effectlow<-quantile(h1bvec,.025)
effectup<-quantile(h1bvec,.975)
h1.ci <- 0
h1.ci[1] <- effectlow
h1.ci[2] <- effectup

# Calculate Bootstrap Empirical Power
countup<-length(h1bvec[h1bvec>=critup])
countlow<-length(h1bvec[h1bvec<=critlow])
power.twotail<-(countup+countlow)/nboot

##### Report Results
# Pearson Correlation and bi-weight midcorrelation
# before outlier is added

cor.nolie
cor.nolie.test
bicor.nolie

# Results after adding outlier
# Mean and Standard Deviation (standard error)
# of Bootstrap Samples for Robust Correlation under H0

mean.h0bvec
stdev.h0bvec

# Normal theory standard error

1/(length(z.x)^.5)

# Mean and Standard Deviation (standard error)
# of Bootstrap Samples for Robust Correlation under H1

mean.h1bvec
stdev.h1bvec
mean.h1bvec-pop.cor

### Pearson r and Bias

cor.empirical
cor.empirical-pop.cor

### Robust Correlation and Bias
```

```
# Function for extracting M-Estimator from the MASS Library
# Calculating Empirical P-values,
# Empirical Power, and Confidence
# Intervals for a Robust Correlation
# Coefficient Population parameters

pop.cor<-.40
pop1.mean<-100
pop1.stdev<-15
pop2.mean<-100
pop2.stdev<-15
sample.size<-40

# Sample from bivariate normal distribution
# with given parameters

# Set random number generator seed
set.seed(12)

z<-mvrnorm(n=sample.size, p=2, u=c(pop1.mean,pop2.mean), s=c(pop1.stdev,pop2.stdev), S=matrix(c(1, pop.cor, pop.cor, 1), ncol=2, nrow=2, byrow=T))

# Correlation before addition of outlier

cor.nolie<-cor(z[,1], z[,2])
cor.nolie.test<-cor.test(z[,1], z[,2])
bicor.nolie<-bicor(z[,1], z[,2])

# Add outlier to Y vector for smallest X value

z[z[,1]==min(z[,1]),2]<-z[z[,1]==min(z[,1]),2]*2.5

#######################################################
# Assign data vectors

z.x<-z[,1]
z.y<-z[,2]

### Set number of bootstrap samples
nboot<-1000

### Set Estimation Method
est<-bicor

# Select "cor" for Pearson r
# Select "bicor" for
# Robust correlation based on bi-weight covariance

# Calculate observed effect

cor.empirical<-cor(z.x, z.y)
diff.empirical<-est(z.x, z.y)

#################################
# Sampling from H0: sample columns independently
# to simulate correlation coefficient under H0

h0datax.list<-vector("list", nboot)

for (i in 1:nboot){
h0datax.list[[i]]<-sample(z.x, size = length(z.x), replace=T)
}

h0datay.list<-vector("list", nboot)

for (i in 1:nboot){
h0datay.list[[i]]<-sample(z.y, size = length(z.y), replace=T)
}

# Calculate correlation for each bootstrap sample
# under H0 sampling distribution

h0bvec<-unlist(lapply(1:nboot, function(i, x, y)est(x[[i]], y[[i]]), x = h0datax.list, y = h0datay.list))
mean.h0bvec<-mean(h0bvec)
stdev.h0bvec<-var(h0bvec)^.5

# Calculate Critical cutoffs on H0 Sampling distribution

critlow<-quantile(h0bvec,.025)
critup<-quantile(h0bvec,.975)

# Calculate Confidence Intervals from H0;
# shift null sampling distribution by observed effect

h0.ci <- 0
h0.ci[1] <- critlow + diff.empirical
h0.ci[2] <- critup + diff.empirical

#### Calculate P-value for H0
count<-length(h0bvec[abs(h0bvec)>=abs(diff.empirical)])
pvalue.empirical<-count/nboot

#################################
# Sample from H1: Sample with rows
# intact; bootstrap sampling of row indices

data.index <- matrix(sample(length(z.x), size = length(z.x) * nboot, replace = T), nrow = nboot)

# Use bootstrap sampled row indices # to extract the same row from both X
# and Y vector

datax <- matrix(z.x[data.index], nrow = nboot, ncol=length(z.x))
datay <- matrix(z.y[data.index], nrow = nboot, ncol=length(z.y))
h1datax.list<-vector("list", nboot)

for (i in 1:nboot){h1datax.list[[i]]<-datax[i,]}

h1datay.list<-vector("list", nboot)

for (i in 1:nboot){h1datay.list[[i]]<-datay[i,]}

# Calculate the correlation under H1 sampling
# distribution

h1bvec<-unlist(lapply(1:nboot, function(i, x, y) est(x[[i]], y[[i]]), x = h1datax.list, y = h1datay.list))
mean.h1bvec<-mean(h1bvec)
stdev.h1bvec<-var(h1bvec)^.5

# Calculate confidence Intervals from H1

effectlow<-quantile(h1bvec,.025)
effectup<-quantile(h1bvec,.975)
h1.ci <- 0
h1.ci[1] <- effectlow
h1.ci[2] <- effectup

# Calculate Bootstrap Empirical Power
countup<-length(h1bvec[h1bvec>=critup])
countlow<-length(h1bvec[h1bvec<=critlow])
power.twotail<-(countup+countlow)/nboot

##### Report Results
# Pearson Correlation and bi-weight midcorrelation
# before outlier is added

cor.nolie
cor.nolie.test
bicor.nolie

# Results after adding outlier
# Mean and Standard Deviation (standard error)
# of Bootstrap Samples for Robust Correlation under H0

mean.h0bvec
stdev.h0bvec

# Normal theory standard error

1/(length(z.x)^.5)

# Mean and Standard Deviation (standard error)
# of Bootstrap Samples for Robust Correlation under H1

mean.h1bvec
stdev.h1bvec
mean.h1bvec-pop.cor

### Pearson r and Bias

cor.empirical
cor.empirical-pop.cor

### Robust Correlation and Bias
```

```
# Function for extracting M-Estimator from the MASS Library
# Calculating Empirical P-values,
```

Run Program

Results

The output from the S script above are listed below.

```
> #####   Report Results
>
> # Pearson Correlation  and  bi-weight midcorrelation
> # before outlier is added
>
> cor.nolie
[1] 0.4094627
> cor.nolie.test

        Pearson's product-moment correlation

data:  z[, 1] and z[, 2]
t = 2.7667, df = 38, p-value = 0.008698
alternative hypothesis: true correlation is not equal to 0
sample estimates:
      cor
0.4094627

> bicor.nolie
[1] 0.4222197
```

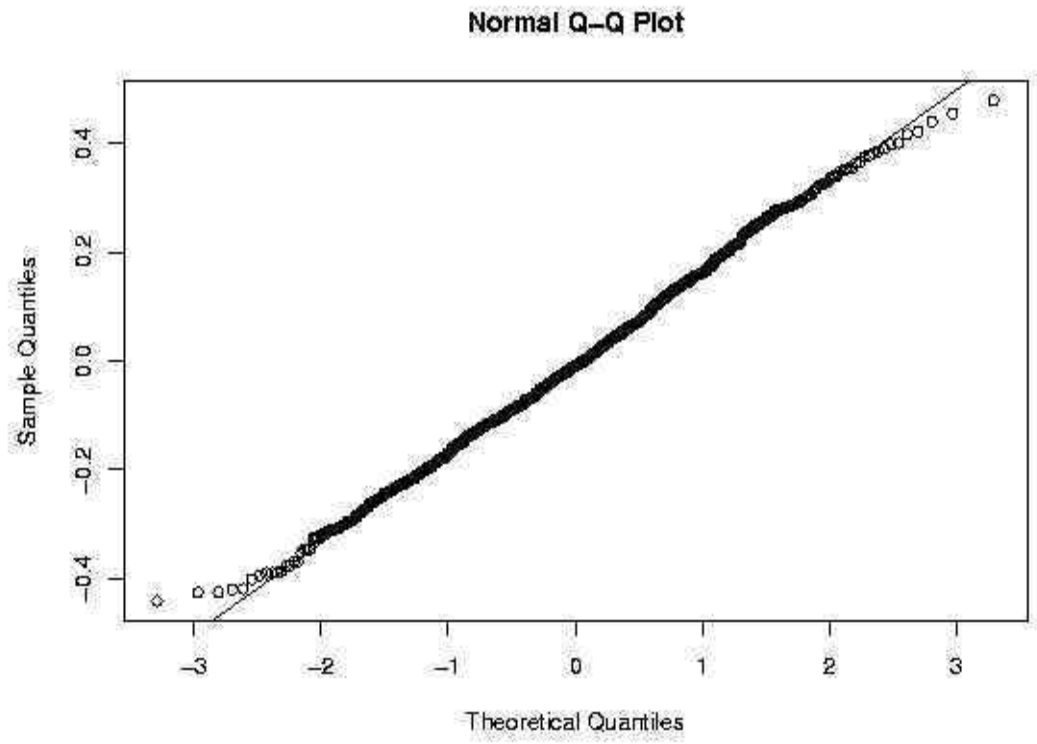
```
> # Results after adding outlier
>
> # Mean and Standard Deviation (standard error)
> # of Bootstrap Samples for Robust Correlation under HO
>
> mean.hObvec
[1] -0.003922042
> stdev.hObvec
[1] 0.1657244
>
> # normal theory standard error
> 1/(length(z.x)^.5)
[1] 0.1581139
>
> # Mean and Standard Deviation (standard error)
> # of Bootstrap Samples for Robust Correlation under H1
>
> mean.h1bvec
[1] 0.3206665
> stdev.h1bvec
[1] 0.1531287
> mean.h1bvec-pop.cor
[1] -0.07933351
```

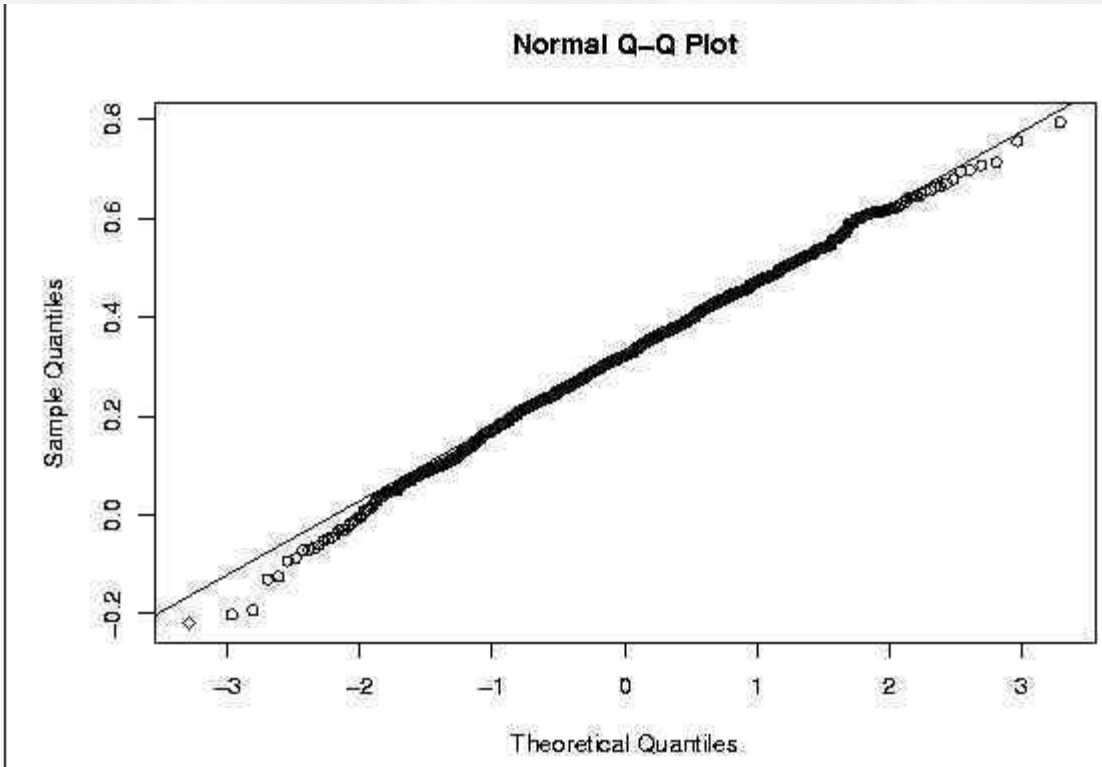
```
> ### Pearson r and Bias
>
> cor.empirical
[1] -0.07661869
> cor.empirical-pop.cor
[1] -0.4766187
>
> ### Robust Correlation and Bias
>
> diff.empirical
[1] 0.3373465
> diff.empirical-mean.h1bvec
[1] 0.01667996
> diff.empirical-pop.cor
[1] -0.06265355
```

```
> # Bootstrap empirical power for two-tail test
> # of the robust correlation
>
> power.twotail
[1] 0.494
>
> # Bootstrap empirical p-value for the
> # robust correlation
>
> pvalue.empirical
[1] 0.041
```

```
> # Two alternative ways of calculating confidence
> # intervals for Robust Correlation (H0 is preferred)
>
> # Confidence intervals based on the
> # bootstrap sampling distribution for the null
> # and alternate sampling distribution
>
> list(h0.ci = h0.ci)
$h0.ci
[1] 0.02247948 0.66213757

> list(h1.ci=h1.ci)
$h1.ci
[1] 0.009956474 0.614598842
```


Normal Q–Q Plot
Sample Quantiles
Theoretical Quantiles
−0.4
−0.2
0.0
0.2
0.4
−3
−2
−1
0
1
2
3


Normal Q–Q Plot
Sample Quantiles
Theoretical Quantiles
−0.2
0.0
0.2
0.4
0.6
0.8
−3
−2
−1
0
1
2
3

References

Hall P., & Wilson S.R. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics, 47,* 757-762.

Wilcox, Rand (1997). *Introduction to Robust Estimation and Hypothesis Testing*. Academic Press, New York.

Research and Statistical Support
University of North Texas

SAS Corner

By Dr. Karl Ho, Research and Statistical Support Services Manager

Data Capability: Transporting data from the Mainframe

In the 2001 November issue of *Benchmarks Online*, Dr. Baczewski pronounced the shutdown of the academic mainframe in 2003 (proposed, read original). It is never too early to start packing up. In this article, I introduce the method of moving SAS data to local platforms. There are many ways to do that but the following method is relatively easier to implement and safe for data.

1. Converting to transport file

Since transporting data involves using file transfer programs, it is better to convert the SAS system files into a portable format so the transportation process will not contaminate the data. In SAS the format is called transport file. SAS has a specialized engine that handles the transport file. To convert to transport file, the PROC COPY procedure is used:

```
LIBNAME TRANS XPORT 'FILENAME';
 PROC COPY IN=WORK OUT=TRANS;
 SELECT COUNTRY;
 RUN;
```

This short program creates a transport file using the XPORT engine in the LIBNAME statement. Note that this LIBNAME statement is pointing to a file instead of a directory as we usually do.

The PROC COPY procedure copies the file WORK.COUNTRY to the output transport file. You can use any FTP program to transfer the file to other platforms including PC and UNIX. Note that binary mode is used.

The following is a complete MVS program that converts the SAS system file into transport file:

```
//IDXXORTZ JOB  (IDXX,2,20),'YOURNAME',CLASS=A,PASSWORD=XXXXXX,
// USER=IDXX
/*ROUTE  PUNCH UNTVM1.IDXX
/*ROUTE  PRINT UNTVM1.IDXX
// EXEC IEFBR14
//TRANS DD DSN=USER.IDXX.TEST.TRANS,UNIT=SYSDA,
//       VOL=SER=ACAD03,DISP=(OLD,DELETE)
// EXEC SAS
//TRANS DD DSN=USER.IDXX.TEST.TRANS,UNIT=SYSDA,
//     VOL=SER=ACAD03,DISP=(NEW,KEEP),
//     SPACE=(CYL,(50,20),RLSE),
//     DCB=(RECFM=FB,LRECL=80,BLKSIZE=8000)
//DATA1 DD DSN=USER.IDXX.TEST.SAS,UNIT=SYSDA,
//     VOL=SER=ACAD03,DISP=SHR
//SYSIN DD *
LIBNAME TRANS XPORT;
PROC COPY IN=DATA1 OUT=TRANS MEMTYPE=DATA;
run;
```

Your JCL

IEFBR14 procedure clears any file with same name to make room on MVS for data file.

Preparing room on MVS disk ACAD03 for writing data file.

SAS system file

SAS procedures

This program converts a SAS system file located on MVS system disk (ACAD03) into a transport file on the same disk. The IEFBR14 procedure cleans up any file previously created in the same name. Note that the boldfaced DCB (Data Control Block) specification, which fixes the LRECL (Logical Record Length) to 80, is the key for the file conversion. The new transport file MUST BE a sequential file in 80 columns.

2. Downloading file to CMS

Now, the transport file is on MVS disk. Before downloading to PC, we need to get the file to CMS. The IEBGENER program does this job:

```
//IDXXORTZ JOB (IDXX,12,999),'YOURNAME',
//         CLASS=N,PASSWORD=XXXXXX,USER=IDXX
```

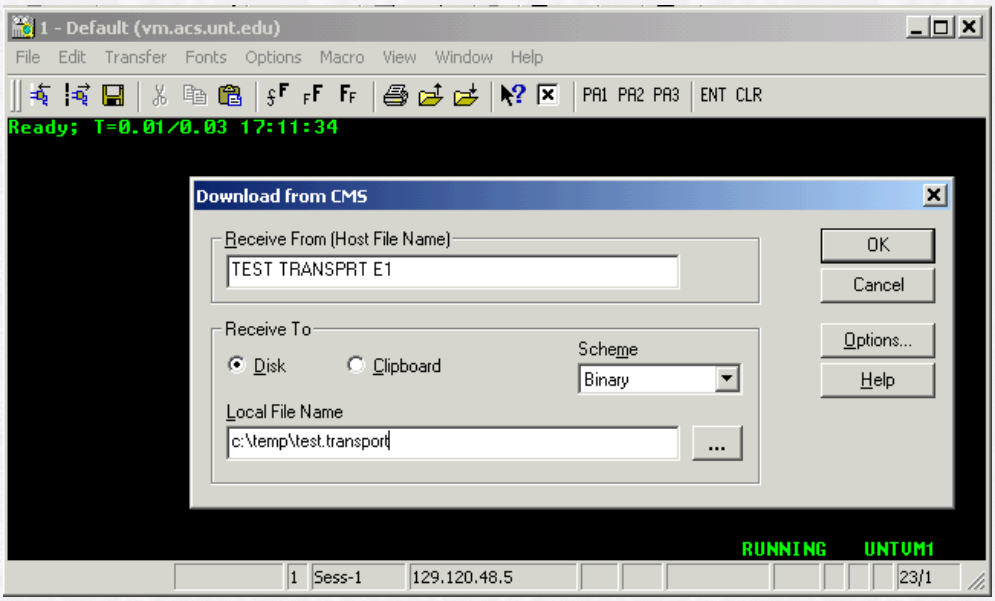
```
/*JOBPARM CARDS=9999999
/*ROUTE PRINT UNTVM1.IDXX
/*ROUTE PUNCH UNTVM1.IDXX
//STEP1  EXEC PGM=IEBGENER
//SYSPRINT DD SYSOUT=A
//SYSIN DD DUMMY
//SYSUT1 DD DSN=USER.IDXX.TEST.TRANS,DISP=(OLD,KEEP),UNIT=SYSDA,
//          VOL=SER=ACAD03
//SYSUT2 DD SYSOUT=B
```

This program downloads the transport file from the MVS system disk ACAD03 to the CMS environment and the PUNCH output (the SAS transport file) will be routed to the user's READER LIST. Preparing for a large data set, this program specifies CLASS=N and a large number for the JOBPARM parameter. It is recommended to create a temporary disk before receiving the file to CMS. To create a temporary disk in CMS, type: **tempdisk** at the Ready mode. Specify the number of cylinders for temporary storage. By default, the temporary 193 disk will be created as the E disk. Copy the file from the READERLIST to E disk by receiving the transport file into TEST TRANSPRT E1. At the command column, type:

RECEIVE / TEST TRANSPRT E1

3. Downloading the transport file to local host (UNIX or PC)

Most of the TN3270 programs (e.g. Hummingbird's HostExplorer) has built-in file transfer functions. Remember to transfer the file using binary format:



4. Reading the SAS transport file in SAS for Windows

Now the file is at bay and you can read it in SAS. The following program does the trick:

```
libname trans xport 'c:\temp\test.transport';  /* transport file */
libname new 'c:\temp'; /* output library */
proc copy in=trans out=new;
run;
```

This process may be a bit tedious but it works well with all my data. And you will find the time well spent. Regardless, my suggestion is what my mom always tells me: DO NOT WAIT TILL THE LAST MINUTE!

Have a merry Christmas and happy new year.

References

SAS Institute Inc., *SAS Companion for the MVS Environment, Version 6, Second Edition*, Cary, NC: SAS Institute Inc., 1996. 503 pp.

Research and Statistical Support (RSS). FAQ: How to convert a SAS data set from the mainframe to SAS for Windows? (http://www.unt.edu/rss/faq/sas0001.htm)

Network Connection

By Dr. Philip Baczewski, Associate Director of Academic Computing

This is the second in a series of "Back to Basics" articles. The first can be found here: "Back to Basics: DNS." - Ed.

Back to the Basics: Domain Names

Over the last five years, the term "dot-com" has entered into the common vocabulary. It is a designation for businesses which provide services or offer products for sale exclusively via the Internet. In the last year, the phrase "bankrupt dot-com startup" has become equally as common. The "dot-com" we say is derived from the ".com" suffix that is the end of most commercial sites' E-mail or Web addresses, that is, the very top level of their fully-qualified domain name (FQDN).

Domain names have been around since the early days of the Internet. The early standards were .mil -- military sites, .edu -- educational institutions, .gov -- government sites, .org -- non-profit organizations, and .com -- commercial entities. A bit later, .net came into being to indicate networks and network organizations and .int was created for organizations established by international treaties between governments. A domain name applies a human language appellation to an address on the Internet and also provides a way to categorize the function of that address.

Fully qualified domain names

To make a fully qualified domain name, you start with the top level suffix, such as .edu, and prepend qualifiers as you need them. For example, the primary domain for the University of North Texas is "unt.edu". Prepend an additional descriptor and you've got an FQDN -- www.unt.edu is UNT's Web server address, ftp.unt.edu is UNT's ftp server address, and eaglemail.unt.edu is the address of UNT's EagleMail student E-mail service. An FQDN can be a simple way to indicate the what, the where, and the why of a particular online service.

An FQDN always maps to a specific numeric address that represents a computer on the Internet (see last month's column for more on how this actually happens). In large networks like UNT's, FQDNs are usually hierarchically organized to match the organization of the institution: cc.unt.edu is used for network addresses in the Computing Center, while cas.unt.edu is used for addresses in the College of Arts and Sciences, etc.

FQDNs usually include no more than four parts. They can have more, but beyond four, it gets harder to keep track of all the pieces. Also, since numeric addresses have four parts, there is a tendency for FQDNs to match that level of hierarchy (this is a human rather than technical tendency). For example, sol.acs.unt.edu matches 129.120.220.42. Since the numeric address hierarchy is the reverse of the FQDN, we can see that unt.edu maps to 129.120, acs maps to 220, and 42 is specifically the computer named "sol."

Top-Level Domains

The number of domains has dramatically increased since the inception of the Internet. For

one thing, as the Internet became international, country Top-Level Domains (TLDs) were created. These correspond to the name of the country and are two-letter designations like .us for United States, .ca for Canada, .uk for the United Kingdom, .de for Germany, etc. The small south Pacific island country of Tuvalu has increased it's coffers by selling the rights to it's country TLD which happens to be ".tv".

In addition to the country TLDs, several new general domains have been either proposed or already implemented: .aero, .biz, .coop, .info, .museum, .name, and .pro are all in various stages in the approval process. The newest domains which are already available are .biz and .info. As you might expect, .biz is specifically for businesses, but .info can be for any site which is intended to provide information (which is a pretty wide category). The other TLDs are fairly obvious, except maybe for .aero, which is specifically for the air travel industry, and .pro, which is intended for professions such as doctors, lawyers, accountants, etc. The .name domain may become one of the most popular. It is specifically for individuals to obtain an address which is your name, such as john.smith.name.

ICANN?

With all of these new and old domains around the Internet, you know that somebody's got to be in charge of it all. Generally, that responsibility falls upon the Internet Corporation for Assigned Names and Numbers, otherwise known as ICANN. ICANN is a non-profit organization that took over the name and address management functions of the Internet from the U.S. Government and its contracted organizations. ICANN has oversight of the approval of new TLDs, provides an accreditation service for domain registrars, and runs the root-level servers which ultimately can point to where to find what FQDN goes with what numeric Internet address.

If you want an address to use exclusively for your business, organization, or even your personal use, you must register a name that is unique within your desired domain. There are a number of accredited registrars which can be found on the ICANN Web site (www.icann.org) and for a small fee (usually $25-$35 per year), they will verify that the name is available and register it on their site along with the appropriate contact information and Internet addresses. Usually, you will only register a domain name if you have an Internet address that it can point to. Some Internet Service providers will sell this to you as a package along with Web space and other services. On the other hand, if you have acquire a fixed Internet address from a service provider, you can register a name to point to that address, and assuming it's OK with your provider, you can set up your own network and servers using the domain name.

One of the more annoying "businesses" which have appeared in recent years has been "virtual real estate" or perhaps more accurately, "cybersquatting." Essentially, someone or some organization will register names which it thinks will be popular or which are similar to known products or organizations. They will then offer to register it or give up the name for a fee. Most accredited registrars have name dispute resolution procedures in place to ensure that someone else cannot use a name which is copyrighted by someone else, however, in some cases it may be necessary to go to court to prove such name ownership. ICANN has created a Domain-Name dispute resolution policy (http://www.icann.org/udrp/udrp.htm) which most accredited registrars must follow when a conflict exists.

The privatization of domain name management via the creation of ICANN and other organizations has resulted in the expansion of the domain name space with an increasing number of top-level domains from which to choose. It has also provided for competition in registration services which has resulted in a slight lowering of the average cost of a registered domain. You may never need a domain name for yourself, but the trend is now to

make domain names even more descriptive by adding TLDs like .biz or .info. I am guessing, however, that just like Herseys (tm) bar or Kleenex (tm), it will be a long while before "dot-com" disappears from our language.

Link of the Month

Each month we highlight an Internet, USENET Special Interest Group (SIG), or similar mailing list(s) or Website(s).

Got some time on your hands over the Winter Break? Why not do some online learning? Like it says on the UNT SmartForce Website, http://www.unt.edu/smartforce/ , "SmartForce Campus is an innovative computer-based training system offered free of charge to all UNT faculty, staff, and students. Smartforce offers courses on a large variety of information technology topics such as Java Programming, Novell NetWare, Microsoft Office, and UNIX on your own machine through the Internet."

We have published articles about SmartForce in *Benchmarks Online* frequently this past semester, most recently "Update on SmartForce CBT," which appeared in the October issue. If you are interested in using SmartForce, reading this short article is a good place to start.

WWW@UNT.EDU

How much do you care about security?

By Shane Jester, Campus Web Administrator

Are you concerned about your privacy? Most people when asked that question will say "of course!", but when it comes to doing something about it, most people take convenience over privacy. Take Portable phones for example. Any time you use a cordless phone in your home, you are increasing the odds that someone can eavesdrop on your conversation. Granted, most people are never exploited in this manner, but it is still a convenience that opens up the possibility.

Let me give you a more pertinent example of a convenience that is becoming more common place in our lives and will continue to grow in popularity in the near future. I recently installed some wireless networking devices in my home so that I can easily roam with my laptop and still have internet access without cables. With the right hardware, it is extremely easy to set up this system, however, the default settings of these devices implement no security measures. The big problem is that anything you do on that wireless network is very exploitable. Granted, traffic sent over the physical network is not too hard to intercept , but at least you have to have a physical presence on the network to intercept any data. If you choose not to implement wireless security, your next door neighbor can see everything that you are doing by simply installing a wireless card in their computer and scanning until they find what channel you are using. Possibly even more concerning, they can tap into your network and get free internet access using your connection. This could be a frustrating consumption of your bandwidth that may take you a while to even discover.

The sad thing is, it is not hard to lock down your network with a little time spent configuring your software. You don't even need an extensive background in data communications concepts to implement your security. Most wireless access points and cards come with software and documentation on encrypting your data and filtering your network from unwanted eavesdroppers or bandwidth poachers. It just takes a little extra time and patience, but believe me, the peace of mind is worth the extra time. Unfortunately, the attitude I see in most people is that convenience outweighs prudence. At least until they get "burned" once. My advice: Don't wait until you get burned. Do it right the first time.

Page One

Campus Computing News

Winter Break Hours

Forget about Santa Claus; 'GAUSS' is coming to town!

EagleMail News

Lab-of-the-Month: SLIS

IDE RAID Technology

Today's Cartoon

RSS Matters

SAS Corner

The Network Connection

Link of the Month

WWW@UNT.EDU

Short Courses

IRC News

Staff Activities

Subscribe to Benchmarks *Online*

Short Courses

By Claudia Lynch, *Benchmarks Online* Editor

ACS Short Courses are over for this semester. We will be offering similar courses in the spring.Please consult the Short Courses page to get an idea of the types of courses that will be offered.

Customized Short Courses

Faculty members can request customized short courses from ACS, geared to their class needs. Other groups can request special courses also. Contact ACS for more information (ISB 119, 565-4068, lynch@unt.edu).

Especially for Faculty and Staff Members

In addition to the ACS Short Courses, which are available to students, faculty and staff, staff and faculty members can take courses offered through the Human Resources Department, the Center for Distributed Learning, and the UNT Libraries' Multimedia Development Lab. Additionally, the Center for Continuing Education and Conference Management has a new program for interdepartmental training in business computer literacy. These classes are offered for a fee but discounts are given to those associated with UNT, and Inter-departmental Orders are accepted.

GroupWise Training

If would like to have a Basic GroupWise seminar for your area, please contact Jason Gutierrez, Campus Wide Networks, jasong@unt.edu . .Seminar Topics: Basic GroupWise, HTML Messages, FAQ. Tentative dates for next semester are:

1. February 14, 2002 10 a.m. - 11:50 a.m.
2. March 21, 2002 10 a.m. - 11:50 a.m.
3. April 25, 2002 10 a.m. - 11:50 a.m.

All seminars are in ESSC Room 152. For signup information, go to http://www.unt.edu/hr/training/treg.htm or E-mail Bhavna Vaswani at bvaswani@unt.edu

Center for Distributed Learning

The Center for Distributed Learning offers courses especially for Faculty Members. A list of topics and further information can be found at http://www.unt.edu/cdl/training_events/index.htm The center also offers a "Brown Bag" series which meets for lunch the first Thursday of each month at Noon in ISB 204. The purpose of this group is to bring faculty members together to share their experiences with distributed learning. One demonstration will be made at each meeting by a faculty member with experience in distributed learning. More information on these activities can be found at the Center for Distributed Learning Web site.

UNT Libraries'

The UNT Libraries' Multimedia Development Lab has also offered free training to all University of North Texas faculty and staff in the basics of FrontPage and information architecture in the past. For more information see http://www.library.unt.edu/media/services.htm#Distributed.

Technical Training

Technical Training for campus network managers is available, from time to time, through the Campus-Wide Networks division of the Computing Center. Check the CWN site to see if and when they are offering any training.

UNT Mini-Courses

These are a variety of courses offered, for a fee, to UNT faculty, staff and students as well as the general public. For additional information surf over to http://www.unt.edu/ccecm/cont_ed/Minicourse/Courses/UNT_Minicourse_Page.htm .

Alternate Forms of Training

Many of the General Access Labs around campus have tutorials installed on their computers. For example, the College of Education recently acquired some Macromedia Tutorials for Dreamweaver 4.0, Flash 5.0 and Fireworks 4.0.

The Training Web site has all sorts of information about alternate forms of training. Training tapes, Computer Based Training (CBT) and Web-based training are some of the alternatives offered. Of particular interest are courses available via SmartForce (formerly CBT Systems). See http://www.unt.edu/smartforce/ for more information.

There are also handouts for computer training on the following topics:

- GroupWise 5.2 Handout for Win95/NT
- FAQ for GroupWise 5.2
- Computers - Back to the Basics
- Introduction to Windows 95 /98/NT
- Introduction to Word 97
- Advanced Word 97 - MailMerge It Together
- Introduction to PowerPoint 97 (Creating a Slide Show)
- Introduction to Remedy (THE Call-Tracking Program)
- AND, the award winning Introduction to Excel 97

Adobe Acrobat Reader Format only for the following:

- Introduction to Microsoft Word 2000
- Introduction to Microsoft Excel 2000
- Creating a Slide Show with PowerPoint 2000
- Using Netscape Communicator & the UNT Home Page

Use the Internet to search for answers to Microsoft Office problems. See

http://www.zdnet.com/zdhelp/filters/office/ December 1999's "List of the Month" offers links to free Microsoft Word and Excel information also.

Page One

Campus Computing

Winter Break Hours

Forget about Santa Claus; 'GAUSS' is coming to town!

EagleMail News

Lab-of-the-Month: SLIS

IDE RAID Technology

Today's Cartoon

RSS Matters

SAS Corner

The Network Connection

Link of the Month

WWW@UNT.EDU

Short Courses

IRC News

Staff Activities

Subscribe to Benchmarks *Online*

IRC News

Minutes provided by Sue Ellen Richey, Recording Secretary

IRC Regular and Ex-officio Voting Members: *Judith Adkison, College of Education; Ginny Anderson, Fiscal Affairs; Donna Asher, Administrative Affairs; Craig Berry, School of Visual Arts; Cengiz Capan, College of Business, GALC; Bobby Carter, UNT Health Science Center; Christy Crutsinger, Faculty Senate; Jim Curry, Academic Administration; VACANT, Student Association; Duncan Engler, University Planning Council; Don Grose, Libraries; Jenny Jopling, Instruction Program Group; Joneel Harris, EIS Project Group; Elizabeth Hinkle-Turner, Standards and Cooperation Program Group; Abraham John, Student Affairs; Christine Mitchamore,Graduate Student Council; Ramu Muthiah, School of Community Services; Jon Nelson, College of Music; Robert Nimocks, Director, Information Technology,UNTHSC; Patrick Pluscht, Distributed Learning Team; Mark Rorvig, Research Program Group (Acting Chair); Paul Schlieve, Communications Program Group; Kathleen Swigger, College of Arts and Sciences; Philip Turner, School of Library and Information Science and University Planning Council (Chair, IRC); Virginia Wheeless, Chancellor for Planning.* ***IRC Ex-officio Nonvoting Members:*** *VACANT, Telecommunications; Charles Andrews, GALMAC; Bill Buntain, Computing Center Networking; Jim Curry, Microcomputer Maintenance Shop; Richard Harris, Computing Center and University Planning Council; Coy Hoggard, Computing Center/Administrative; VACANT, UNT Health Science Center; Maurice Leatherbury, Computing Center/Academic; Sue Ellen Richey, Computing Center (Recording Secretary). [As of 9/2001]*

November 20, 2001

VOTING MEMBERS PRESENT: PHILIP TURNER, Chair, RICHARD HARRIS, LOU ANN BRADLEY (for DON GROSE), ELIZABETH HINKLE-TURNER, JUDITH ADKISON, JONEEL HARRIS, MARK RORVIG, CENGIZ CAPAN, NANCY MCCRAY (for VIRGINIA WHEELESS), DUNCAN ENGLER, WIL CLARK (for JOHN PRICE), CRISTINE MITCHAMORE, CHRISTY CRUTSINGER, JIM CURRY, CRAIG BERRY, JENNY JOPLING **NON-VOTING MEMBERS PRESENT:** DOUG MAINS, RICHARD HARRIS, MAURICE LEATHERBURY, COY HOGGARD, CHARLES ANDREWS, SUE ELLEN RICHEY (Recording Secretary) **MEMBERS ABSENT:** PAUL SCHLIEVE, RAMU MUTHIAH, KATHLEEN SWIGGER, ROBERT NIMOCKS, JON NELSON, DONNA ASHER, ABRAHAM JOHN, BOBBY CARTER, GINNY ANDERSON, BILL BUNTAIN, BECKY MORGAN, PATRICK PLUSCHT **GUESTS:** JENNIFER LAFLEUR, RICH ANDERSON, CHARLOTTE RUSSELL, CATHERINE HARDY

The minutes of the October 16th meeting were approved as distributed.

DCSMT

Maurice Leatherbury reported for the DCSMT that they are still reviewing Office XP and will report to the IRC when that review is completed.

Instruction Program Group

Jenny Jopling reported for the Instruction Program Group that Counseling & Testing has decided to move its ETS to the Gateway Center, and originally only had room for 8 testing workstations for other than ETS usage but have agreed to increase that number to 12 workstations. It is not anticipated that their facility will be very useful for the University

campus as a whole because it wouldn't accommodate an entire class for testing. It can be used for CLEP testing as well as a place where distributed learning students could be tested, at least on a small scale.

EIS Planning Group

Joneel Harris reported for the EIS Planning Group that the three vendor overview presentations have been held, and were well attended, and said she was not sure of the general response to those presentations. The Planning Group is about to begin holding the detailed presentations which will involve the more functional aspects of the three systems. The schedule for all but one of the presentations is final and can be viewed at the web site. The only time conflict to be resolved is a presentation planned for the day UNT football team plays in the New Orleans Bowl game. An alternative date will be found and announced later. Joneel welcomed IRC members to attend at least the introductory presentations.

Cengiz Capan asked Joneel to explain the implementation approach, specifically did they plan to change UNT's business processes to match the vendors' software products, as opposed to changing the software to match UNT's existing business processes. Joneel replied that the Family Product Team Leaders are committed to looking at the product and best practices and adapting our way of doing business as much as possible to the software products; however, when it comes to compliance issues, that may not work. It is clear that some software modifications will have to be made. It was acknowledged that there will be problems across the campus when existing business processes are changed.

Research Program Group

Mark Rorvig announced that the Research Program Group will be meeting in the adjoining Conference Room immediately following the IRC meeting and invited IRC members to attend. They will discuss the increasing need for computing resources at UNT.

Standards & Cooperation Program Group

Elizabeth Hinkle-Turner reported for the Standards & Cooperation Program Group that as a result of comments at the October IRC meeting, there has been a revision of the UNT Information Resources Security Policy and she turned the discussion over to Maurice Leatherbury. Electronic copies of the revised policy as well as an accompanying draft of a UNT Information Security Handbook for Faculty, Staff, and Students were distributed prior to the meeting. Maurice explained that after reviewing state regulations and looking at what other universities have done, a revised draft of the Security Policy and a Handbook were prepared for consideration by the IRC. He asked that the handbook, when and if approved by the IRC be considered a "living document" so that IRC approval would not be required for changes to be made in it from time to time. This new Security Policy will replace the existing Microcomputer Data Integrity and the Information Resources Security policies. A vote on the acceptance of the new policy and handbook will be on the agenda at the December IRC meeting.

Distributed Learning Center

Jenny Jopling reported for Patrick Pluscht and the Distributed Learning Center that they are about to fill their Media Specialist position with someone who is a former producer for PBS in Wisconsin. The Center is working on completing the approximately 30 courses under production for Web-based delivery in the spring, and getting ready for on-line training for the Excellence in Teaching On-Line course which will be offered to faculty over WebCT in

the spring.

IRC Charge and Composition

The Chair asked for a motion to adopt the revised IRC Charge and Composition, which was presented at the October meeting. Cengiz Capan moved for approval of the revised document; Craig Berry seconded the motion. Richard Harris stated that the three changes that were suggested at the October IRC meeting were incorporated into the new document which is being voted on today. The motion passed and the new IRC Charge and Composition was approved.

IRC Meeting Schedule

The IRC generally meets on the third Tuesday of each month, from 2-4 p.m., in the Administration Building Board Room. From time to time there are planned exceptions to this schedule. This fiscal year, the December meeting has been changed to December 11th, and the May meeting to May 7th. All meetings of the IRC, its program groups, and other committees, are open to all faculty, staff, and students.

Page One

Campus Computing News

Winter Break Hours

Forget about Santa Claus; 'GAUSS' is coming to town!

EagleMail News

Lab-of-the-Month: SLIS

IDE RAID Technology

Today's Cartoon

RSS Matters

SAS Corner

The Network Connection

Link of the Month

WWW@UNT.EDU

Short Courses

IRC News

Staff Activities

Subscribe to Benchmarks *Online*

Staff Activities

Transitions

The following are new employees:

- **Mario Cauley**, Programmer III, Voice and Web Strategic Applications, MTS.
- **Amanda James**, I/O Operator, Print Services, Production Services, MTS (part-time).
- **Cristine McClure**, University Information Operator.

The following people no longer work in the Computing Center:

- **John Trautman**, I/O Consultant, Printing Services, Production Services, MTS (part-time).
- **Kristine Young**, I/O Consultant, Printing Services, Production Services, MTS (part-time).
- **Krysti Ripaldi**, Microcomputer Consultant, Helpdesk, ACS (part-time).
- **Rebecca Saunders**, Microcomputer Consultant, Helpdesk, ACS (part-time).
- **J. Chris Bates**, I/O Consultant, Printing Services, Production Services, MTS (part-time).

Awards, Recognition

- **Maurice Leatherbury**, Senior Director of Academic Computing, discussed UNT's policy regarding Internet piracy in the November 12-18 *Dallas-Fort Worth Tech Biz*.
- **Scott Blackwell** and **Eanen Cohn**, Helpdesk Microcomputer Consultants, both graduated with their Bachelor's degrees December 15. The good news for the Helpdesk is they are continuing their education at UNT, so they will continue working for the Helpdesk!

According to the December 2001/January 2002 *Human Resources Newsletter*, the following people have been nominated as Soaring Eagles and will receive their award at the President's Staff Sack Lunch on February 26:

- **Philip Baczewski**, Associate Director, Academic Computing Services, "recently helped a co-worker set up a new computer and transfer important data from the old one to the new one. Way to Soar!"
- **Robert Jones**, Programmer/Analyst on the Student Records Data Systems Team, and **Linda Wallace**, Programmer/Analyst on that same team, soar for their judging of the pumpkin carving contest for the ESSC.
- **Maurice Leatherbury**, Senior Director of Academic Computing, was

recognized for his extraordinary efforts in employee morale building including this: he has brought doughnuts virtually every Friday since 1995!

Several people are celebrating anniversaries this month here at UNT. **Sandy Burke**, Manager of Computing Center Help Desk Support Services, and **Linda Terry**, Campus Information Operator Supervisor, are both celebrating 30 years of service! **Nevin Ellis**, Systems Programmer/Analyst, recently marked 10 years here at the University.

Holiday Greetings

From the Campus Campus Wide Networks Computing Team (Keep viewing for a holiday bonus!): *tm*

http://cwn.unt.edu/cwn/ccard

(Please wait: it may take several minutes to view using a modem)

Campus Computing News

By Dr. Maurice Leatherbury, Senior Director of Academic Computing

2002-2003 Teaching with Technology Grants Announced

On December 13th, Dr. Kesterson sent out announcements of the latest round of Teaching with Technology Grants to all tenure-track faculty members. This is the seventh consecutive year that the Vice President for Academic Affairs has funded this important grant program, and like those in the past this year's program is directed toward the use of technology to improve instruction at UNT. Also like last year's program, the TwT Grants are limited to *distributed learning courses*, that is, those that are taught more than 50% of the time in a mode other than face-to-face in the classroom.

The full grant request for proposal can be found at the Center for Distributed Learning's Web site: both the call for proposals as well as a downloadable Word application form can be accessed there. If you are submitting a proposal, be sure to use the form because that's the only format in which the evaluation committee will consider your submission.

Dr. Kesterson is providing up to $150,000 in funding for the TwT Grants this year. Something new in this year's program is funding for summer faculty release time in order to allow faculty to spend the summer developing distributed learning courses. $50,000 of the $150,000 has been set aside for the release time funding, which is limited to a maximum of $5,000 per faculty member. The remainder of the $150,000 will be awarded in grants of up to $8,000 to support distributed learning course development, primarily in the form of student assistants to convert existing course material to electronic form. As in the past, Teaching with Technology Grant funds can't be used to pay for traditional teaching assistants to conduct course research in the library, to develop course content, or to tape a course lecture. Specifically, this portion of the grants is restricted to "programming" because the funds come from HEAF sources and their expenditures must result in machine-readable files.

Jenny Jopling, Associate Director of the Center for Distributed Learning, (extension 4462) and I (x 3854) can provide additional information or assistance in the grant proposal process.

Winter Break Hours

By Claudia Lynch, Benchmarks Online Editor

Following are the hours for Computing Center-managed facilities for the Winter Break. All staff offices will be closed Monday, December 24 through Tuesday, January 1, 2002. The Helpdesk, ACS General Access Lab and Mainframe Print Services will maintain some services during that period, however.

- **Print Services** will be open from 6 a.m. - 2 a.m. M-F, 10 a.m. - 12 midnight Saturday and Sunday, throughout winter break ***except*** for the week of Christmas/New Years. They will be **closed** December 24 & 25. On December 26-30 they will be open 8 a.m. - 8 p.m. each day. Monday, December 31 they will be open from 10 a.m.- 4 p.m., and they will open at 10 a.m. on New Year's Day. The winter break schedule (above) resumes at that time.

- The **Helpdesk** will be open Monday, December 24 - 8 a.m.- 8 p.m. then:

  Tuesday, December 25 - **Closed**
  December 26-28, 2001 - 8 a.m. - 8 p.m.
  Saturday, December 29 - 9 a.m.- 6 p.m.
  Sunday, December 30 - 1 p.m. - 10 p.m.
  Monday, December 31 - 8 a.m. - 8 p.m.
  Tuesday, January 1, 2002 - **Closed**
  January 2, 2002 - resume normal hours (M-TH 8am-midnight; F 8am-8pm, SA 9am-5pm, SU 1pm-midnight)

- The **ACS General Access Lab** (ISB 110) will be open Monday Dec. 17 - Friday Dec. 21 from 9:00 a.m. - 5:00 p.m. then:

  Saturday, December 22 - Tuesday January 1, 2002 - **Closed**
  Wednesday, January 2 - Friday, January 4: Lab is open 9:00 a.m. - 5:00 p.m.
  Saturday, January 5 and Sunday, January, 6 - **Closed**
  Monday, January 7 - Saturday January, 12: Lab is open 9:00 a.m. - 5:00 p.m.
  On Sunday January 13, resume normal semester hours (Sundays -- 1:00 p.m. - 11:45 p.m.; Mondays -- Thursdays -- 8:00 a.m. - 11:45 p.m.; Fridays -- 8:00 a.m. - 8:45 p.m.; Saturdays -- 9:00 a.m. - 8:45 p.m.)

Hours for Other Campus Facilities

The University is officially closed for Winter Break Monday, December 24 through Tuesday, January 1, 2002. Classes for the spring semester begin January 14.

General Access Labs

- WILLIS:

  December 21 – January 1 **Closed**
  January 2 – 4 Open 8 a.m. – 6 p.m
  January 5 – 6 **Closed**
  January 7 – 10 Open 8 a.m. – 6 p.m
  January 11 Open 8 a.m. – 3:30 p.m
  January 12 Open 9 a.m. – 6 p.m
  January 13 Starting at 1 p.m normal daily hours resume

- SLIS:

  December 15 – January 13 **Closed**
  January 14 Normal daily hours resume

- MUSIC:

  December 15 – January 13 **Closed**
  January 14 Normal daily hours resume

- SCS:

  Normal hours through December 21
  December 22 – January 13 **Closed**
  January 14 Normal Daily hours resume

- SOVA:

  December 16 – January 13 **Closed**
  January 14 Normal daily hours resume

- COE:

  December 15 – January 11 **Closed**
  January 12 – 13 Maintenance 8 a.m. – 6 p.m
  January 14 December 15 – January 13 **Closed**
  January 14 Normal daily hours resume

- COBA:

  December 15 – January 13 **Closed**
  January 14 Normal daily hours resume

- CAS:

  December 15 – January 13 **Closed**
  January 14 Normal daily hours resume

Forget about Santa Claus; 'GAUSS' is coming to town!

By Dr. Elizabeth Hinkle-Turner, Student Computing Services Manager

During the winter break, a new server will be brought online which promises to significantly contribute to statistical research here at UNT. Housing of statistics applications and data has typically been on the server ACSLAB, but the increasing size of current stat packages coupled with the heavy use of CPU services by the ACS General Access Lab, has prompted the purchase of a machine specifically devoted to needs of statistics scholars.

Ever wondered what a server looks like? Well, now you know!

Named GAUSS after one of the early statistics and number theorists (To read more about Johann Carl Friedrich Gauss click here), this Compaq Proliant ML530 server sports a 1 Ghz Xeon processor and 512MB of RAM. Most importantly, with 72 GB of drive space available, there is plenty of room for all the statistical research applications utilized on campus, all the data collected and needed by UNT scholars, and the inclusion of a "hot spare" drive for greater hardware integrity. Additional drive slots are available should even more disk space be needed.

The migration from ACSLAB to the GAUSS server will be gradual. All ACSLAB statistics services will remain in place until Fall 2002 (and beyond if necessary) in order to give network personnel plenty of time to reconfigure shortcuts and permissions. Researchers should not be affected at all as their rights to files and folders will be transferred to the new machine. Most importantly, software and hardware use will be carefully tracked from the new server so that licensing and disk needs can be more easily assessed for expansion purposes. GAUSS will be a Novell NetWare 5.x server with Zenworks 3 for maximum machine administration capabilities.

With more drive space will come more room to offer new statistics software and services. Dr. Karl Ho has already made several purchasing requests in anticipation of this and testing will begin on new products during the winter break. As the server administrator I am excited about bringing a new research server online for the university community and look forward to continuing to provide updates on its installation and services. For more information about GAUSS itself and the installation and migration process, you may contact me at ehinkle@unt.edu. For more information about software licenses and new statistics plans and products, contact Dr. Karl Ho at kho@cc.admin.unt.edu.

EagleMail News

By Dr. Philip Baczewski, Associate Director of Academic Computing

One E-mail account coming up - no waiting!

The student EagleMail account application found at http://getlogin.unt.edu/ now provides almost immediate access to E-mail and other services. Previously, accounts were not activated until 8:30 p.m. after the online application was completed. Academic Computing Staff have upgraded the EagleMail application page so that it creates a password entry and EagleMail mailbox as soon as the the application is successfully completed. The EagleMail account will be fully functional an hour after the application is completed.

EagleMail in an Instant

One side benefit of this instant access is that other services which employ the EUID and EagleMail password, such as WebBills or personal Web page publishing, are now immediately accessible upon completion of the EagleMail account application. To eliminate the need to apply for services twice, the process which allows students to set a password for WebBills now also activates EagleMail at the same time - one stop shopping all around. Some students may need to set a WebBills password again so that their EagleMail account will be activated.

EagleMail - it's not just a good idea...

UNT policy number 18.5.7 (http://www.unt.edu/planning/UNT_Policy/volume3/18_5_7.html) requires that students activate their UNT E-mail accounts and that they read "their E-mail frequently enough to receive important communications from the University." It also allows for use of external E-mail systems: "Students can forward their mail from the UNT-provided mail service to a preferred account on some other E-mail service. Students are responsible for ensuring that any forwarding address they set is accurate and operational." Students can set a forwarding address as well as control or monitor other features of their EagleMail service by visiting the UNT Internet Services page (http://people.unt.edu/manage/).

Looking forward

While students can use Hotmail or Yahoo to read their messages, we still think that EagleMail has a lot to offer (accessibility and reliability being a couple of those things). In the next couple of months, we will be testing and implementing a **new version** of the software that runs the EagleMail Web client. The new version includes a number of new features that allow you to customize various aspects of the E-mail interface and operation.

You can help us evaluate the new interface. To do so, go to the EagleMail login page (http://eaglemail.unt.edu) and click on the "New EagleMail Interface" link.

You will then login with your EUID and password. Please note that the new interface no longer uses frames in order to comply with ADA standards. The new interface does, however, provide access to all of the same features as the old interface with many new additions. Please try out all of the new features (including the newly expanded "options" area) and report any problems to the Helpdesk.

Page One

Campus Computing News

Winter Break Hours

Forget about Santa Claus; 'GAUSS' is coming to town!

EagleMail News

Lab-of-the-Month: SLIS

IDE RAID Technology

Today's Cartoon

RSS Matters

SAS Corner

The Network Connection

Link of the Month

WWW@UNT.EDU

Short Courses

IRC News

Staff Activities

Subscribe to Benchmarks *Online*

Lab-of-the-Month: The Graduate Studies General Access Lab

By Dr. Elizabeth Hinkle-Turner, Student Computing Services Manager

Grad school getting you down? Has your advisor rejected Chapter Three of your fascinating and insightful thesis once again? Failed that Music History portion of quals for the fifth time? Believe it or not, there are some great perks to being a graduate student and one of them is that you have a general access lab of your very own! Located in ISB205c, the Graduate Studies General Access Lab was established specifically to cater to the needs of UNT's upper level student population. Alan Livingston, manager of the lab, notes "Most graduate students that come to our lab find it seems more comfortable than other labs they have used on campus. I think it's mostly because it is smaller than other labs, which cuts down on the noise, crowded work areas, and more personal service from the lab's staff. I have tried to put together a staff that not only can help the average lab user, but can make a stressed out graduate student feel like things might work out."

The Graduate Studies Lab offers a quiet and productive work setting

Patrons of the lab will find the latest hardware and software available for their needs. Boasting 16 Pentium III 1 Ghz stations, a Macintosh G3 and a Macintosh G4 station, and laserprinters and scanners, this lab has some of the best technology on campus. Software available includes Office 2000 and Microsoft Project; all of the major statistics applications; Microsoft Visual Studio; Adobe Acrobat, PageMaker, Illustrator and PhotoShop; and Java 2.0 SDK. Additionally, the lab management has begun working with the graduate reader, Julian Long, to provide any student that is preparing their thesis or dissertation, a place to come and create the final PDF product, work on statistics with SAS or SPSS, and other tools that not all of the labs provide.

The Graduate Studies General Access Lab is located in the School of Library of Information Science (SLIS) because this school only offers graduate degrees. However, the current usage of the lab is equally divided between SLIS majors and other graduate fields and Livingston especially wants to emphasize that this lab is for *all* graduate students at UNT. Word is getting out quickly; usage of the

lab has literally doubled in the past two years.

The Graduate Studies General Access Lab is open Mondays - Thursdays 8:00a.m. to 10:00p.m.; Fridays and Saturdays 8:00a.m. to 6:00p.m.; and Sundays from noon to 8:00p.m. A very thorough Website detailing all of the lab's services is accessible at http://www.unt.edu/slis/gradlab/

IDE RAID Technology

By Duane Gustavus, UNIX Research Analyst

Abstract

This paper documents an investigation of a new disk storage array strategy that offers an attractive price/storage point as compared to current commercial SAN or NAS products. The discussion includes a general description of the technology, practical details of implementation and an itemized price list.

Introduction

One of the salient features of a great deal of contemporary scientific research is the reliance on significant amounts of computing power. The data that results from these computations must be stored for analysis and reference in later research. It is not surprising, then, that new terms like "disk farm" and "terabyte" are being added to the technical jargon of the day.

Current commercial answers to this vexing problem come in a variety of flavors, but share the general feature that significant amounts of storage come at significant prices. Because the definition of "significant" is at the fulcrum of the topic under consideration, I propose that one terabyte of disk space (1,000 gigabytes or one million megabytes) and $10,000 are both significant values for individual research projects. When changing the definition of significant (say one petabyte of storage or $1,000 investment), the options available for investigation might certainly be expected to change as well.

Unfortunately the terabyte/10k price point is not a hotly contested slot in the current commercial market. Some of the national supercomputer centers[1] and other research groups[2] have published results of their attempts to provide some solution near that price point, and those investigations lent impetus to a similar exercise here at UNT. The results presented here indicate the terabyte/10k price point is not wishful thinking, but there are also compromises that should be understood before building your own.

RAID Arrays

Many computer users already have experience using disk arrays, often referred to as RAIDs (an acronym for Redundant Arrays of Inexpensive Disks). The RAID concept was a reaction to the problems and expense involved with building continually larger (in terms of storage capacity) disk devices. If one could make several smaller devices appear functionally as a single disk drive, the problems of bit density and physical size could be neatly side stepped. This nut was cracked, and there are many varieties of RAID controllers on the market now, providing both increased reliability and performance.

RAIDs can be configured in a variety of ways which highlight the required compromise between data integrity and storage size. The trade-off is generally

to increase data integrity by writing the data in more than one place (data redundancy) which in turn reduces the amount of storage available by a factor of the redundancy. The most obvious version of this trade-off is between RAID0 and RAID1. In RAID0, multiple disks are interleaved to appear as one disk which is the size of all disks combined (called striping); in RAID1, half the disks are used for data redundancy and are simply copies of the other half (called disk mirroring). RAID0 maximizes the amount of space available, but at the cost of reliability (the probability of disk failure increases with the number of disk drives, and any single drive failing breaks the entire array); RAID1 increases reliability (two disks would have to fail to loose data), but at the cost of total available storage which is cut in half. Some attempts have been made to provide a compromise between the two extremes with features of both. RAID10 is the combination of RAID0 and RAID1 where a set of striped disks are mirrored. This provides the redundancy of mirroring with the large file system sizes available to RAID0 configurations. Another approach to a compromise between size and redundancy is RAID5. In a RAID5 disk array, one drive can be considered logically as a parity drive (parity is a methodology for reconstructing data). When a single disk fails, the data can still be reconstructed from the parity information. This provides redundancy at a lower storage cost than mirroring, but has slower performance due to the nature of constructing and reconstructing the parity information.

Up to this point, RAID drive arrays have primarily been specialty products aimed at data centers where big systems (and big budgets) are the rule. One cost factor involved with these RAID products has been the type of disk drive supported. Due to design limitations, the IDE disk drives commonly used in desktop computers were not useful for RAID configurations, so more expensive SCSI-interfaced hard disks were and are the norm. This is unfortunate from a cost perspective because the economy of scale involved in the huge desktop computer market constantly drives down the cost of IDE drives while the competition for a share in that market has continually increased the performance.

Several companies have recently introduced IDE RAID controllers (sometimes called storage switches) which are designed to use the newer ATA100 and ATA133 specification IDE drives. This approach offers considerable savings due to the lower cost of the hardware involved and the necessity of establishing a new price point in order to compete with established vendors already controlling NAS/SAN market share. Nonetheless, entire IDE-based RAID disk systems are already appearing in commercial form, so you may be able to buy something similar to the system described in this paper by the time you read this, thus avoiding the hassles (and savings) of building your own. Even if you decide to buy a complete system from a commercial integrator, these experiences might be useful in helping you understand the compromises you must make.

Components

In order to employ one of the new IDE RAID controllers, a server class computer system will be needed. The controller selected for this project is a 64-bit PCI card, so nothing exotic is required; a system was specified which is generally representative of current[3] high-end desktop technology. The exceptional part of the specification is that a case/power supply combination

must be selected which can meet the requirement of running at least twelve disk drives. These drives generate heat and can be expected to generate a substantial power surge when the system is first turned on. To reduce costs further, existing hardware could be used or other components selected, but the ability to power and cool a dozen disk drives is not a minor issue and should not be discounted.

Items on the original system purchase order are shown in the following table. These were reasonable prices at the time of order, but will almost assuredly have changed in price since this investigation:

Description	Qty	Unit price	Subtotal
ASUS A7M266 mbd	1	169.00	169.00
AMD Athlon 1.4g cpu	1	189.99	189.00
Thermaltake cpu cooler	1	22.99	22.99
256 mbyte DDR RAM	2	99.99	199.98
Maxtor 100g ATA-100 disk drives	12	261.00	3,132.00
ATI Radeon DDR 32meg video card	1	159.99	159.99
Antec SX1480B case w/ dual p/s	1	667.00	667.00

Some required items were not on this PO because they were available locally (CD-ROM drive, network card, video monitor, mouse and keyboard). The video subsystem is not particularly important on a file server, so this is a good area to look for savings. The components actually delivered substituted an 80g Maxtor drive due to delivery problems with the higher density drive and the desire not to pend the project on delivery of that product (and at a price reduction to $229.99 each). With the 20% decrease in drive storage, the terabyte target slips to 800 gigabytes for this implementation, but the 100g drives seem to be available now (and possibly larger ATA-133 drives). The final component necessary for this investigation is the IDE RAID controller. The model selected was a product from the 3ware corporation named the Escalade Storage Switch model 7810. There are several others, and this study is not an endorsement of the 3ware product over any other.

This particular controller supports up to eight drives (which is why two were ordered) in various RAID configurations. Support is provided for IDE drives that meet the ATA/100 specification (or lower). The company also claims "hot swap" and "hot spare" capabilities, but these features have not been tested in this implementation. The controllers were purchased for a unit price of $385.00, and were the component (there's always one) which was the slowest piece to arrive, delaying the project by several weeks.

The system was constructed while awaiting delivery of the RAID controllers using one of the drives for system software with the motherboard's IDE controller. The software installed was the RedHat Linux 7.2 distribution which comes standard with 3ware drivers. The primary installation issue involved the physical layout of the case selected. In order to make the IDE disk cables reach from the controller to the disk drive bay, it was necessary to machine two "windows" in the slide-out tray that holds the motherboard for cable routing. This modification was a fairly trivial machining job if you have a machine shop, but cable routing should be carefully planned in advance if you don't. Most of

these cases will be designed with SCSI cabling in mind, which means one or two fairly long cables with multiple connectors; for this controller, each disk drive as a separate 80-pin ribbon cable (supplied with the controller) which requires a little more forethought. In addition, the controller comes with a set of cables which "Y" the power connectors, but your case selection should provide enough power leads as an indication the manufacturer designed the system to support several drives. When the controllers arrived, one was placed in the system and connected to eight drives. The instructions provided with the controller are minimal, but there was no difficulty in putting the controller and drives into service.

The total cost of the final configuration as implemented, including shipping and handling charges, was $4998.83. This does not include keyboard,mouse, NIC, video monitor or miscellaneous hardware (disk drives do not come with mounting screws etc). The potential data storage capacity is 880g (one of the twelve drives is used for system software and therefore not placed on the RAID controller) for a price very close to $5,000.

Implementation

The primary thrust of this implementation was low-cost, network accessible disk storage. To meet that system criterion, it was only necessary to see if file systems could be built on the array and then exported to the network. The initial tests were done with a configuration using one controller and eight disk drives. The drives are represented to the user as a single SCSI disk (ie the 3ware controller looks like a SCSI controller to Linux). In the first case, the entire capacity of all eight drives was used to build a single file system. From the user perspective, it looks like:

```
Filesystem           1k-blocks      Used Available Use% Mounted on
/dev/hda1               132207     86550     38831  70% /
/dev/hda10            71837756   1379752  66808844   3% /export
/dev/hda6              2016016     36268   1877336   2% /home
none                    256408         0    256408   0% /dev/shm
/dev/hda9               194443        24    184380   1% /tmp
/dev/hda5              2016016    890712   1022892  47% /usr
/dev/hda8               194443     17522    166882  10% /var
/dev/sda1            615383612        24 584123912   1% /3W
/dev/cdrom                9158      9158         0 100% /mnt/cdrom
```

Notice the file system named 3W which shows up as device /dev/sda1 contains 615 gigabytes (the reported sizes are in one kilobyte blocks). The eight 80g drives have a maximum advertised storage of 640g, but the file system requires some of that storage (about four percent in this case) for metadata. Under the Available column, this file system has only 584g showing. That is the result of saving back five percent of the space in the file system which only root can utilize (this is a standard practice which can be modified with special directives). From the user perspective then, we have a file system with just over half a terabyte, however, notice that this is about 91% of the theoretical maximum storage available for this configuration (one controller, eight drives,

640g). In tests with different types of RAID, redundant storage will make 90% utilization of advertised space look very attractive indeed.

The file system is now operational, but not very useful because there is only local access. This file system can be exported for network access using NFS (Network File System). The following screen capture shows the mount command used to make the new file system available from a different machine by using the path /3W (in other words, the same path but on a different computer):

```
[root@jonvon /]# mount -t nfs tbyte:/3W /3W
[root@jonvon /]# df -k
Filesystem           1k-blocks      Used Available Use% Mounted on
/dev/sda1               132207     57031     68350  46% /
/dev/sdb3              6467288   3736116   2402644  61% /bak
/dev/sda10             3921404   1346944   2375256  37% /export
/dev/sda6              2016016    392140   1521464  21% /home
/dev/sdb2              2016044    392048   1521584  21% /home_bak
/dev/sda8               194443       134    184270   1% /tmp
/dev/sda5              2016016   1424616    488988  75% /usr
/dev/sda9               194443     23617    160787  13% /var
tbyte:/3W            615383616        24 584123912   1% /3W
```

Comparing the two views of the file system will show that only the device name has changed with the NFS mount showing the hostname from which the file system was shared. The configuration which defines which file systems are exported (that is capable of being shared) also controls where they will be available (allows access control by host). In other words, the file system may be shared with only specific hosts, or entire subnets, or any system on the network capable of performing an NFS mount. It is important to remember when using NFS that file sharing in this fashion presumes a common user base (ie the user named linus on the file server will have the same user ID on all hosts that share the server's exported file systems).

The exercise was to provide large amounts of inexpensive disk space which was accessible over the network. To this point a file system of 584 usable gigabytes has been provided for a cost of $5,000, or about 116 megabytes per dollar (remember this cost includes the computer and all hardware necessary to use the disks). Having established a "beachhead" on the storage/price criterion, some benchmarks are in order to characterize the performance of the result.

Benchmarks

The benchmark numbers provided in this section are intended to be neither authoritative nor exhaustive. In order to provide more reproducible results, the tests would need to be run over a private network in a much more controlled environment. On the other hand, it is often quite difficult to reproduce the performance documented in this type of benchmark testing outside a controlled environment. These tests were run over the UNT Academic Computing Services subnet during "off" hours (which means early morning here). In addition to the

described system, some commercial alternatives were tested and the numbers provided for perspective. The environment was *not* equal for all tests simply because some of the systems are not available on the public network. This additional information is still useful to get the "flavor" of the compromises required.

To test disk performance, the publicly available bonnie benchmark was used. This program tests sequential read/write performance in both character and block mode, and also provides a random seek test. The file size used for these tests was 100 megabytes with results averaged over five runs as kilobytes/second. These test were run on the standard Linux file system type (ext2)[4]. The following table summarizes the test runs on tbyte[5]:

Type	char in	blk in	char out	blk out
IDE	14,465	325,499	15,236	285,120
RAID0	9,540	330,983	9,442	250,264
RAID5	9,182	331,235	9,428	260,339
RAID5-NFS	6,050	10,710	5,384	229,958
RAID10	9,555	336,194	9,475	276,232
RAID10-NFS	6,001	224,985	5,309	10,677

The test labeled IDE was run to the IDE disk attached to the standard IDE controller providing information about the performance capabilities of a single drive; the remainder were run on groups of disks attached to the RAID controller. The RAID0 configuration was a single stripe over eight disks with RAID5 also using a full eight disk array. The RAID10 and RAID10-NFS tests were both run with an array configuration using four drives (a two drive stripe which is mirrored). The RAID10-NFS test was run on a remote host using the RAID10 array over an NFS mount point. The two NFS tests are probably the most useful in terms of predicting performance because they were run over the UNT network. Both of the NFS tests exhibit the performance decrease characteristic of network file accesses. In addition, the RAID5-NFS figures show the combined effects of network access and the read-before-write design (two network accesses) typical of RAID5 configurations. The trade-off between RAID1 (or RAID10) and RAID5 is more obvious when the data size available to users is considered for eight disk arrays: 511g for RAID5 and 292g for RAID10. In other words, for redundant array types, the RAID5 will provide more storage[6] at lower performance.

In order to get the flavor of the performance compromise between this approach and some commercial products, similar tests were run on two different "disk farms" employed by ACS in production mode. This is not an apples-apples comparison for several reasons. First, the network over which these tests were run is not the campus network (this should affect only the NFS test). In the case of the MetaStor NAS, the file system involved was a proprietary Veritas file system using RAID5. The Nstore is a SAN, so the "network" attachment is a fiber technology SCSI network and not the common campus network. Finally, the cost of the Metastor with 375g was about $130,000 while the Nstore with 500g was about $50,000 with an additional $2k/node charge for the fiber interface card..

Type	char in	blk in	char out	blk out
MetaStor	11,502	116,503	13,146	48,269
MetaStor-NFS	10,898	465,338	6,999	6,649
Nstore	13,360	509,140	13,262	161,215

Both of these systems provided generally improved performance on these tests, but at significantly increased price. Part of the compromise here is due to the projected use of the storage. With the MetaStor, the physical configuration will support many more disk drives than are currently employed (growth capacity). While the cost of these drives will be high, the hardware described in this project is simply not capable of the storage capacities which are supported by the MetaStor. In addition, some of the costs involved in the MetaStor are for large buffers which will sustain these performance numbers under much greater load. The Nstore SAN is is high-performance technology most comparable to a local disk drive because the data transport is not over a general purpose network. The performance comes at the cost of specialized interface cards and fiber cable connects which are an additional expense to the normal NIC already part of the system.

Summary

The search for storage space will become an increasingly common venture as more data is churned out by computational research in every field. In some cases where the amount of space is prioritized over performance concerns, the IDE RAID technology can provide impressive amounts of storage per dollar and, perhaps more importantly, is within reach of modest departmental or even project level budgets. The implementation is not particularly difficult if local UNIX competence is available, but could be an issue if that is not the case. This technology is new enough that long-term stability should be a concern, though there was no sign of any instability in the process of running these tests. In addition, NFS is not an appropriate solution in all contexts. If, however, the requirements can be met by this approach, a very impressive price/storage point is possible with an IDE RAID system.

The decision about which type of technology to employ should not be generalized into a "one size fits all" methodology. The data collected in this project argues that at the current time there are several viable approaches, each with its own set of compromises. The premise of deploying a system with "room to grow" should be greeted with special skepticism because the "upfront" costs of this growth potential are high, while the rate of technological change can easily devalue the usefulness of growing a two-year old technology.

About this document ...

IDE RAID Technology

This document was generated using the **LaTeX**2HTML translator Version 2K.1beta (1.47)

The command line arguments were:
latex2html `-no_subdir -split 0 -show_section_numbers /tmp/lyx_tmpdir28395Hp3nNC/lyx_tmpbuf28395Uogozw/tbyte.tex`

The translation was initiated by Duane Gustavus on 2001-12-19

Footnotes

... centers[1]
http://staff.sdsc.edu/its/terafile/

... groups[2]
http://www.research.att.com/~gjm/linux/ide-raid.html

... current[3]
The project PO was dated 8/14/01.

... (ext2)[4]
Preliminary tests over the newer ext3 journaling file system seem to indicate a 3% degradation in performance on all tests but block write where the degradation was 40%.

... tbyte[5]
The raw data is available on request.

... storage[6]
How well RAID5 utilizes storage is a function of the number of spindles.

Page One

Campus Co

Winter Break Hours

Forget about Santa Claus; 'GAUSS' is coming to town!

EagleMail News

Lab-of-the-Month: SLIS

IDE RAID Technology

Today's Cartoon

RSS Matters

SAS Corner

The Network Connection

Link of the Month

WWW@UNT.EDU

Short Courses

IRC News

Staff Activities

Subscribe to Benchmarks *Online*

Today's Cartoon

From "Today's Cartoon by Randy Glasbergen", posted with special permission. For many more cartoons, please visit www.glasbergen.com.