**Design of Experiments**

**Running Time Estimates**

Using the historical data, a simple linear regression model can be used to estimate the time to finish for 150 runs. The R-square 0.99, indicating a nearly perfect fit. The regression line is as follows.

MinutesPerRun=0.1016+0.1116\*TotalGB

The plot of the data fit is shown as follows.



**Suggested Configurations for a Full Coverage of Design Space**

Based on the information from Jon and the use of equal allocation to provide a maximum coverage of the design space, the following design configurations are suggested.

*CPU Frequency:*

The CPU Frequency in GHz are: 1.2, 1.6, 2.0, 2.4, 2.8, 3.2, 3.5. Seven points in total.

*Number of Threads:*

The number of threads are: 1, 8, 16, 24, 32, 40, 48, 56, 64. Nine points in total.

*Per Thread File Size:*

The per thread file sizes are: 4K, 16K, 64K, 256K, 1M, 4M, 8M, 12M, 16M. Nine points in total. I am not sure the small file size points are necessary but they are not time consuming and provide a good coverage on the log scale of the file size. With 16M per thread and 64 threads, the largest total file size is 16M\*64=1G.

*Per Thread Record Size:*

The per thread file sizes are: 4K, 16K, 64K, 256K, 1M, 4M, 8M, 12M, 16M. Nine points in total, with the constraints that record size can not exceed the file size and it must be a factor of the file size.

*Overall Configurations:*

The following two plots illustrate the scatter of the design points.





*Running Time Estimation:*

There are 2709 total configurations (see attached .csv file). For 150 runs, the total estimated time is 5.1 weeks. With three nodes (each run 50 runs), it can be finished within 1.7 weeks.