Journal of Information Technology Impact
Vol. 10, No. 2, pp. 87-98, 2010

High Performance Computing Instrumentation and Research Productivity in U.S. Universities

Amy Apon[1]
University of Arkansas
USA

Stanley Ahalt[2]
University of North Carolina
USA

Vijay Dantuluri[3]
University of North Carolina
USA

Constantin Gurdgiev[4]
IBM & Trinity College
Ireland

Moez Limayem[5]
University of Arkansas
USA

Linh Ngo[6]
University of Arkansas
USA

Michael Stealey[7]
University of North Carolina
USA

Abstract

This paper studies the relationship between investments in High-Performance Computing (HPC) instrumentation and research competitiveness. Measures of institutional HPC investment are computed from data that is readily available from the Top 500 list, a list that has been published twice a year since 1993 that lists the fastest 500 computers in the world at that time. Institutions that are studied include US doctoral-granting institutions that fall into the very high or high research rankings according to the Carnegie Foundation classifications and additional institutions that have had entries in the Top 500 list. Research competitiveness is derived from federal funding data, compilations of scholarly publications, and institutional rankings. Correlation and Two Stage Least Square regression is used to analyze the research-related returns to investment in HPC. Two models are examined and give results that are both economically and statistically significant. Appearance on the Top 500 list is associated with a contemporaneous increase in NSF funding levels as well as a contemporaneous increase in the number of publications. The rate of depreciation in returns to HPC is rapid. The conclusion is that consistent investments in HPC at even modest levels are strongly correlated to research competitiveness.

Keywords: US doctoral-granting institutions, research competitiveness.

Introduction

Modeling and simulation are central to modern science and engineering. The National Science Foundation, the Office of Science, and many other agencies and foundations identify computational science as the third leg of science, after analysis and experimentation. More recently, data-driven science has been called out as a fourth paradigm of science. Modeling and simulation, as well as data driven science, rely heavily on high-performance computers, also known as supercomputers. However, HPC investments can be costly, requiring substantial ongoing capital and operational investments. Furthermore, investments in HPC may additionally require investments in data center space, power and electricity, air conditioning, high performance network access, and highly skilled staff support.

Thus, it is important that the value realized through investments in HPC, as quantified by research productivity, be investigated carefully. This paper specifically attempts to quantify research productivity as it is related to investment in large scale computational resources. This research studies the relationship between the investments in HPC systems and the changes in outcomes of research activities of an academic institution. Researchers at many institutions will have access to small sized computing clusters or high-end workstations that are used for computational research. These systems do not typically represent large investments, and the number and size of these small systems that are used by researchers on campuses is difficult to measure. For this study, an institution's investment in HPC systems is measured by considering the relatively large investments that are represented by entries by that institution on the Top 500 HPC List.

The Top 500 list by been published each year in June and November since 1993, and contains a compilation of the fastest 500 computers in the world as measured by the performance of a particular dense matrix algebra calculation, High Performance LINPAC (HPL). In each list the fastest computer at that time in the world has rank #1. An entry by an institution on the top 500 list indicates a substantial monetary investment in a powerful HPC system, and also signifies a significant commitment by the institution for HPC with the efforts to run the top 500 list's benchmarks as well as to report the results to the list.

Since an entry on the Top 500 list is voluntary, it does not include all the computational resources available to academic researchers in the United States. An institution may have made a significant investment in computational resources without ever having had an entry in the Top 500 list if it does not report its benchmark results. In spite of this shortcoming, the Top 500 list represents an historical record without peer of the performance of supercomputers that have been located at many top academic institutions in the United States. No other set of data describes the location of these top performing computers as well as the Top 500 list.

There were 43 U.S. academic institutions that appeared on the first Top 500 list in 1993. By 2009, more than 100 U.S. academic institutions had appeared on a Top 500 list. Figure1 illustrates the number of unique U.S. academic institutions on each list (bottom line) and the number of entries by U.S. academic institutions on each list (top line). The number of unique U.S. academic institutions as they appear cumulatively is shown by the shaded region.

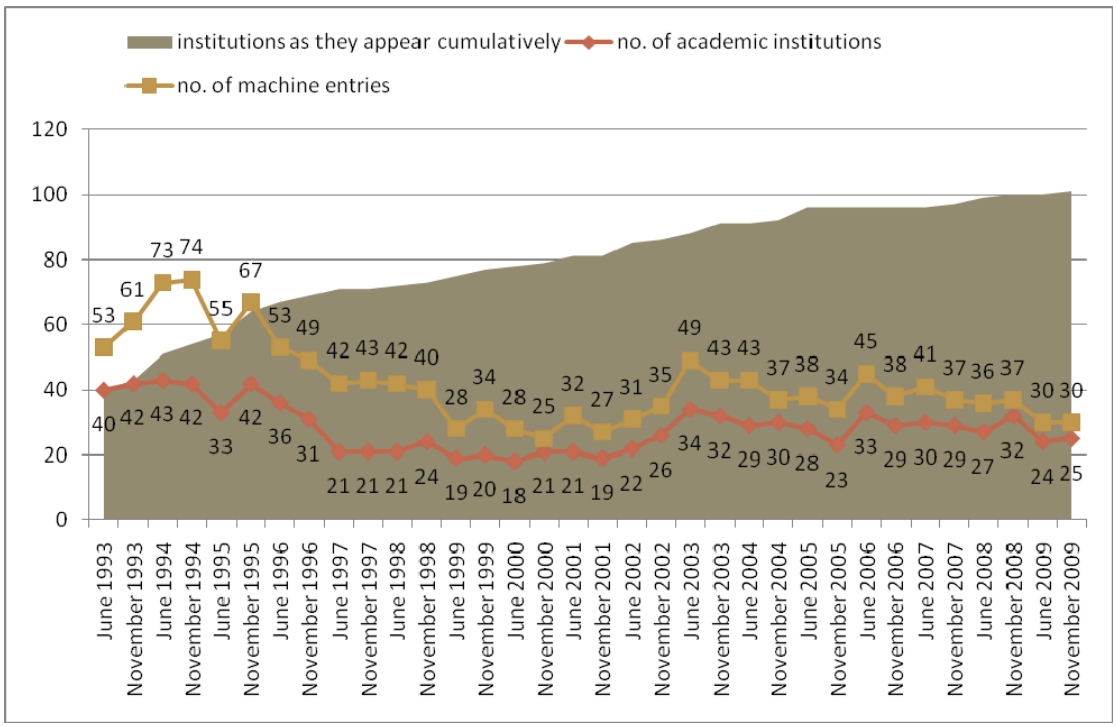

Figure 1. Number of U.S. Academic Institutions on Each Top 500 List, by List Year and Month

Previous papers that examined the Top 500 list cover many topics ranging from technologies, architectures, and to future trends of the systems on the list. Some of the literature gives an overview of the diversity of architectural approaches as well as the vendors for the systems (Simon, 1995; Dongerra & Simon, 1996; Dongerra & Simon, 1997; Strohmaier, Dongarra, Meuer & Simon, 1997; Dongarra, Meuer, Simon & Strohmaier, 2001; Strohmaier & Meuer, 2004). The work of Feitelson provides a statistical analysis of the usage pattern and evolutionary trends of the systems on the Top 500 list (Feitelson, 1999). Ripeanu reports in 2006 that it is becoming more rewarding to invest in an aggregation of small machines' computing power (Ripeanu, 2006). Meuer offers a comprehensive analytical look at the history of the Top 500 list and discusses the need for additional benchmarks in order to satisfy the different style of computation requirements (Meuer, 2008).

The remainder of this paper is organized as follows. Section 2 describes research measures that have been obtained and their use in measuring research competitiveness. Section 3 is a detailed statistical analysis of the inputs and outcomes of research that is supported and enabled by HPC. Section 4 provide discussion and conclusions, and describes areas for future investigation.

Research Measures

The set of institutions considered in this study are taken from the Carnegie Foundation list of approximately 200 colleges and universities that have very high or high research activity, and five additional institutions that have also made investments in HPC as documented by entries on the Top 500 list.

While the data from the Carnegie Foundation (Carnegie Foundation) is a list of institutions and institutional characteristics, the entries in the Top 500 list do not include the institutional location of the machine. That is, each entry in a Top 500 list is a computer and an associated "supercomputer site". The supercomputer site can be a university, such as "Mississippi State University", or, the supercomputer site may be a supercomputer center such as the "Ohio Supercomputer Center". The Top 500 entry information does not provide the mapping from the supercomputer site to the institution. To further complicate the matters, entries in the Top 500 list are entered by different individuals over time, the names of some supercomputer sites have changed over time, and the data contains misspellings and abbreviations. Finally, supercomputer sites may be affiliated with institutions via relationships that cannot be known from any of the available datasets. In these cases, anecdotal information from the supercomputer community has been used to associate a supercomputer center with its institution.

The names of the supercomputer sites in the Top 500 list were matched with associated universities using an automated process augmented by a manual process. The "unitid" from the Carnegie Foundation data set is used as the unique identifier of the institution, and the "name" from the Carnegie Foundation data set is used as the institution name. Approximately half of the institutions in the study set have had an entry on the Top 500 list at some time, and the other half have not.

One of the key metrics in this study is external research funding and its correlation to institution competitiveness. The National Science Foundation (NSF, www.nsf.gov) provides summarized funding data from federal sources not including National Institute of Health (NIH, www.nih.gov) funding. The NIH provides data as part of the award search information. Institutions whose names closely match the Carnegie foundation institution names were selected to be in the study set.

Another key metric in this study is the count of publications by researchers at an institution. The Institute for Scientific Information (ISI) has maintained some of the most detailed citation databases of scientific journals (Thompson Reuters, 2010). The ISI database, provided by Thomson Reuters through the Web of Knowledge portal, is used to determine the publication counts of different institutions.

In order to establish a reliable value for the count of publications that can be duplicated, a two-pass search procedure was used (Toutkoushian, Porter, Danielson, & Hollis, 2003). The first pass identifies the appropriate (official) institution name used in ISI system. This is done by a searching on a combination of institution names and addresses as well as the utilization of ISI's analysis tool. Secondly, a search is performed using this ISI designated name. The number of results acquired from this second search is the value recorded for the count of publications for institutions in the study set, and used in the subsequent analysis.

An additional metric of research competitive that is considered in this study is the National University Rankings from USNews.com (www.usnews.com/rankings). The US News and World Report list of college rankings judges an institution's relative effectiveness in a broad spectrum of categories based on quantitative measures that experts have proposed as reliable indicators of academic quality. Institutions with ranks up to 128 for the year 2009 have been mapped to Carnegie Foundation institution names. Other institutions that have a rank lower than 128 fall into a classification tier and this metric is not used.

Data Analysis

The basic hypothesis of this study is that investments by an institution in HPC lead to higher research outcomes. That this would be true seems natural. HPC resources are high-end research tools, and institutions that use them effectively should be better able to produce high quality research.

The approach is to first consider a correlation analysis of the summary data that is available. In a second step a regression analysis is performed on the aggregate data, again to provide evidence of the impact of HPC on various research outcomes. Two models are presented.

The variables of interest are listed below. Variables of interest that are derived from Top 500 list entries are the derived rank, and the count of lists on which an institution appears. The derived rank is defined as 501-Top 500 rank. The most capable computer on any particular Top 500 list has derived rank equal to 500. The least capable computer on any particular Top 500 list has a derived rank equal to 1, which is still very fast. Since an institution may have multiple entries in any particular list, the sum of the derived ranks for an institution for a particular list can be more than 500. The complete list of variables for the study is described as follows:

Variable	Description
dRankSum	Sum of derived ranks
Counts	The count of lists on which an institution has appears
NSF	Sum of NSF funding for the institution
Pubs	Sum of publications
FF	Sum of federal funding
DOE	Sum of DOE funding
DOD	Sum of DOD funding
NIH	Sum of NIH funding
USNews	US News and World Report ranking from 2009

Correlation Analysis

A correlation analysis is used to measure the strength of the relationship between two variables. Table 1 shows the correlation coefficient for each pair of variables. With the exception of USNews, which is the value in the year of 2009, the remaining data values considered in the correlation analysis are the accumulated values from the year 1993 until 2007, the years for which funding and other data is available, for each institution in the study set.

Table 1. Correlation Analysis for Institutional Summary Data, 204 Institutions

	Counts	NSF	Pubs	FF	DOE	DOD	NIH	USNews
dRankSum	0.8198	0.6545	0.2643	0.2566	0.2339	0.1418	0.1194	-0.243
Counts		0.6746	0.4088	0.3601	0.3486	0.1931	0.2022	-0.339
NSF			0.7123	0.6542	0.5439	0.2685	0.4830	-0.540
Pubs				0.8665	0.4846	0.3960	0.8218	-0.588
FF					0.4695	0.6836	0.9149	-0.543
DOE						0.1959	0.3763	-0.384
DOD							0.4691	-0.252
NIH								-0.500

Table 1 shows that for the 204 institutions in the study set, the dRankSum and Counts have a high correlation with NSF funding levels, .6545 and .6746, respectively. This result is consistent with expectations. Since the NSF supports science and engineering research in U.S. academic institutions, and HPC has traditionally been utilized mostly in areas of science and engineering research, it is expected that NSF funding will be highly correlated to the presence of HPC resources. Table 1 also shows that the correlation to both NSF funding and publication counts is somewhat higher for Counts than for dRankSum. This suggests that the presence of HPC resources is somewhat more significant for research productivity than the presence of a system with high rank on the Top 500 list.

The correlation of dRankSum and Counts to other federal funding measures is smaller. For example, the correlation of dRankSum to NIH funding is low, 0.1194. The correlation of Counts to NIH funding is similarly low, at 0.2022. This is not surprising since during the period of 1993-2007, HPC resources were not commonly used to support medical research.

The correlation of dRankSum to non-NSF funding is also low. Specifically, the correlation of dRankSum to all federal funding not including NIH (FF) is 0.2566, to DOE funding is 0.2339, and to DOD funding is 0.1418. The correlation of Counts follows a similar trend. These relatively low correlations suggest that academic HPC systems do not have a large impact on DOE and DOD funding to these institutions.

The correlations of all variables to US News and World Report rankings are negative because all variables are positive indicators (a higher number is better) except for the US News and World Report rank, where a lower number is better. For US News and World Report rankings the #1 school is considered better than the #128 school. The high negative correlation of publication counts to US News and World Report rankings suggests that, as expected, publication counts are a strong factor in achieving a better US News and World Report ranking.

The modest correlation of dRankSum to US News and World Report rankings of -0.243 and the modest correlation of Counts to US News and World Report rankings of -0.339 suggests that the presence of high performance computing resources at a U.S. academic institution is less responsible for institutional ranking as measured by US News and World Report rankings than other factors such as publication counts and NSF funded research.

There are other correlations in Table 1 that are not directly related our hypothesis, but which may warrant additional study. For example, the high correlation of NIH funding to both publication count and other federal funding is an indication that schools that do medical research are successful in other types of science as well. Perhaps less obviously, the higher correlation of US News and World Report rankings to publications, NSF funding, and NIH funding as compared to the lower DOE or DOD funding suggests that academic reputation, as measured by US News and World Report rankings, is less dependent on DOE and DOD funding than other types of research.

Regression Analysis

In this section, Two Stage Least Squares (2SLS) regression is used to analyze the research-related returns to investment in High Performance Computing. To do this, we looked at the overall sample of institutions with and without appearances on the Top 500 list. We model the two relationships between:

1. NSF Funding (NSF) as a function of contemporaneous and lagged Appearance on the Top 500 List Count (APP) and Publication Count (PuC), and
2. Publication Count (PuC) as a function of contemporaneous and lagged Appearance on the Top 500 List Count (APP) and NSF Funding (NSF)

Thus, Model 1 uses NSF as a dependent variable of PuC and APP. In Model 1, APP enters as contemporaneous variable, and one and two year lagged variables. Model 2 uses PuC as dependent variable of NSF and APP, with APP entering as contemporaneous variable, and one and two year lagged variables.

Original tests revealed significant problems with endogeneity of PuC and NSF. To correct for this, we deployed a 2SLS estimation method, with number of undergraduate Student Enrollments (SN) acting as an instrumental variable in the first stage regression for PuC (model 1) and NSF (model 2). In both cases, SN was found to be a suitable instrument for endogenous regressors.

The results in Table 2 show the estimation for Model 1.

Table 2. Model 1 for NSF funding

2SLS with fixed effects	Dependent variable	Number of observations	Number of groups	R-squared within	R-squared between	R-squared
	NSF funding (NSF)	2058	193	0.0180	0.0341	0.0272
	Coefficient	Std. Errors	t	P > |t|	95% Confidence Interval	
2SLS SN	9.211185	3.656232	2.52	0.012	2.040437	16.38193
APP (L0)	2419.682	841.5259	2.88	0.004	769.248	4070.116
APP (L1)	-1284.936	905.502	-1.42	0.156	-3060.842	490.9704
APP (L2)	-3121.393	852.5755	-3.66	0.000	-4793.498	-1449.288
Constant	888.5068	6351.752	0.14	0.889	-11568.8	13345.81

F(4,1861)=8.55 Prob>F=0.0000

We find both economically and statistically significant effects of contemporaneous APP on NSF funding levels. Each 1 point increase in overall Top500 ranking score is associated with a contemporaneous increase in NSF funding of USD 2,419,682 at the mid-point of estimated range of USD 769,248-4,070,116, relative to an institution's own past average funding.

However, this positive effect is associated with rapid depreciation of the overall returns to HPC investment as measured by NSF funding. Statistically, the previous year rank score within the Top 500 list has zero effect on NSF funding in the current year, while two years lagged rank score for HPC capability has a negative effect on NSF funding in the current year. Institutions that fail to have a persistent and consistent investment in HPC see a lack of persistent positive effect to NSF funding levels.

This rate of depreciation in returns to HPC can be potentially explained by a combination of factors. First and foremost, the above results deal with the returns due to HPC investment relative to institution-own past historical average of NSF funding. This means that an increase in the NSF funding in year 1 of investment in HPC increases the institutional average for subsequent years significantly enough to have a large long term effect. In subsequent years, therefore, it is natural to expect a decline in overall new NSF funding attributable to the HPC facility. Second, NSF funding relates to multi-annual grants which are captured at the point of award. Third, collaborative projects whereby NSF funding might be allocated to a number of institutions that jointly utilize one of the institutions' HPC facilities will tend to reduce overall future returns to APP if such collaborative grants are more likely to take place in years following original installation (and ranking in Top 500) of HPC facilities. Fourth, NSF grants applications are often pre-planned and can precede actual deployment and ranking of HPC facilities.

The above results are also confirmed with respect to returns on HPC investment measured by a dummy variable that takes 0 if the institution has no rank presence on Top 500 list and 1 if the institution has some presence on the list. These results are reported in Table 4. The result here is

slightly stronger, suggesting that any ranking on the Top 500 list (whether a rank of 1 or 2) improves NSF funding returns by USD 2,974,426 on average, relative to an institution's own past historical funding average.

The results in Table 3 show estimation for Model 2.

Table 3. Model 2 for Publications Count

2SLS with fixed effects	Dependent variable	Number of observations	Number of groups	R-squared within	R-squared between	R-squared
	Publications Count (PuC)	2105	193	0.0711	0.0675	0.0673
	Coefficient	Std. Errors	T	P > |t|	95% Confidence Interval	
2SLS SN	0.0750336	0.0067632	11.09	0.000	0.0617696	0.0882976
APP (L0)	59.57417	20.66175	2.88	0.004	19.05217	100.0962
APP(L1)	12.93856	22.08729	0.59	0.558	-30.37921	56.25632
APP (L2)	-62.58879	20.65567	-3.03	0.002	-103.0989	-22.07872
Constant	508.9459	110.8077	4.59	0.000	291.6289	726.2628

F(4,1908)=36.54 Prob>F=0.0000

Turning to the number of publications as a metric for return on HPC investment, Table 3 above reports the main findings of the model.

As before, we find that APP has an economically and statistically significant effect on overall publications produced by the investing institution, with each 1 point increase in overall Top 500 ranking scores associated with contemporaneous increase in the number of publications (relative to institution own past average number of publications) of approximately 60 at the mid-point of estimated range of 19-100. However, as before, we find this effect short-lived, with previous period APP score increases yielding statistically insignificant change in the number of publications (and increase of 13) and the overall effect turning negative for institutions with 2 periods of lagging improvements in APP. Potential reasons for this rapid depreciation of new HPC investment are also similar to those discussed in the case of Model 1 above.

At this stage in research, four conclusions are warranted from the data analyzed above:

1. HPC investment yields economically and statistically significant immediate returns in terms of new NSF funding available, relative to institution-own past historical average;
2. HPC investment yields economically and statistically significant immediate returns in terms of the increased number of academic publications produced, relative to institution-own historical past average number of publications;
3. It appears that HPC investments suffer from fast depreciation over the 2 year horizon; and
4. More research and data collection is needed to precisely determine the rate of depreciation of HPC investments

Overall, both models indicate that investment in high performance computing as measured by the entries on the Top 500 list is a good predictor of research competitiveness at U.S. academic institutions as measured by NSF research funding and Publications Counts. It is important to notice that PuC measure includes all publications, not just publications that are specific to HPC. Therefore, the current data on publications can be improved in the future by explicitly identifying publications related to HPC. Nonetheless, the last two results listed above suggest that institutions that have attained in the past significant returns from investment in HPC cannot rest on laurels. Maintaining strong investment in High Performance Computing is associated with strong, but quickly deprecating returns in terms of both new funding and new publications.

Table 4. Additional Model estimation for NSF Funding

2SLS with fixed effects	Dependent variable	Number of observations	Number of groups	R-squared within	R-squared between	R-squared
	NSF Funding	2058	193	0.0177	0.0219	0.0183
	Coefficient	Std. Errors	T	P > |t|	95% Confidence Interval	
2SLS SN	9.38458	3.657153	2.56	0.011	2.185906	16.53101
APP (L0)	2974.426	1436.823	2.07	0.039	156.4726	5792.379
APP(L1)	-2150.7	1494.166	-1.44	0.150	-5081.119	779.718
APP (L2)	-5616.667	1407.145	-3.99	0.000	-8376.414	-2856.919
Constant	881.2712	6353.069	0.14	0.890	-11578.62	13341.16

F(4,1861)=8.38 Prob>F=0.0000

Discussion and Conclusions

We have studied the relationship between investments in High-Performance Computing (HPC) and research competitiveness. Using publically available data drawn from a number of sources, we have shown that consistent investments in HPC at even modest levels are strongly correlated to research competiveness. The correlation between the capability of the machines that are purchased with the investments and indicators of research competitiveness is positive but less strong. The capability of the machines seem to strongly moderate the value of persistent investments. From these data we conclude that modest, but consistent investment in HPC results in measureable increases in research competiveness, and a corresponding increase in research funding and publication counts.

Acknowledgement

Acknowledgement is given to Conor O'Toole (Trinity College, Dublin) for research assistance. This work was supported by Grant #0946726 from the National Science Foundation.

References

Dongarra, J., Meuer, H., Simon, H., & Strohmaier, E. (2001). Biannual TOP-500 Computer Lists Track Changing Environments for Scientific Computing. *SIAM News*, 1 ff.

Dongerra, J., & Simon, H. (1996). High performance computing in the US in 1995 - An analysis on the basis of the TOP500 list. *Supercomputer*, 16-22.

Dongerra, J., & Simon, H. (1997). High performance computing in the US in 1996 - An analysis on the basis of the TOP500 list. *Supercomputer*, 19-28.

Feitelson, D. (1999). On the interpretation of TOP500 Data. *International Journal of HIgh Performance Computing Applications*, 146-153.

Meuer, H. (2008). The TOP500 project: Looking back over 15 years of supercomputing experience. *Informatik-Spektrum*, 203-222.

Ripeanu, M. (2006). A note on zipf distribution in top500 supercomputers list. IEEE Distributed Systems Online.

Simon, H. (1995). High performance computing in the US in 1994 - An analysis on the basis of the TOP500 list. *Supercomputer*, 21-30.

Strohmaier, E., & Meuer, H. (2004). Supercomputing: What have we learned from the TOP500 project? *Computing and Visualization in Science*, 227-230.

Strohmaier, E., Dongarra, J., Meuer, H., & Simon, H. (1997). High-performance computing in industry. *Supercomputer*, 74-88.

Thompson Reuters. (2010). *Web of Knowledge*. Retrieved from http://isiwebofknowledge.com/

Toutkoushian, R., Porter, S., Danielson, C., & Hollis, P. (2003). Using publications counts to measure an institution's research productivity. *Research in Higher Education*, 121-148.

---

[1] Dr. Amy Apon is a Professor of the Department of Computer Science and Computer Engineering at the University of Arkansas and Director of the Arkansas High Performance Center. She can be reached at 515 JBHT, University of Arkansas, Fayetteville, Arkansas 72701. Email: aapon@uark.edu; Phone: 479 575 6794.

[2] Dr. Stanley Ahalt is the Director of the Renaissance Computing Institute (RENCI) of the University of North Carolina at Chapel Hill. He can be reached at 100 Europa Drive, Suite 540, Chapel Hill, NC 27517. Email: ahalt@renci.org

[3] Vijay Dantuluri is the System Specialist of the Renaissance Computing Institute (RENCI) of the University of North Carolina at Chapel Hill. He can be reached at 100 Europa Drive, Suite 540, Chapel Hill, NC 27517. Email: vijayd@renci.org

[4] Dr. Constantin Gurdgiev is the Head of Macroeconomics with the Global Central of Economic Development, IBM Institute for Business Value and a lecturer in Finance with Trinity College, Dublin. He can be reached at Email: gurdgiev@ie.ibm.com; Phone: +353 1 815 4793

[5] Dr. Moez Limayem is the Associate Dean for Research & Graduate Studies at the Walton College of Business at the University of Arkansas. He can be reached at 328K WCOB, University of Arkansas, Fayetteville, Arkansas 72701. Email: mlimayem@walton.uark.edu

[6] Linh Ngo is a Ph.D. candidate at the the Department of Computer Science and Computer Engineering at the University of Arkansas in Fayetteville. He can be reached at 448 JBHT, University of Arkansas, Fayetteville, Arkansas 72701. Email: lngo@uark.edu

[7] Michael Stealey is the Senior Research Software Developer of the Renaissance Computing Institute (RENCI) of the University of North Carolina at Chapel Hill. He can be reached at 100 Europa Drive, Suite 540, Chapel Hill, NC 27517. Email: stealey@renci.org