A BENCHMARKING INDEX TO COMPARE HIGH-PERFORMING COMPUTING SYSTEMS

Corrado lo Storto and Benedetta Capano
Department of Industrial Engineering, University of Naples Federico II, Naples, Italy
E-Mail: corrado.lostorto@unina.it

ABSTRACT

An index to compare supercomputers is proposed in the study. This index is based on the concept of technical efficiency and is developed adopting a non-parametric technique, e.g. Data Envelopment Analysis. The index is used to calculate the technical efficiency of 500 high-performing computing systems listed in the TOP500 supercomputers database. Finally, statistical analysis is performed to assess the weight that some supercomputers characteristics have on their efficiency.

Keywords: high-performance computing, benchmarking, non-parametric analysis, data envelopment analysis, efficiency.

INTRODUCTION

Over the last 20 years, the global high-performance computing landscape has evolved rapidly. The USA, Japan and China have driven the supercomputer technology development over time in order to satisfy the need of computationally intensive tasks in science, engineering and business fields [1]. The history of supercomputers dates back to the 1960s, when Seymour Cray designed the Atlas at the University of Manchester and, later, a series of computers at Control Data Corporation (CDC). These systems were defined as "supercomputers" because of the innovative design and parallelism used by Cray to achieve superior computational peak performance. In 1972 Cray left CDC to form his own company, Cray Research. In 1976 he introduced the Cray-1 supercomputer into the market and later, in 1985 Cray-2. The latter was the world's fastest supercomputer until 1990. After the 1990s, faster supercomputer systems were designed in the USA and Japan by using thousands of processors and, as a result, new computational performance records were set. China emerged later on the scene, developing high performing computers and contributing substantially to the progress of microelectronics and the computing technology. The advancement of supercomputers can be tracked by looking at the lists drawn up for the TOP500 supercomputer project, which was launched at the University of Mannheim (Germany) in 1993. Since then, two TOP500 lists have been published every year providing a computer ranking based on their performance [2-3]. Today, the TOP500 supercomputer lists show that the United States are still the clear leader of the market with 46% of the top supercomputer systems, followed by China (12%) and Japan (6%).

One of the main aspects linked to the computational task of supercomputers is their architecture. Supercomputers are not designed according to the Von Neumann's architecture used by standard computers. Instead they use alternative architectures, such as massively parallel processing (MPP) and cluster computing, which allow a small number of operations on a large number of data to be performed more efficiently [4]. The MPP architecture has multiple processors running in parallel and linked with the motherboard. In other words a MPP is a single computer with many networked processors. Cluster computing, on the other hand, consists of a set of connected computers that work together as a single system [5]. With regard to the architectures of supercomputers, the TOP500 list from 2008 to 2014 shows that 85% of the total number of supercomputers uses clusters while only 15% uses MPP.

The evolution of the architecture of supercomputers has affected performance capabilities, such as the number of floating point operations per second (FLOPS). In the 1960s, when the first supercomputers were introduced, they were characterized by one million FLOPS (megaflops). By increasing the number of processors, the number of FLOPS increased too. By the end of the 20th century, MPP systems achieved computing teraflop ranges with trillions of operations and, later, with the introduction of cluster computing, the first supercomputers characterized by a quadrillion floating point calculations per second (petaflops) had been introduced [6-7]. If the progress of supercomputers continues at this rate, by 2020 the first exascale machines will appear with speed performance at least 30 times faster [8]. The number of FLOPS depends on the total number of cores and the processor speed (in MHz) which are both given in the TOP500. It must be noted that the faster the clock speed (in Hertz) the processor has and the more cores there are, the more flops the processor will perform. In order to describe system performance, the TOP500 provides other measures such as the maximal achieved performance (Rmax) and the theoretical peak performance (Rpeak) scores of supercomputers. Rpeak is a measure of the theoretical maximum number of floating point operations that the system can perform, and Rmax is the maximal rate at which the system can run the Linpack benchmark by solving a dense system of linear equation [2-4]. Both Rmax and Rpeak are measured in FLOPS.

Benchmarking has become an important tool to improve product and process performance in several industries [10-15]. In the high tech industries such as the high-performance computing systems it is particularly useful to measure the progress of the technology and compare performance changes of products over time. The

TOP500 supercomputer project provides a ranking of systems, but it is based only on the ratio Mflops to Watt (i.e., speed to power) and, consequently, unable to take into account several attributes of the system in the same time. This paper adopts a non-parametric method to develop an efficiency index that compares high-performance computing systems and generates useful benchmarks. It has the following organization. The second section presents the non-parametric approach to measure the supercomputer efficiency. In the third section, the study setting and model specifications to implement the method to compare computers in the TOP500 list are illustrated, while in the fourth section the results of the benchmarking analysis are discussed. Finally, the last section provides conclusions and summarizes this work.

A NON-PARAMETRIC APPROACH TO MEASURE EFFICIENCY

Introduced in 1978 by Charnes *et al.* [16], Data Envelopment Analysis (DEA) is a technique for decision making which provides the relative efficiency of a homogeneous set of units, known as decision making units (DMUs). When DMUs use multiple inputs to produce multiple outputs (MIMO problems), the efficiency scores are defined as the ratio of the weighted sums of outputs to the weighted sums of inputs. In particular, the Charnes-Cooper-Rhodes (CCR) model is formulated in terms of the following fractional programming to calculate the relative efficiency of the DMU *p* [16]:

$$Max \quad \frac{\sum_{k=1}^{s} v_k y_{kp}}{\sum_{j=1}^{m} u_j x_{jp}}$$

$$s.t. \quad \frac{\sum_{k=1}^{s} v_k y_{ki}}{\sum_{j=1}^{m} u_j x_{ji}} \le 1 \qquad i = 1,...,n \tag{1}$$

$$v_k, u_j \ge 0 \qquad k = 1,....,s, j = 1,...,m$$

where *n* is the number of DMU, *m* is the number of inputs and *s* is the number of outputs. Also, y_{ki} and x_{ji} represent the amount of output *k* and input *j* produced by DMU *i*, respectively. Lastly, v_k and u_j are the weights given to output *k* and input *j*, respectively.

The above fractional program can be converted into the primal CCR model and the dual CCR model, which are linear programs (LP). The following dual model is easier to solve because of its reduced calculation size:

$$Min \quad \theta + \varepsilon \, [\sum_{j=1}^{m} S_j^- + \sum_{k=1}^{s} S_k^+]$$

$$s.t. \quad \sum_{i=1}^{n} \lambda_i x_{ji} + S_i^- = \theta x_{jp} \qquad j = 1,...,m$$

$$\sum_{i=1}^{n} \lambda_i y_{ki} - S_k^+ = y_{kp} \qquad k = 1,....,s \tag{2}$$

$$\lambda_i \ge 0 \qquad i = 1,...,n$$

$$S_j^-, S_k^+ \ge 0 \qquad j = 1,...,m, \; k = 1,....,s$$

where θ is the efficiency score and λ_i the dual variables.

DEA formulations can be either input or output oriented. In the case of an input-oriented model, DEA measures the ability of DMUs to produce a given set of outputs with the minimum amount of inputs. The output-oriented model maximizes the amount of outputs while controlling the set of inputs. The obtained efficiency score, in both input and output orientation, is denominated technical efficiency (TE) [17].

In order to identify the relative efficiency scores of all the DMUs, the problem has to run *n* times. Efficiency scores range from 0 to 1; if a DMU has a score of 1 it will be considered efficient, whereas if the score is lower than 1 the DMU will be deemed inefficient. For each inefficient DMU, efficient DMUs can be used as benchmark units for improving performance and overcoming inefficiencies [18].

The main criticisms the described DEA model presents are related to the random noise of the data and the difficulty in applying statistical inference on efficiency scores [19]. In order to overcome these limitations, Simar and Wilson introduced the bootstrapped approach [20] and provided an ad hoc algorithm for estimating the bias corrected efficiency scores and confidence intervals [21].

STUDY SETTING AND MODEL SPECIFICATION

Method

The supercomputers technical efficiency TE is measured by implementing an input-oriented DEA model and assuming both constant returns to scale (CRS) and variable returns to scale (VRS). In order to have variable returns to scale, an additional convexity constraint $\Sigma\lambda_i = 1$ had to be considered in the CCR model that assumes constant returns to scale to formulate the Banker-Charnes-Cooper BCC model [22]. As the VRS model measures the pure technical efficiency the efficiency scores always satisfies the inequity $\theta(VRS) \ge \theta(CRS)$. In this study the DEA-bootstrapping is carried out by using the R FEAR package and performing 2,000 replications with alpha equal to 0.05 [23]. The bootstrapped VRS DEA efficiency is used to perform statistical analysis in order to assess the weight that the supercomputers characteristics (inputs and outputs) have on their technical efficiency. To this aim, the truncated regression technique has been performed as suggested in literature, assuming 0 and 1 as lower and upper limitations [21].

Table-1. Input and output variables.

type	variable	description
input	Power	the average power consumption (measured in kilowatts or kW) of a supercomputer while achieving R_{max}
	Number of cores	the total number of central processing units (called cores)
output	Performance	geometric mean of R_{max} (the maximal LINPACK performance achieved, measured in GFLOPS) and R_{peak} (the theoretical peak performance, measured in GFLOPS)

Sample

Data have been collected in January 2015 for a sample including 500 high-performance computers from the Top500 database freely available online [24]. This database contains supercomputers ranked by their performance on the LINPACK Benchmark [25].

Variables

The input and output variables displayed in Table-1 have been used to specify the DEA model implemented to calculate technical efficiency.

In particular, both input variables were directly available in the Top500 database, while the output variable was obtained as the geometric mean of R_{max} and R_{peak}. The choice to use only one output variable obtained as a combination of two output variables was suggested by the high correlation of these latter.

RESULTS

Table-2 shows statistics relative to the input and output variables used to calculate supercomputer technical efficiency. Data indicate that supercomputers differ to a large extent as all variables have a great standard deviation. For instance, power varies from 35 kW to 19,431 kW, while R_{max} varies from 153,381 GFLOPS to 33,862,700 GFLOPS.

Table-2. Statistics relative to the whole sample.

Variable	Mean	St.dev.	Maximum	Minimum
Power	**1,185**	**1,763**	**19,431**	**35**
No. of cores	46,288	168,455	3,120,000	2,992
R_{max}	615,726	2,024,264	33,862,700	153,381
R_{peak}	907,005	3,036,893	54,902,400	170,394

In order to perform DEA variables have been preliminarily normalized, generating measurements between 0 and 1. Table-3 displays the supercomputer efficiency measurements calculated adopting both the CRS and VRS approach. Table-3 also contains the bootstrapped efficiency scores for both approaches (CRSboot and VRSboot). The mean CRS and VRS efficiency scores are 28.9% and 31.7%, while the lower CRS and VRS efficiency scores are 6.5% and 7.2%, respectively. Differences between CRS and VRS efficiencies indicate that the computing systems size (i.e., scale) has an effect on their technical efficiency. The number of 100% efficient supercomputers is 9 in the VRS approach and only 3 in the CRS approach.

Table-3. Efficiency measurements.

	CRS	CRSboot	VRS	VRSboot
mean	**0.289**	**0.272**	**0.317**	**0.277**
st.dev	0.160	0.143	0.184	0.141
maximum	1.000	0.882	1.000	0.858
minimum	0.065	0.056	0.072	0.055
no. 100% efficient supercomputers	3	-	9	-

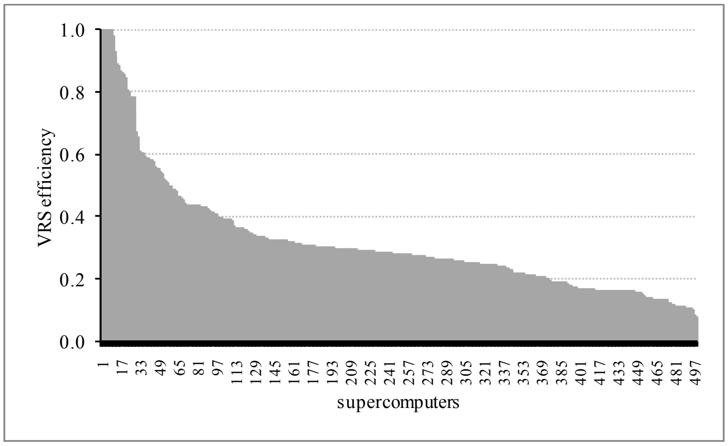

Figure-1. VRS efficiency plot.

Figure-1 plots the distribution of the whole sample VRS efficiency measurements sorted by size (grey area). The shape of the area is indicating that a large amount of supercomputers in the sample have technical efficiency between 40% and 20%, while only a small number of them have efficiency higher than 60%.

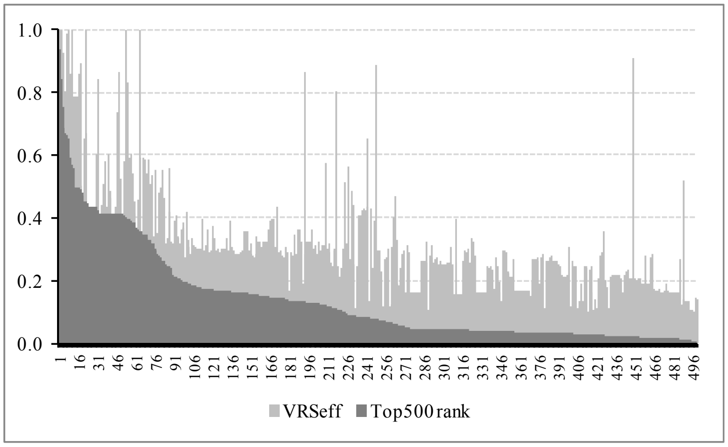

Figure-2. VRS efficiency vs TOP500 rank plot.

Figure-2 shows both the calculated VRS efficiency and the Top500 rank measurements (light and dark grey areas, respectively). Scores have been sorted according to the Top500 rank. Even though the two measurements have similar trends for the aggregate sample, the shapes of the two areas indicate that there may be important differences at the micro level (e.g., the individual supercomputer).

Finally, Table-4 reports the outcome of the truncated regression analysis conducted adopting the bootstrapped VRS efficiency as dependent variable, and the input and output variables of the DEA model as independent variables. As expected, both the coefficients of the “No. of cores” and “Power” variables have negative signs, while the coefficient of the “R_{max}” variable has a positive sign. The “R_{peak}” coefficient is not statistically significant. In terms of size of effects, R_{max} contributes more than No. of cores and Power to the technical efficiency measurement, while Power is less important than No. of cores.

Table-4. Truncated regression analysis.

Variable	Coefficient	z	Prob. |z|>Z*
constant	**-.00030*** **	**-35.44**	**.0000**
No. of cores	-2.36446***	-4.42	.0000
Power	-.87856***	-6.46	.0000
R_{max}	4.37089***	5.76	.0000
R_{peak}	-1.01040	-1.37	.1712
Sigma	.13478***	25.72	.0000
Log likelihood function		346.35932	
Inf.Cr.AIC		-680.7	
AIC/N		-1.361	

*** indicates significant at the one percent level. Threshold values for the model are 0, 1.

CONCLUSIONS

This paper has suggested an efficiency index useful to compare high-performing computers and conduct benchmarking studies. This index is calculated by implementing a non-parametric technique denominated Data Envelopment Analysis (DEA). A benchmarking analysis was conducted using the Top500 database. Two different DEA formulations have been adopted in the empirical study - the CRS and the VRS - the first one assuming constant returns to scale and the second one assuming variable returns to scale in order to take into account and assess the effect of the size of the system upon efficiency. To gain a more in depth knowledge of the single supercomputer characteristics that most influence their efficiency a regression analysis was carried on adopting the bootstrapped VRS DEA score as the dependent variable of a truncated regression equation.

The benchmarking study shows that the DEA VRS efficiency index is between 7.2% and 100%, while the CRS efficiency index is between 6.5% and 100%. The mean VRS and CRS efficiencies are 31.7% and 28.9%, respectively. Findings also indicate that the scale effect influences the efficiency score. The regression analysis suggests that the supercomputers characteristics contribute differently to the efficiency measurement. The efficiency score increases when R_{max} increases, while decreases when either the No. of cores or Power increases. Particularly, the effect of R_{max} is greater than the No. of cores and Power, while Power is less important than No. of cores.

The efficiency index is alternative to the ratio Mflops to Watt (i.e., speed to power) which is used in the Top500 database to rank supercomputers. Even though the proposed efficiency index generates only a relative ranking of the supercomputers, it provides a more comprehensive measurement of their performance as it takes into account several attributes of the system in the same time, supporting the benchmarking practice and developing useful information about how to compare supercomputers in multiple-dimension variables space and improve existing supercomputers or develop new supercomputers that more closely fit technology opportunities and trends, particularly taking into account both the market demand for greater performance and quality and the concerns for energy saving and sustainability [26-27].

REFERENCES

[1] Alspector, J., Brenner, A.A., Leheny, R.F. and Richmann, J.N. 2015. China - A New Power in supercomputing Hardware. (www.ida.org/~/media/Corporate/Files/Publications/IDA_Documents/ITSD/ida-document-ns-d-4857.ashx, retrieved on July 2, 2015).

[2] Feitelson, D.G. 2005. The supercomputer industry in light of the Top500 data. Computing in Science & Engineering, 7, pp. 42-47.

[3] Meuer, H.W. and Gietl, H. 2013. Supercomputers - Prestige Objects or Crucial Tools for Science and Industry? PIK - Praxis der Informationsvararbeitung und Kommunikation, 36.

[4] Dongarra, J., Meuer, H. and Strohmaier, E. 1997. TOP500 supercomputer sites. Supercomputer, 13, pp. 89-111.

[5] Marksteiner, P. 1996. High-performance computing - an overview. Computer Physics Communications, 97, pp. 16-35.

[6] Lim, D.J. and Kocaoglu, D.F. 2011. China - Can it move from imitation to innovation? Proceedings of the PICMET'11, Technology Management in the Energy Smart World, Portland (USA), pp. 1-13.

[7] National Research Council 2005. Getting Up to Speed: The Future of Supercomputing. Washington D.C: National Academies Press.

[8] Ashby, S., Beckman, P., Chen, J., Colella, P., Collins, B. and Crawford D. 2010. The opportunities and challenges of Exascale computing. U.S. Department of Energy Office of Science, pp. 1-71.

[9] lo Storto, C. 2013. Are Public-Private Partnerships a Source of Greater Efficiency in Water Supply? Results of a Non-Parametric Performance Analysis Relating to the Italian Industry. Water, 5, pp. 2058-2079.

[10] lo Storto, C. 2013. Evaluating ecommerce websites cognitive efficiency: An integrative framework based on data envelopment analysis. Applied Ergonomics, 44, pp. 1004-1014.

[11] lo Storto, C. 2014. A <value for money> framework to study product competitiveness in the automotive market. Journal of Engineering Science and Technology Review, 7, pp. 158-168.

[12] lo Storto, C. 2014. Benchmarking operational efficiency in the integrated water service provision: Does contract type matter? Benchmarking: an International Journal, 21, pp. 917-943.

[13] lo Storto, C. and Capano, B. 2014. Productivity changes of the renewable energy installed capacity: an empirical study relating to 31 European countries between 2002 and 2011. Energy Education Science and Technology Part A: Energy Science and Research, 32, pp. 3061-3072.

[14] lo Storto, C. and Capano, B. 2015. A dynamic efficiency analysis of the European renewable energy capacity between 2002 and 2011. Advanced Materials Research, 1079-1080, pp. 1274-1279.

[15] lo Storto, C. and Ferruzzi, G. 2013. Benchmarking Economical Efficiency of Renewable Energy Power Plants: A Data Envelopment Analysis Approach. Advanced Materials Research, 772, pp. 699-704.

[16] Charnes, A., Cooper, W.W. and Rhodes, E. 1978. Measuring the efficiency of decision making units. European Journal of Operational Research, 2, pp. 429-444.

[17] Farrell, M.J. 1957. The Measurement of Productive Efficiency. Journal of Royal Statistical Society Series A, 120, pp. 253-281.

[18] Sexton, T.R., Silkman, R.H. and Hogan, A.J. 1986. Data envelopment analysis: critique and extensions. In Silkman R.H. (ed.), Measuring Efficiency: an Assessment of data Envelopment Analysis, San Francisco, CA: Jossey-Bass, pp. 73-105.

[19] Simar, L. and Wilson, P.W. 2000. Statistical inference in nonparametric frontier models: the state of the art. Journal of Productivity Analysis, 13, 49-78.

[20] Simar, L. and Wilson, P.W. 1998. Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models. Management Science, 44, pp. 49-61.

[21] Simar, L. and Wilson, P.W. 2007. Estimation and inference in two-stage, semi-parametric models of productive efficiency. Journal of Econometrics, 136, pp. 31-64.

[22] Banker, R.D., Charnes, A. and Cooper, W.W. 1984. Some models for estimating technical and scale inefficiencies in data envelopment analysis. Management Science, 30, pp. 1078-92.

[23] Wilson, P.W. 2008. FEAR 1.0: A Software Package for Frontier Efficiency Analysis with R. Socio-Economic Planning Sciences, 42, pp. 247-254.

[24] TOP500. 2015. http://www.top500.org (list retrieved on January 25, 2015).

[25] Dongarra, J., Luszczek, P. and Petitet, A. 2003. The LINPACK Benchmark: past, present and future. Concurrency and Computation: Practice and Experience, 15, pp. 803-820.

[26] Savino, M.M. and Batbaatar, E. 2015. Investigating the resources for Integrated Management Systems within resource-based and contingency perspective in manufacturing firms. Journal of Cleaner Production, 104, pp. 392-402.

[27] Savino, M.M. and Mazza, A. 2014. Toward Environmental and Quality Sustainability: An Integrated Approach for Continuous Improvement. IEEE Transaction on Engineering Management, 61, pp. 171-181.