On Measuring Fidelity of Estimation Models

Haris Javaid   Sri Parameswaran

University of New South Wales, Australia
{harisj,sridevan}@cse.unsw.edu.au

Technical Report
UNSW-CSE-TR-1010
March 2010

THE UNIVERSITY OF
NEW SOUTH WALES

School of Computer Science and Engineering
The University of New South Wales
Sydney 2052, Australia

Abstract

Estimation models play a vital role in many aspects of day to day life. Extremely complex estimation models are employed in the design space exploration of SoCs, and the efficacies of these estimation models are usually measured by the absolute error of the model compared to a known actual result. Such absolute error based metrics can often result in over-designed estimation models, with a number of researchers suggesting that fidelity of an estimation model should be examined in addition to, or instead of, the absolute error. In this paper, for the first time, we propose four metrics to measure the fidelity of an estimation model, in particular for use in design space exploration. The first two are based on two well known rank correlation coefficients. The other two are weighted versions of the first two metrics, to give importance to points nearer the Pareto front. The proposed fidelity metrics were calculated for a single processor estimation model and a multiprocessor estimation model to observe their behavior, and were compared against the models' absolute error.

1 INTRODUCTION

The increasing SoC design productivity gap has necessitated the use of comprehensive design automation methodologies to ensure in-time delivery of reliable and flexible embedded devices at reduced prices. Design space exploration is a crucial part of all the design automation methodologies. In design space exploration, various algorithms and heuristics are used to search the design space for some global minima or maxima. Researchers heavily rely on estimation models to estimate the values of design points, especially where billions of design points are present in the design space, to create systems in short design times [1]. Thus, estimation is an important and critical part of most design space exploration methodologies.

To ensure that exploration algorithms provide optimal or near-optimal solutions, there is an expectation that the underlying estimation models need to be as accurate as possible. However, estimation models can be just as valid, if they exhibit good fidelity instead of just absolute accuracy. In fact, the authors in [2, 3] stated that the fidelity of an estimation model is more important than its absolute accuracy. In absolute accuracy, each estimated value is compared with the corresponding actual value and the absolute error is calculated. This is done for all the estimated values to calculate the average absolute error incurred by an estimation model to evaluate its suitability. Fidelity, on the other hand, measures the correlation between the ordering of the actual values and the ordering of the estimated values. A high correlation means the estimation model has a high fidelity relative to the actual values. Fidelity measures how well the estimated values track the actual values across different design points. An estimation model with 0% absolute error will also have a fidelity of 100%. An estimation model with non-zero absolute error can result in a fidelity value ranging from 0 to 100. An estimation model with low absolute accuracy but high fidelity (for example, estimated values are around twice the actual values but in the same order) will suffice the purpose of the design space exploration, where ordering of the design points is more important than the absolute accuracy to properly guide exploration algorithms. On the other hand, an estimation model with high absolute accuracy but low fidelity (for example, estimated values are very close to the actual values but are in highly erratic order relative to the ordering of the actual values) can misguide the exploration algorithms. Thus, measuring fidelity of an estimation model is more important from the perspective of exploration algorithms. Typically, designers use absolute accuracy to evaluate an estimation model, and ignore fidelity. In some cases, designers use a few design points or a graphical representation to visualize the correlation between the actual values and the estimated values [2, 3, 4, 5]. However, there exists no defined metric to measure the fidelity of an estimation model.

In this paper, for the first time, we propose fidelity metrics for measuring the fidelity of estimation models. Four fidelity metrics are shown which can be used to evaluate the ordering of the estimated values with respect to the ordering of the actual values. The first metric is the direct application of Spearman's rank correlation coefficient [6], ρ, introduced in 1904 by Charles Spearman, while the second metric is the direct application of Kendall's tau correlation coefficient [7], τ, introduced in 1938 by Maurice Kendall. In Spearman's ρ, first actual and estimated values are assigned ranks. Then, the differences between the ranks of the corresponding actual and estimated values are calculated to measure the disordering of the estimated values with respect to the actual values. Kendall's τ correlation coefficient, on the other hand, works on the principle of concordant and discordant pairs (in the set of estimated values) obtained with respect to the actual values.

The last two metrics are weighted metrics and are derived by augmenting ρ and τ to take into account the effect of Pareto front of a design space. A Pareto front is the set of the dominant points in the design space and reflects the optimal points of the design space [8]. Each actual value is assigned a weight depending on its distance from the Pareto front. Thus, actual values close to Pareto front are assigned higher weight than the ones further away from the Pareto front. In Spearman's ρ, the rank difference is multiplied by the corresponding weight of the actual value to suppress the effect of points that are far from the Pareto front on the fidelity metric. Similarly, in Kendall's τ, each concordant or discordant pair is multiplied by the corresponding weight to mitigate the effect of pairs that are far from the Pareto front. Since exploration algorithms typically search for the Pareto front of a design space or a point lying on the Pareto front, these weighted metrics are more intuitive and suitable for evaluating estimation models' fidelity. We evaluated these metrics on estimation models from two different domains: a single processor estimation model and a multiprocessor estimation model. Section 7 includes an insight of the results to show how the proposed metrics can be used to measure the efficacy of an estimation model.

The rest of the paper is organized as follows. Section 2 provides the necessary literature review. Section 3 provides a motivational example to emphasize the importance of fidelity in addition to absolute accuracy of estimation models. Section 4 provides the background knowledge on rank correlation coefficients, with the four fidelity metrics explained in Section 5. Section 6 and 7 present the experimental setup and the results, with the conclusion presented in Section 8.

2 RELATED WORK

Design space exploration is widely addressed with plenty of existing literature. Interested readers are referred to [1], where the authors have provided a good survey of estimation methods typically used for evaluating design points of a design space.

Typically, designers plot few design points' actual and estimated values to visualize the fidelity (correlation) between them [2, 3, 4, 5]. In [2], the authors proposed a system-level performance estimation methodology, [3] presented a performance estimation methodology for component-based embedded systems, [4] proposed an analytical estimation model for computation of delay under the transmission line model, while [5] introduced a novel substrate noise estimation technique to guide the floorplanning and layout optimization. All these papers plotted a few design points with their actual and estimated values to observe fidelity. The authors in [2] and [3] also emphasized the fact that relative ordering of the design points is more important than the absolute accuracy for design space exploration. However, none of these papers introduced any metrics to calculate the fidelity of estimation models.

Faria et al. [9] proposed a system-level performance evaluation methodology for network processors, where the fidelity of the proposed model is measured as the ratio of the absolute accuracies. Eyerman et al. [10] used a similar concept where the relative error between two design points is measured as the difference of the ratios of estimated values and actual values of the two points. The authors in [11], on the other hand, used Spearman's ρ to calculate the correlation between the ordering of the performance values obtained through cycle accurate simulation and statistical simulation, focusing on evaluating the efficacy of statistical simulation only. None of these works [9, 10, 11] have proposed to measure the fidelity of an estimation model in general. In contrast to all these works, we have adopted Spearman's ρ and Kendall's τ as fidelity metrics, which are widely used in the information retrieval domain [12, 13] to compare the rankings of the information retrieved through different methods. Furthermore, we proposed two more metrics based on Spearman's ρ and Kendall's τ to account for the effects of Pareto front of a design space, as finding the Pareto front (or a point lying on the Pareto front) is one of the most important objective of the design space exploration algorithms.

2.1 Our Contribution

In this paper, we propose four metrics to measure the fidelity of an estimation model. The first two metrics are the direct application of the standard correlation coefficients, Spearman's ρ [6] and Kendall's τ [7]. The last two metrics assign a weight to each point depending on its distance from the Pareto front of the design space, mitigating the effects of points far from the Pareto front. Using these metrics, designers can measure the fidelity of their estimation model(s), which they are using in their design space exploration frameworks, in addition to the measurement of absolute accuracy. To the best of our knowledge, this is the first work to adopt Spearman's ρ [6] and Kendall's τ [7] as fidelity metrics, and their augmentation with respect to Pareto fronts to evaluate the efficacy of estimation models used in design automation of embedded systems.

Please note that the calculation of fidelity metrics requires both the estimated values and the actual values. Calculation of absolute error too requires the availability of estimated and actual values. The fidelity metrics and absolute error are calculated for representative benchmarks and extrapolated for use in real designs.

3 MOTIVATION

In this section, we present a motivational example to emphasize the importance of the requirement of a fidelity metric for estimation models, which are widely used in design space exploration and design automation fields.

Let us examine the example given in Table 3.1, where the first column shows the actual values of a parameter, for example, the runtime of an application on a processor. The next two columns show the estimated runtimes using two different estimation models. The last two rows show the average absolute error and the fidelity error of both the estimation models respectively. The average absolute error is calculated by averaging the absolute error for all the six points, where the absolute error for the first point of estimation model 1 is $\frac{20{,}000-16{,}380}{16{,}380} \times 100 = 22.1\%$.

For calculating the fidelity error, the actual values are assigned ranks in the increasing order starting from 1 as shown in the parentheses in the first column. Then, the estimated values are sorted in increasing order and assigned ranks as well, which are shown for both the estimation models in parentheses in columns 2 and 3. Intuitively, it can be seen that the ordering of the estimated points from model 1 is identical to the ordering of the actual values, and thus the fidelity error is 0%. However, in the second estimation model, the points are in a different order, leading to a non-zero fidelity error. The fidelity error is calculated using,

$$FE \quad = \quad \frac{\sum_{i=1}^{n} r_i^2}{\sum_{j=1}^{n} (r_j^o)^2} \times 100 \tag{3.1}$$

Actual Values	**Model 1**	**Model 2**
16,380 (1)	20,000 (1)	18,800 (5)
16,900 (2)	20,600 (2)	16,550 (1)
18,100 (3)	21,800 (3)	18,700 (4)
18,800 (4)	22,600 (4)	18,650 (3)
19,500 (5)	23,000 (5)	18,600 (2)
20,100 (6)	24,000 (6)	20,200 (6)
Abs. Error (Avg.)	21.33%	4.34%
Fidelity Error	0%	40%

Table 3.1: Motivational Example - Comparing absolute error and fidelity of two estimation models

where n is the total number of points, r_i is the difference in the rank of the actual value and the rank of the corresponding estimated value for point i, and r_j^o is the difference in the rank of the actual value and the rank of the corresponding estimated value for point j such that the ranking sequence of estimated values is an exact opposite of the ranking sequence of actual values (o stands for exact opposite), being the worst case scenario. In other words, exact opposite means the estimated values are ranked in decreasing order. Thus, in this example, the denominator of Equation 3.1 will be $5^2 + 3^2 + 1^2 + 1^2 + 3^2 + 5^2 = 70$. For estimation model 2, the numerator will be $4^2 + 1^2 + 1^2 + 1^2 + 3^2 + 0^2 = 28$, making FE = 40%. Thus, an erratic relative ordering of the estimated points will lead to an increased fidelity error.

In this example, in the absence of fidelity error values, designers will choose estimation model 2 because of its low average absolute error. However, it should be noted that even though the estimation model 1 has an average absolute error of 21.33%, the estimated values are in the same order as the actual values. Thus, design space exploration algorithms will find the same global minima or maxima by using either the actual values or the estimation model 1. However, the fidelity error of estimation model 2 will probably result in a misguided solution. Thus, it can be concluded that an estimation model with low fidelity error but high absolute error can still be a good choice. Furthermore, this example signifies the importance of measuring the fidelity of estimation models from the perspective of design space exploration.

Traditionally, researchers used absolute error to measure the efficacy of an estimation model, which does not account for the fidelity of an estimation model. By fidelity, we mean the correlation between the ordering of the estimated values and the actual values. Such a correlation reflects how well the estimation model tracks the trend in actual values. The proposed metrics are useable in conjunction with the absolute error measurement to evaluate an estimation model more rigorously. Once a model with low fidelity error is found, designers do not need to improve its absolute accuracy. In addition, various estimation models can easily be compared and evaluated in terms of both absolute accuracy and fidelity with the help of the proposed metrics to choose the best model for later use in the design space exploration framework.

4 BACKGROUND

In this section, Spearman's rank correlation coefficient [6] and Kendall's tau correlation coefficient [7], two most widely used rank correlation coefficients from the statistics domain are described.

4.1 Spearman's rank correlation coefficient

Spearman's rank correlation coefficient, denoted as ρ, works on the same principle as the one shown in Section 3 for calculating the fidelity error of an estimation model. In general, it takes two data sets, X and Y. The raw values in X and Y, that is X_i and Y_i are converted into ranks X_i^r and Y_i^r, through sorting the data sets X and Y. Sum of the squared differences between the ranks of each pair (X_i, Y_i), that is, $\sum (X_i^r - Y_i^r)^2$ is calculated, which is then divided by the maximum possible sum of the squared rank differences between X and Y . The maximum possible sum of the squared rank differences occurs when the ordering of the points in X is opposite to the ordering of the points in Y, that is, the ranks in X^r are in increasing order while the ranks in Y^r are in decreasing order. Thus, ρ is defined as:

$$\rho \quad = \quad 1 - \frac{2 \times \sum_{i=1}^{n} r_i^2}{\frac{n(n^2-1)}{3}} \tag{4.1}$$

where $r_i = (X_i^r - Y_i^r)$ and n is the total number of points in each data set (both X and Y will have same number of points). The denominator $\frac{n(n^2-1)}{3}$ gives the maximum possible sum of the squared rank differences. Spearman's ρ always lies in the range $-1 \leq \rho \leq 1$ where a value of 1 signifies a perfect agreement between X and Y (correctly ordered), while a value of -1 signifies a perfect disagreement between the two sets (oppositely ordered).

4.2 Kendall's tau correlation coefficient

Kendall's tau correlation coefficient, denoted as τ, is based on the number of concordant and discordant pairs present in Y compared to X. A pair in X is defined as the combination of two points from X, (X_i, X_j) such that $i < j$. A pair in Y, (Y_i, Y_j), is concordant with respect to the corresponding pair in X, (X_i, X_j), if $sgn(X_j - X_i) = sgn(Y_j - Y_i)$ and discordant if $sgn(X_j - X_i) = -sgn(Y_j - Y_i)$ where the sgn function is defined as:

$$sgn(x) = \begin{cases} -1 : x < 0 \\ \;\;\,0 : x = 0 \\ \;\;\,1 : x > 0 \end{cases}$$

Thus, τ is defined as:

$$\tau \quad = \quad \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \tag{4.2}$$

where n_c is the number of concordant pairs, n_d is the number of discordant pairs, and n refers to the total number of points in each data set. The denominator $\frac{1}{2}n(n-1)$

gives the total number of pairs, resulting in a range of $-1 \leq \tau \leq 1$ for Kendall's τ. If all the pairs in Y are concordant with the corresponding pairs in X, meaning the points in Y are in the same order as the points in X, then $n_c = \frac{1}{2}n(n-1)$ and $n_d = 0$ making $\tau = 1$. Similarly, if all the pairs in Y are discordant, meaning the points in Y are in opposite order as the points in X, then $n_c = 0$ and $n_d = \frac{1}{2}n(n-1)$ making $\tau = -1$.

5 FIDELITY METRICS

As stated, fidelity correlates the ordering of the estimated values to the ordering of the actual values. In this section, the use of Spearman's ρ and Kendall's τ as the basis of fidelity metrics is demonstrated. For the sake of simplicity, the discussion in this section assumes a typical 2-dimensional design space, where each design point is associated with a 2-tuple number (Pf, Ar) – Pf and Ar represent the performance and area values respectively. In such a design space, each actual design point P_i^a has a corresponding estimated design point P_i^e. The set of all the actual design points is referred to as P^a while P^e refers to the set of estimated design points. In the discussion here, only performance values are estimated, which means that the area values of both P_i^a and P_i^e are the same, that is, actual area values are used with both actual and estimated performance values.

5.1 FM$_\rho$

FM_ρ is equal to Spearman's ρ explained in Section 4.1, where the performance values of P^a form the data set X, while the performance values of P^e form the Y data set. Since the area value of P_i^a and the corresponding P_i^e is the same, only the fidelity of the performance estimation model is calculated. The fidelity is calculated on the given X and Y sets using Equation 4.1. For example, in Table 3.1, for estimation model 2, column 1 becomes the X data set while column 3 becomes the Y data set. Given these X and Y sets, $\sum r_i^2 = 28$ while $n = 6$, resulting in $FM_\rho = 0.2$. Since estimation model 1 provides $FM_\rho = 1$, estimation model 2 is inferior to estimation model 1 with respect to fidelity.

FM_ρ provides a good measure of the fidelity of an estimation model. However, FM_ρ does not take into account the number of points that have been displaced in Y relative to X (the number of points whose corresponding ranks are different). Thus, for an estimation model where more than 90% of the points have a rank difference, but the difference in each rank is minor, the value of ρ will still be close to 1 due to a large value in the denominator. This discrepancy is reflected by the use of Kendall's τ correlation coefficient.

5.2 FM$_\tau$

FM_τ, as the name suggests, is the adoption of Kendall's τ, explained in Section 4.2, as the fidelity metric by utilizing performance values of P^a and P^e to form data set X and Y respectively. The fidelity is then calculated on these X and Y sets using Equation 4.2. For the estimation model 2 in Table 3.1, again the data set X is obtained from column 1 and the data set Y is obtained from column 3. For these X and Y sets, $n_c = 8$ and $n_d = 7$, resulting in $FM_\tau = 0.067$. This again shows that the estimation model 2 is inferior to estimation model 1 ($FM_\tau = 1$).

FM_τ inherently takes into account the effect of the number of points that have been displaced in Y relative to X. An ordering of the estimated performance values where more than 90% of the points have been displaced, but the displacement for each point is minuscule, will result in increased number of discordant pairs, which reduces the number of concordant pairs as well, in turn affecting the value of FM_τ to a larger extent compared to FM_ρ. For estimation model 2 in Table 3.1, 5 out of 6 points have been displaced (except the 6th point), resulting in a lower value for FM_τ compared to FM_ρ. Usually, Kendall's τ is lower than Spearman's ρ [14].

5.3 Weighted Metrics

Both FM_ρ and FM_τ are the result of the direct adoption of Spearman's ρ and Kendall's τ as fidelity metrics. However, FM_ρ assigns the same weight to all the points with a rank difference, while FM_τ assigns the same weight to all the concordant and discordant pairs. When exploring a design, typically the goal is to perform multi-objective optimization, which directly translates to finding the Pareto front or a point lying on the Pareto front of the design space. Intuitively, one can argue that an estimation model providing more design points that are close to the Pareto front in the correct order is better than a model providing more correctly-ordered design points that are far from the Pareto front. Such effects of the Pareto front can be accounted, by assigning a weight to each point based upon its distance from the Pareto front. A point closer to Pareto front is assigned a weight higher than the one far from the Pareto front. This also allows to extend the simple fidelity metrics (FM_ρ and FM_τ) which measure the fidelity for single-objective exploration algorithms, to target the measurement of fidelity from the multi-objective exploration algorithms' perspective.

A Pareto front is the set of dominant points from the design space and reflects the trend of the design space [8]. The calculation of Pareto front of a design space is usually referred to as the maximal vector computation problem [8]. There are numerous ways to obtain the Pareto front of an n-dimensional design space, a survey of which is provided in [8]. The work in this paper is not limited to any particular method of finding the Pareto front.

Let us assume the availability of the Pareto front of our typical 2-dimensional design space, which is shown in Figure 5.1, where the circles represent the actual design points, $P_i^a s$, while the asterisks connected through straight lines show the Pareto front of the design space. The Euclidean distance of each actual design point is calculated from all the lines on the Pareto front separately, and the minimum of all these distances is obtained. For example, in Figure 5.1, the distance of one of the design points is calculated separately for each of the 17 lines present in the Pareto front, and the minimum of all these 17 distances is obtained, represented as *d1* in the figure. Similarly, the distance of another point, further away from the Pareto front, is marked as *d2* in the figure. In this way, the minimum distance of each P_i^a is calculated to be used in a weight function. It should be noted, however, that the distance calculated as mentioned above may not be suitable for a weight function if the unit of measurements on both the axes differ by significant amounts. For example, if the performance is measured in seconds and area is measured in gates then the variations on y-axis may be very minute compared to the variations on the x-axis. Thus, the distance of all the points may be very close to each other, giving almost identical weights to all the points. To avoid such problems, we normalize the x and y values of the distance of each point from the lines in the Pareto front by the maximum range of values on x-axis and y-axis respectively. This normalizes the x and y values of the distance to the range of 0 to 1, giving a range

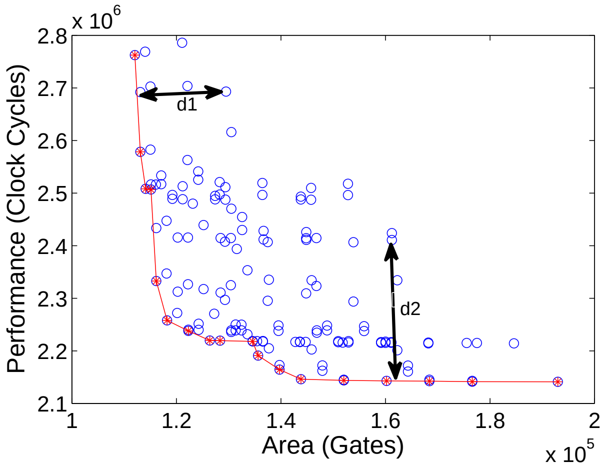


Figure 5.1: Pareto front of the design space consisting of actual design points

of 0 to $\sqrt{2}$ for the distance of each point. One may argue that the set of Pareto points be curve-fitted and then the distance of each actual point be calculated from the fitted curve. In such a case, it is possible that the fitted curve may not pass through all the Pareto points, thus will not reflect the actual Pareto front of the design space.

Once the distance of each actual point, P_i^a, has been calculated, a weight function can be used to assign different weights to different points depending on their calculated distances. The following weight function was used:

$$W = \frac{1}{1 + s \times d^k} \tag{5.1}$$

where s and k are constants, used to vary the amount of weight, and d is the minimum distance of the point from the Pareto front. A point with $d = 0$, that is a point on the Pareto front, will be given a weight of 1, which is the maximum possible weight. Points not on the Pareto front are assigned weights less than 1, decreasing the weights as the points move further away from the Pareto front. The values of s and k determine the decreasing nature of the weight function, and determine the suppression applied to points while moving away from the Pareto front. We explored different values of s and k and found that $s = 1000$ and $k = 1$ provide a reasonable weight function (exploration methodology is not described here due to the lack of space). Thus, all the results presented in Section 7 use $W = \frac{1}{1+1000d}$ as the weight function for the calculation of weighted fidelity metrics (explained later). If required, exploration of s and k can be performed by a designer in order to choose different values.

WFM$_\rho$

The procedure to calculate WFM_ρ is very similar to the one shown for FM_ρ. For WFM_ρ, first the Pareto front of the design space consisting of actual design points is obtained (this can be done by using any of the algorithms from [8]). Once the Pareto front is available, each actual design point is assigned a weight according to its distance from the Pareto front (the distance is calculated as explained in the last section) using Equation 5.1. As was the case with FM_ρ, the performance values in the set of actual design points, P^a, form the data set X, while the performance values of P^e form the data set Y. These X and Y sets are converted into ranks, X_i^r and Y_i^r, and then Equation 5.2 is used to calculate the value of WFM_ρ.

$$WFM_\rho = 1 - \frac{2 \times \sum_{i=1}^{n} W_i r_i^2}{\sum_{j=1}^{n} W_j (n + 1 - 2j)^2} \tag{5.2}$$

where W_i is the weight of the i^{th} point, $r_i = (X_i^r - Y_i^r)$, and n is the total number of points. The denominator gives the weighted sum of the squared rank differences such that the Y data set is ranked in decreasing order. $WFM_\rho \leq 1$ where a value of 1 means perfect ordering of Y with respect to X, while a value of -1 means the points in Y are in opposite order to X. The value of WFM_ρ can go below -1 in some cases because normalization of WFM_ρ in the range -1 to 1 is very difficult due to the presence of a product term ($W_i r_i^2$) in the numerator. Due to the lack of space, the details of different normalization techniques is not presented here. As most of the estimation models are developed intuitively, the value of WFM_ρ will typically be positive for any useful model, and Equation 5.2 will suffice for the purpose of measuring fidelity[1]. More points in the wrong order closer to the Pareto front will decrease the value of WFM_ρ, while more correctly-ordered points closer to the Pareto front will increase its value. In addition, $WFM_\rho = FM_\rho$ when $s = 0$ in Equation 5.1.

$\mathbf{WFM_\tau}$

The weighted version of Kendall's τ, WFM_τ, is based on a similar idea to WFM_ρ. The Pareto front of the actual design space is obtained and weights assigned to each actual point using Equation 5.1. Then the performance values of actual design points form the X data set, with the performance values of estimated design points forming the Y data set. The concordant and discordant pairs are calculated in the same way as it was calculated for FM_τ in Section 4.2. WFM_τ is then calculated as:

$$WFM_\tau = \frac{\sum_{i=1}^{n_c} W_{c,i} - \sum_{j=1}^{n_d} W_{d,j}}{\sum_{k=1}^{\frac{n(n-1)}{2}} W_k} \tag{5.3}$$

where $W_{c,i}$ is the weight of the i^{th} concordant pair, $W_{d,j}$ is the weight of the j^{th} discordant pair, and W_k is the weight of the k^{th} pair irrespective of being concordant or discordant. n_c and n_d refer to the total number of concordant pairs and discordant pairs respectively, while n is the total number of points in each data set. A pair is decided as concordant or discordant based on the two points which make up that pair. Thus, W_i of a pair is calculated as the minimum of the weights of the points that make up that pair. In contrast to WFM_ρ, the denominator in Equation 5.3 is the sum of the weights of all the pairs, resulting in a range of $-1 \leq WFM_\tau \leq 1$ for WFM_τ. More discordant pairs closer to the Pareto front will reduce the value of WFM_τ, while more concordant pairs closer to Pareto front will increase its value. In addition, $WFM_\tau = FM_\tau$ when $s = 0$ in Equation 5.1.

Comparing weighted metrics (WFM_ρ and WFM_τ) with the non-weighted ones (FM_ρ and FM_τ), area values of the points in P^a (same as the area values of the points

[1] Note that the fidelity of -1 can be just as good as 1 for the purpose of design space exploration. However, typically estimation models exhibit positive fidelity.

in P^e) are now used to compute the Pareto front of the design space. The weights assigned to each point depend on the Pareto front, thus using area values indirectly for the calculation of WFM_ρ and WFM_τ, which is not the case with FM_ρ and FM_τ. Excluding the complexity of computing the Pareto front and d of each point, the complexity of calculating FM_ρ and WFM_ρ is $O(n \log n)$ (assuming $n \log n$ sorting is used), and $O(n^2)$ for FM_τ and WFM_τ. The complexity of calculating d for all the points in P^a is $O(n^2)$.

5.4 Generalization of Fidelity Metrics

Thus far, the assumption has been that the design space under consideration is a 2-dimensional design space. We further assumed that only performance values are estimated in our typical performance-area design space. Now, the fidelity metrics are generalized to n dimensions.

There are no limitations to the number of dimensions of the design space for the calculation of the fidelity metrics. An n-dimensional design space can just be considered as well – this will require the computation of the Pareto front of an n-dimensional design space for which algorithms exist [8]. In addition, the range of d in Equation 5.1 will be $0 < d < \sqrt{n}$ for an n-dimensional design space. However, only one dimension can be estimated at a time from the perspective of measuring the fidelity. For example, in our typical 2-dimensional design space, only one dimension was estimated, that is, the performance values were estimated and the actual area values were used for both actual and estimated design points. This allows us to evaluate the performance estimation model's fidelity only. Typically, designers use various estimation models for estimating different dimensions of their design space. For example, in our typical performance-area design space, we can use two estimation models to estimate performance and area separately. In this case, the fidelity of performance estimation model and area estimation model should be calculated separately. The fidelity of performance estimation model is calculated by considering the performance estimation values with actual area values for both actual and estimated design points. The fidelity of area estimation model is calculated by considering the area estimation values with actual performance values for both actual and estimated design points. Since the Pareto front is obtained from the design space consisting of actual design points, it should be noted that the weight of each actual design point will be the same when calculating the fidelity metrics for either the performance estimation model or the area estimation model – only the ranks and the number of concordant and discordant pairs will change depending on which estimation model's (area or performance) fidelity is being computed. As explained, the proposed metrics are applicable to design spaces where estimation of one dimension is considered at a given instant. However, all the estimation models used in obtaining an n-dimensional design space can be evaluated separately. It should be noted that a design space from any domain can be considered and is not limited to just performance-area design spaces.

6 EXPERIMENTAL SETUP

To evaluate the proposed fidelity metrics, we chose two estimation models: a single processor performance estimation model, and a multiprocessor performance estimation model. The single processor model estimates the runtime of an application being executed on an in-order processor, with separate L1 instruction and data caches, and separate instruction and data memories. The multiprocessor estimation model estimates the runtime of an application that is partitioned into separate tasks, which are assigned to the processors in the multiprocessor system. The runtime of the multiprocessor system is estimated using the runtime of individual processors in the system. Further details about the estimation models can be found in [15]. It should be noted that the details of estimation models are not required to calculate the fidelity metrics – only the estimated values are needed. We refer to these models as SP (Single Processor) and MP (MultiProcessor) models.

We evaluated the proposed metrics with 2-dimensional design spaces, creating performance-area design spaces. In these design spaces, the performance values are estimated using the SP and MP models, while no estimation is used for the area measurement which means that the actual area values are used for both the actual and estimated design points. The fidelity of the SP and MP models is calculated using the created design spaces (which are used for obtaining the Pareto front). As explained in Section 5.3, all the results shown in the next section use $W = \frac{1}{1+1000d}$ as the weight function for the calculation of WFM_ρ and WFM_τ.

7 RESULTS & ANALYSIS

Firstly, we present the results for the SP model. We used the SP model to calculate the estimated runtime of different applications (JPEG Encoder and JPEG Decoder) on 16 different processors, where several different implementations are available for each processor. Thus, an application running on two different implementations of a processor will result in different runtimes. Actual runtime values were obtained through cycle-accurate simulations, and the area values were obtained as was explained by the authors in [15]. Once these values were available, the four fidelity metrics were computed. Table 7.1 shows the computed fidelity metrics of SP model for all 16 processors.

In Table 7.1, the second and third column show the average and maximum absolute error, computed by comparing the estimated runtimes (calculated through SP model) with the actual runtimes for all the available implementations of a processor. For example, the SP model encountered an average absolute error of 0.15%, with a maximum absolute error of 0.50% across all the available implementations of P2 (row 3). The fidelity metrics were calculated as explained in Section 5, shown in columns 4 – 7. For all the 16 processors, all the fidelity metrics (FM_ρ, WFM_ρ, FM_τ and WFM_τ) are above 0.80. It is interesting to note that P6, which encountered a maximum absolute error of only 3.17%, had the lowest fidelity ($FM_\tau = 0.828$) amongst all the processors. On the other hand, P15 encountered the worst maximum absolute error of 17.07% amongst all the processors, and still had a better fidelity than P6. Thus, it can be concluded that low absolute errors does not necessarily mean the best fidelity. This shows the significance of measuring the fidelity of estimation models. Another interesting result is the value of 1 for all the fidelity metrics for P2 and P8. Thus, for P2 and P8, the exploration algorithms will find the same global minima or maxima. For some processors, for example P14, the values of weighted metrics are lower than the non-weighted ones, suggesting that wrongly-ordered points are closer to the Pareto front than the correctly-ordered points. For other processors, for example P6, $WFM_\tau = 0.922$ compared to $FM_\tau = 0.828$, suggesting that more correctly-ordered points are closer to the Pareto front, thus SP model will allow an exploration algorithm to make better choices in the vicinity of the Pareto front. Using the proposed fidelity metrics, designers can easily observe the usefulness of their estimation models in terms of how well the ex-

Processor	Avg.(%)	Max.(%)	FM$_\rho$	WFM$_\rho$	FM$_\tau$	WFM$_\tau$
P1	1.19	2.94	0.979	0.991	0.896	0.928
P2	0.15	0.50	1.000	1.000	1.000	1.000
P3	1.38	15.65	0.988	0.991	0.906	0.912
P4	2.45	2.56	0.995	0.993	0.874	0.882
P5	0.37	1.74	0.999	0.999	0.990	0.989
P6	0.72	3.17	0.954	0.990	0.828	0.922
P7	1.40	3.11	0.985	0.994	0.913	0.946
P8	0.16	0.96	1.000	1.000	1.000	1.000
P9	1.32	4.36	0.989	0.978	0.882	0.903
P10	1.29	8.66	0.979	0.992	0.926	0.945
P11	0.36	1.41	0.991	0.998	0.928	0.974
P12	6.65	13.92	0.983	0.991	0.901	0.920
P13	7.02	15.37	0.970	0.993	0.867	0.909
P14	7.41	16.10	0.984	0.955	0.893	0.893
P15	8.21	17.07	0.978	0.974	0.886	0.884
P16	1.37	4.71	0.994	0.995	0.941	0.942

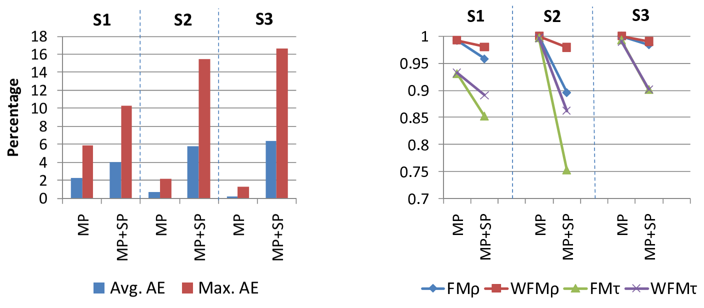
Table 7.1: Computed fidelity metrics for SP model

ploration algorithms will be guided by those estimation models. Furthermore, once a model with good fidelity has been developed, there is no need to improve its absolute accuracy as fidelity is more important than absolute accuracy for proper guidance of the design space exploration algorithms. It can also be seen that the values of FM_τ and WFM_τ are usually lower than the values of FM_ρ and WFM_ρ respectively, as explained in Section 5.2.

In the second set of experiments, we used MP model to estimate the runtime of 3 multiprocessor systems running JPEG encoder and decoder applications. As mentioned earlier, the MP model estimates the runtime of the multiprocessor system by utilizing the runtimes of the individual processors [15]. The runtimes of individual processors can be obtained either through cycle-accurate simulation or through the SP model. Thus, we term the estimation technique that uses the MP model and cycle-accurate runtimes of individual processors as MP model, and the other technique that uses MP model and estimated runtimes of individual processors through SP model as MP+SP model. Obviously, MP+SP model will be faster than MP model, but less accurate. The results for MP and MP+SP model are shown in the table depicted in Figure 7.1, with the same major column titles as Table 7.1. The two sub-columns in each major column show the computed values for MP and MP+SP models respectively. The values show that MP model is very good in predicting the runtime of an application as all the fidelity metrics are above 0.93.

A graphical comparison of the absolute error and fidelity metrics of the two models (MP and MP+SP) is illustrated in Figure 7.1. In Figure 7.1, the results for the 3 multiprocessor systems are separated by the vertical dotted lines and marked as S1, S2 and S3. In Figure 7.1(a), the blue bars show the average absolute error while the red bars show the maximum absolute error. For all the 3 systems, the absolute error has increased, with 17% worst absolute error in the MP+SP model. Examining the fidelity

System	Avg. Error (%)		Max. Error (%)		$\mathbf{FM}_\rho$		$\mathbf{WFM}_\rho$		$\mathbf{FM}_\tau$		$\mathbf{WFM}_\tau$	
	MP	MP+SP	MP	MP+SP	MP	MP+SP	MP	MP+SP	MP	MP+SP	MP	MP+SP
S1	2.28	4.03	5.91	10.28	0.992	0.958	0.996	0.98	0.930	0.852	0.945	0.891
S2	0.69	5.77	2.16	15.44	1.000	0.896	1.000	0.979	0.996	0.753	0.997	0.862
S3	0.21	6.4	1.29	16.61	1.000	0.984	1.000	0.99	0.992	0.901	0.986	0.902



(a)

(b)

Figure 7.1: Comparison of (a) Absolute error (b) Fidelity metrics of MP and MP+SP models

of MP+SP model compared to MP model, shown in Figure 7.1(b), illustrates that the lowest fidelity metrics of MP+SP model for S1, S2 and S3 are 0.852, 0.753 and 0.901 respectively. This suggests that MP+SP model has been affected significantly by the use of SP model for runtime estimation of individual processors. This is more acute for S2 where the fidelity metrics have dropped from 0.99 to 0.75. However, all the weighted metrics are above 0.85 suggesting that estimated values from MP+SP model are better ordered closer to the Pareto front. Selection of an estimation model based on a threshold value is left to designer. Thus, one may choose 0.85 as the threshold value, opting not to use MP+SP model for S2 or to further improve the model. This illustrates another significance of the proposed metrics where designers can evaluate different estimation models quickly in addition to the measurement of absolute error, and choose the best one to be used later in their design space exploration frameworks.

8 CONCLUSION

In this paper, it is shown that measuring fidelity in addition to the measurement of absolute error of an estimation model is important, especially from the perspective of design space exploration algorithms. Four fidelity metrics were proposed, based on Spearman's rank correlation coefficient and Kendall's tau correlation coefficient, to measure the efficacy of estimation models in terms of fidelity. Once a model with good fidelity has been found, designers do not need to work on improving its absolute accuracy. In addition, different estimation models can be evaluated quickly with the proposed fidelity metrics, to choose the best model for use in the design space exploration frameworks. Finally, we showed the calculation of the fidelity metrics on a single processor and a multiprocessor estimation model, and included an insight of the results.

Bibliography

[1] M. Gries, "Methods for evaluating and covering the design space during early design development," *Integr. VLSI J.*, vol. 38, no. 2, pp. 131–183, 2004.

[2] P.-K. Huang, M. Hashemi, and S. Ghiasi, "System-level performance estimation for application-specific mpsoc interconnect synthesis," in *SASP '08: Proceedings of the 2008 Symposium on Application Specific Processors*, (Washington, DC, USA), pp. 95–100, IEEE Computer Society, 2008.

[3] J. T. Russell and M. F. Jacome, "Architecture-level performance evaluation of component-based embedded systems," in *DAC '03: Proceedings of the 40th annual Design Automation Conference*, (New York, NY, USA), pp. 396–401, ACM, 2003.

[4] T.-C. Chen, S.-R. Pan, and Y.-W. Chang, "Performance optimization by wire and buffer sizing under the transmission line model," in *ICCD '01: Proceedings of the International Conference on Computer Design: VLSI in Computers & Processors*, (Washington, DC, USA), p. 192, IEEE Computer Society, 2001.

[5] M. Cho, H. Shin, and D. Z. Pan, "Fast substrate noise-aware floorplanning with preference directed graph for mixed-signal socs," in *ASP-DAC '06: Proceedings of the 2006 Asia and South Pacific Design Automation Conference*, (Piscataway, NJ, USA), pp. 765–770, IEEE Press, 2006.

[6] C. Spearman, "The proof and measurement of association between two things," *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904.

[7] M. G. Kendall, *Rank Correlation Methods*. London: Griffin, 4th ed., 1970.

[8] P. Godfrey, R. Shipley, and J. Gryz, "Algorithms and analyses for maximal vector computation," *The VLDB Journal*, vol. 16, no. 1, pp. 5–28, 2007.

[9] F. De Faria, M. Strum, and W. J. Chau, "A system-level performance evaluation methodology for netwrok processors based on network calculus analytical modeling," in *ISVLSI '07: Proceedings of the IEEE Computer Society Annual Symposium on VLSI*, (Washington, DC, USA), pp. 265–272, IEEE Computer Society, 2007.

[10] S. Eyerman, L. Eeckhout, and K. De Bosschere, “Efficient design space exploration of high performance embedded out-of-order processors,” in *DATE ’06: Proceedings of the conference on Design, automation and test in Europe*, (3001 Leuven, Belgium, Belgium), pp. 351–356, European Design and Automation Association, 2006.

[11] A. Joshi, J. Yi, J. Bell, R.H., L. Eeckhout, L. John, and D. Lilja, “Evaluating the efficacy of statistical simulation for design space exploration,” in *Performance Analysis of Systems and Software, 2006 IEEE International Symposium on*, pp. 70–79, March 2006.

[12] E. Yilmaz, J. A. Aslam, and S. Robertson, “A new rank correlation coefficient for information retrieval,” in *SIGIR ’08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, (New York, NY, USA), pp. 587–594, ACM, 2008.

[13] B. Carterette, “On rank correlation and the distance between rankings,” in *SIGIR ’09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, (New York, NY, USA), pp. 436–443, ACM, 2009.

[14] G. A. Fredricks and R. B. Nelsen, “On the relationship between spearman’s rho and kendall’s tau for pairs of continuous random variables,” *Journal of Statistical Planning and Inference*, vol. 137, no. 7, pp. 2143 – 2150, 2007.

[15] H. Javaid, A. Janapsatya, M. S. Haque, and S. Parameswaran, “Rapid runtime estimation methods for pipelined mpsocs,” in *DATE ’10: Proceedings of the conference on Design, automation and test in Europe*, 2010.