TeamQuest

Reevaluating "Evaluating Scalability Parameters: A Fitting End"

The article, "Evaluating Scalability Parameters:A Fitting End" by Dr. Neil J. Gunther, contains a step by step method to estimate the parameters, σ and λ, of the Super-serial Scalability Law (SSL) using Microsoft Excel. Further investigation of the scheme shows some deviations from expected behavior. In this white paper, the root cause of the deviations from the expected results are explained and an improved scheme is proposed for getting more accurate estimates.

About the Author

Dr. Jayanta Choudhury has a Ph.D in applied mathematics and an MS in computer engineering from the University of Louisiana at Lafayette. He is currently a software engineer at TeamQuest in the area of capacity planning and performance prediction. His research interests include capacity modeling, high performance computing, algorithm development, data analysis, numerical analysis, and numerical solution of PDEs, ODEs, etc.

Original Arrangement

A method to estimate the parameters of the Super-serial Scalability Law was published in the article "Evaluating Scalability Parameters:A Fitting End" by Dr.Neil J. Gunther [1]. For the sake of self containment some of the expressions are reproduced in this section. Eq.(4) in [1] states:

$$\frac{N}{C} = 1 + \sigma\{(N-1) + \lambda N(N-1)\}. \tag{1}$$

Denote the random variables, *X* and *Y* as described in [1]:

$$Y = \frac{N}{C} - 1. \tag{2a}$$

$$X = (N-1). \tag{2b}$$

Substitution of *(N/C)* − 1 and *(N* − 1) in Eq.(1) by the expressions in Eq.(2) gives:

$$Y = \sigma\lambda X^2 + (\sigma\lambda + \sigma)X. \tag{3}$$

The scheme, in [1], proposed to substitute $\sigma\lambda$ by *a* and $\sigma\lambda + \sigma$ by *b* in Eq.(3) to get

$$Y = aX^2 + bX. \tag{4}$$

The above Eq.(4) is the exact statement of Eq.(6) in [1]. The above quadratic polynomial in Eq.(4) can be used directly in the regression-based "Trendline" tool of Microsoft Excel and is stated to be the reason for converting to this form [1].

Eq. (4) was based on Eq.(7) and Eq.(8), in section 3.3 of [1], which suggested the following symbolic substitution

$$\lambda = \frac{a}{b-a} \tag{5}$$

$$\sigma = \frac{a}{\lambda} = \frac{a}{\frac{a}{b-a}} = \frac{a}{1} \times \frac{b-a}{a} = b - a. \tag{6}$$

The suggested relations in (5) and (6) imply that

$$\lambda\sigma = \frac{a}{b-a} \times (b-a) = a. \tag{7}$$

In the same way, the relations (5) and (6) imply that

$$\lambda\sigma + \sigma = \left[\frac{a}{b-a} \times (b-a)\right] + (b-a) = a + (b-a) = b. \tag{8}$$

Once the equation is stated in a suitable form, the scheme in [1] suggested using performance data to use the regression-based "Trendline" tool of Microsoft Excel for a quadratic polynomial to estimate the parameters σ and λ. In the next section we will use two sets of performance data to show the deviation in the results.

Application of Least Square Regression

June 1995 Data

The source SPEC data in Table 1 in [1] was referred to as the June 1995 release but later it was found by Jon Hill of TeamQuest Corporation, that the data was from a December 1994 release. There is no harm in applying the method to both data sets. The June 1995 data [3], along with the processed values for applying the method of [1], is reported in Table 1. Notice specifically the blue values in Table 1. The maximum of N, N_{max}, should be between $N = 36$ and $N = 108$ as indicated by the blue colored data in Table 1.

Table 1: SPEC SDM Benchmark on a 16-way SUN SC 2000 [3]

User Load N	Throughput Scripts/Hour $T(N)$	$X_j = N_j - 1$	$Y_j = \frac{N_j}{C_j} - 1$
1	105.2	$X_0 = 0$	$Y_0 = 0.00$
9	833.0	$X_1 = 8$	$Y_1 = 0.136615$
18	1507.1	$X_2 = 17$	$Y_2 = 0.256453$
27	2074.0	$X_3 = 26$	$Y_3 = 0.369527$
36	2453.3	$X_4 = 35$	$Y_4 = 0.543717$
72	2672.2	$X_5 = 71$	$Y_5 = 1.834518$
108	2582.3	$X_6 = 107$	$Y_6 = 3.399799$
144	2508.5	$X_7 = 144$	$Y_7 = 5.038987$

Table 2: SSL parameters from [3]

SSL Parameter	Parameter Value
σ	0.014
λ	0.014286
$N_{\max} = \sqrt{\frac{1-\sigma}{\sigma\lambda}}$	70.21396

The steps to estimate the parameters σ and λ, described in section 4 of [1], are applied to the data of Table 1 and the results are reported in Table 2, above. The value of $N_{max} = 70$ predicted by the estimated σ and λ is included in Table 2. The following Figure 1 compares the model function specified by the estimates of σ and λ of Table 2 and the actual measurements of Table 1.

The red colored graph in Figure 1 is the model function of Super-serial Scalability Law for the parameters of Table 2. It shows that the maximum performance, $C(N_{max}) \approx 23.87$, for $N = 70.21396$ as predicted by Super-serial Scalability Law, is less than the measured performance for $N = 72$ (C(72) ≈ 25.4) and $N = 108$ ($C(108) \approx 24.6$). This is a subtle but qualitatively significant deviation from the expected behavior suggested by the measured performance data.

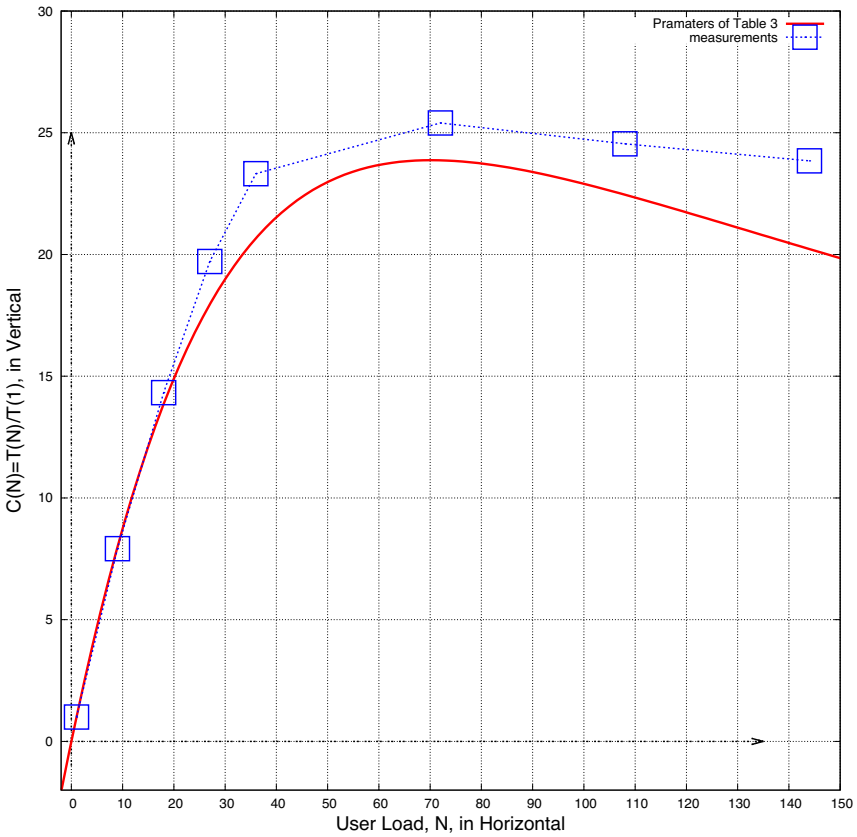

Figure 1: Comparing the Prediction Graphs with Measurements

By observing the data in Table 1, one can rationally expect that $C(N_{max}) \geq C(72)$ and $36 < N_{max} < 108$. In Table 2 the value for N_{max} is 70, which is between 36 and 108, but in Figure 1 the predicted maximum performance, $C(70)$, is less than $C(72)$ and $C(108)$. This starkly defies expectations.

Oct 1994 Data

An Internet search by Jon Hill of TeamQuest Corporation indicated that the data for the same system, reported on Dec 1994 [4], matches with the data of Table 1 in [1]. The data is reproduced in Table 3, below, to indicate an important observation. Notice again the rows colored blue in Table 3. The data in those blue rows indicate that the N_{max} is between $N = 36$ and $N = 108$. In Table 5 in [1], $N_{max} = 111$. The data of Table 5 in [1] is reproduced in Table 4, below. The measured performance for $N = 108$ is less than the measured performance for $N = 72$. Thus, the predicted value of $N_{max} = 111$ (in red) in Table 4, computed from the σ and λ in Table 5 in [1], does not comply with the data in Table 3.

Table 3: SPEC SDM Benchmark on a 16-way SUN SC 2000 [4]

User Load N	Throughput Scripts/Hour $T(N)$
1	64.9
18	995.9
36	1652.4
72	1853.2
108	1828.9
144	1775
216	1702.2

Table 4: SSL parameters from [4]

SSL Parameter	Parameter Value
σ	0.0169
λ	0.0047
$N_{\max} = \sqrt{\frac{1-\sigma}{\sigma\lambda}}$	111

The data in Table 3 suggests that the performance, $C(111) \approx 28.99$, at $N = 111$ should be less than the performance, $C(108)$ at $N = 108$. Hence the prediction that maximum performance will be at $N = 111$ is qualitatively a significant deviation from expected behavior. This is evidence of deviation in the method to estimate the parameters σ and λ in [1].

Analysis with Regression Theory

According to Eq.(9.100) in page 604 of [2], denote the random variables, x_1 and x_2, as following:

$$x_1 = X. \tag{9a}$$

$$x_2 = X^2. \tag{9b}$$

Substitute Eq.(9) into Eq.(4). Then according to Eq.(9.86) in page 589 of [2], $S(a,b)$ needs to be minimized:

$$S(a,b) = \sum_{j=1}^{J} (Y_j - ax_{2j} - bx_{1j})^2. \tag{10}$$

where, $J = 8$ when using data from Table 1 and $J = 7$ when using data from Table 3. To do the regression, according to page 589 of [2], set partial derivative of $S(a,b)$ with respect to a to zero and set partial derivative of $S(a,b)$ with respect to b to zero and solve the system of equations as shown below:

$$\frac{\partial S}{\partial a} = 2\sum_{j=1}^{J} (Y_j - ax_{2j} - bx_{1j}) \left[\frac{\partial}{\partial a}(Y_j - ax_{2j} - bx_{1j})\right] = 0. \tag{11a}$$

$$\frac{\partial S}{\partial b} = 2\sum_{j=1}^{J} (Y_j - ax_{2j} - bx_{1j}) \left[\frac{\partial}{\partial b}(Y_j - ax_{2j} - bx_{1j})\right] = 0. \tag{11b}$$

Notice, the red colored factors in the expressions in Eq.(11). These are the source of the trouble. How these factors cause trouble is explained next.

Remember that from Eq.(7) and Eq.(8), a and b are functions of each other. From Eq.(7) and Eq.(8):

$$\begin{aligned} a = \sigma\lambda \text{ and } b = \sigma\lambda + \sigma. \\ \Rightarrow b = a + \sigma. \\ \Rightarrow \frac{\partial b}{\partial a} \neq 0 \text{ and } \frac{\partial a}{\partial b} \neq 0. \end{aligned} \tag{12}$$

Now expand the red colored factors in Eq.(11a):

$$\frac{\partial}{\partial a}(Y_j - ax_{2j} - bx_{1j}) = \frac{\partial}{\partial a}Y_j - \frac{\partial}{\partial a}ax_{2j} - \frac{\partial}{\partial a}bx_{1j}.$$

$$\Rightarrow \frac{\partial}{\partial a}(Y_j - ax_{2j} - bx_{1j}) = -x_{2j}\frac{\partial}{\partial a}a - x_{1j}\frac{\partial}{\partial a}b.$$

From Eq.(12) the above expression implies that:

$$\frac{\partial}{\partial a}(Y_j - ax_{2j} - bx_{1j}) = -x_{2j} - \frac{\partial b}{\partial a}x_{1j}. \tag{13}$$

Similarly from Eq.(12) the red colored factor of Eq.(11b) implies that:

$$\frac{\partial}{\partial b}(Y_j - ax_{2j} - bx_{1j}) = -\frac{\partial a}{\partial b}x_{2j} - x_{1j}. \tag{14}$$

So, a simplified form of Eq.(11) is:

$$\frac{\partial S}{\partial a} = 2\sum_{j=1}^{7}(Y_j - ax_{2j} - bx_{1j})\left[-x_{2j} - \frac{\partial b}{\partial a}x_{1j}\right] = 0. \tag{15a}$$

$$\frac{\partial S}{\partial b} = 2\sum_{j=1}^{7}(Y_j - ax_{2j} - bx_{1j})\left[-\frac{\partial a}{\partial b}x_{2j} - x_{1j}\right] = 0. \tag{15b}$$

According to page 589 of [2], it is required that

$$\frac{\partial b}{\partial a} = 0, \frac{\partial a}{\partial b} = 0. \tag{16}$$

If the required conditions of Eq.(16) were satisfied then the expression in Eq.(15) would have looked like the following:

$$\frac{\partial S}{\partial a} = -2\sum_{j=1}^{J}(Y_j - ax_{2j} - bx_{1j})(x_{2j}). \tag{17a}$$

$$\frac{\partial S}{\partial b} = -2\sum_{j=1}^{J}(Y_j - ax_{2j} - bx_{1j})(x_{1j}). \tag{17b}$$

The extra terms in the red colored factor in Eq.(15) gives the correct form of the equations for estimating *a* and *b* of Eq.(4) that are the counter part of the final solvable equations, Eq.(17), based on the theory in Page 589 of [2] for regression or least-square-error solution when the conditions of Eq.(16) do not hold because of Eq.(5) and Eq.(6). Further simplification of the equations of Eq.(15) deduces:

$$\sum_{j=1}^{7}(Y_j - ax_{2j} - bx_{1j})\left[x_{2j} + \frac{\partial b}{\partial a}x_{1j}\right] = 0. \tag{18a}$$

$$\sum_{j=1}^{7}(Y_j - ax_{2j} - bx_{1j})\left[\frac{\partial a}{\partial b}x_{2j} + x_{1j}\right] = 0. \tag{18b}$$

Instead, the polynomial option in the regression-based “Trendline” tool of Microsoft Excel solves the following equations that are simplified from Eq.(17):

$$\sum_{j=1}^{7}[Y_j(-x_{2j})] - a\sum_{j=1}^{7}[x_{2j}(-x_{2j})] - b\sum_{j=1}^{7}[x_{1j}(-x_{2j})] \tag{19a}$$

$$\sum_{j=1}^{7}[Y_j(-x_{1j})] - a\sum_{j=1}^{7}[x_{2j}(-x_{1j})] - b\sum_{j=1}^{7}[x_{1j}(-x_{1j})] = \text{(illegible)} \tag{19b}$$

Microsoft Excel assumes that the conditions of Eq.(16) hold, which is not true. So, the regression-based “Trendline” tool of Microsoft Excel cannot be trusted to give the correct estimates of *a* and *b* of Eq.(4). Thus the estimates of σ and λ cannot be trusted as well. The error is introduced when Eq.(5) and Eq.(6) are used to estimate σ and λ from the values of *a* and *b*.

Solution

A method using the regression-based “Trendline” tool of Microsoft Excel that solves the problem from the previous section is proposed in this section.

Theory

The solution of the problem, posed by the presence of the extra term in the red colored factors of Eq.(18), has been published in the proceedings of 2011 IEEE INTERNATIONAL CONFERENCE on ELECTRO/INFORMATION TECHNOLOGY as a peer-reviewed paper [5]. For a detailed reading of the theory behind the proposed method, one can refer to [5]. Table 7 contains the parameters derived using the method described in [5]. Expressions to compute the values in Table 7 are provided next.

Using Microsoft Excel

The flawed regression steps, proposed in [1], can be avoided by implementing the linearization transformations proposed in [5].

STEP:1

For a given set of performance data $\{(X_i, Y_i)|i = 0, 1, 2, ..., M\}$ denote the random variables $\hat{X}$ and $\hat{Y}$ such that

$$\hat{X}_i = Y_1(X_i + X_i^2) - Y_i(X_1 + X_1^2). \tag{20}$$

$$\hat{Y}_i = (Y_i X_1 - Y_1 X_i). \tag{21}$$

The result of this processing on the data sets [4] and [3] are reported in Table 5, below.

Table 5: Processed data for applying linearized regression to estimate λ

data set [4]				data set [3]			
User Load N	Throughput $T(N)$	$\hat{X}$	$\hat{Y}$	User Load N	Throughput $T(N)$	$\hat{X}$	$\hat{Y}$
1	64.9			1	105.2		
18	995.9	0	0	9	833.0	0	0
36	1652.4	91.325	0.982	18	1507.1	23.340	-0.271
72	1853.2	418.852	13.581	27	2074.0	69.298	-0.596
108	1828.9	1132.561	29.640	36	2453.3	132.987	-0.432
144	1775.0	2257.480	47.767	72	2672.2	566.289	4.977
216	1702.2	5820.503	85.806	108	2582.3	1333.933	12.581
				144	2508.5	2450.362	20.776

STEP:2

Get the estimate of λ using the linear regression based "Trendline" tool of Microsoft Excel for the model function:

$$\hat{Y} = \lambda \hat{X}. \tag{22}$$

The procedure is slightly different than the procedure in [1]. In this case, choose "line" instead of the "polynomial" option in the dialog after selecting the "Trendline" tool of Microsoft Excel. Use the rest of the procedure explained in [1] about using the regression-based "Trendline" tool of Microsoft Excel.

STEP:3

Once λ is estimated (estimated λ is reported in Table 7), denote a new random variable, $\tilde{X}$, as

$$\tilde{X}_i = X_i + \lambda(X_i^2 + X_i). \tag{23}$$

The processed data is reported in Table 6.

Table 6: Processed data for applying linearized regression to estimate σ

data set [4], $\lambda = 0.015989$				data set [3], $\lambda = 0.0087$			
User Load N	Throughput $T(N)$	$\tilde{X}$	Y	User Load N	Throughput $T(N)$	$\tilde{X}$	Y
1	64.9	0	0	1	105.2	0	0
18	995.9	21.893	0.173	9	833.0	8.626	0.1367
36	1652.4	55.146	0.414	18	1507.1	19.6622	0.257
72	1853.2	152.736	1.522	27	2074.0	32.107	0.370
108	1828.9	291.769	2.833	36	2453.3	45.962	0.544
144	1775.0	472.246	4.265	72	2672.2	115.474	1.835
216	1702.2	957.529	7.236	108	2582.3	207.537	3.400
				144	2508.5	322.150	5.039

STEP:4

Again, get the estimate of σ using the "line" option of the "Trendline" tool of Microsoft Excel for the model function:

$$Y = \sigma \tilde{X}. \tag{24}$$

The estimated value of σ is reported in Table 7, below.

Table 7: The estimates of σ and λ from [5]

SSL Parameter	Parameter Values [3]	Parameter Values [4]
σ	0.0158	0.00801
λ	0.0087	0.015989
$N_{\max} = \sqrt{\frac{1-\sigma}{\sigma\lambda}}$	84.6162	88.007

In Figure 2 and Figure 3 the parameters σ and λ are taken from Table 7. Notice that N_{max} in Table 7 complies with the data in Table 1 and Table 3.

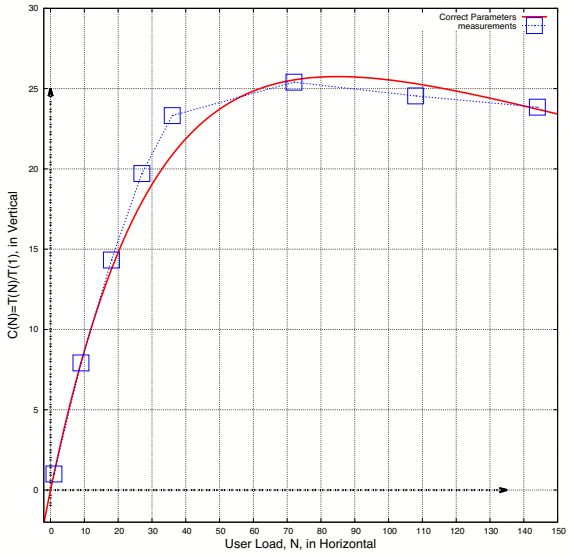

Figure 2: Comparing the Prediction Graphs with Benchmark Data of [3]

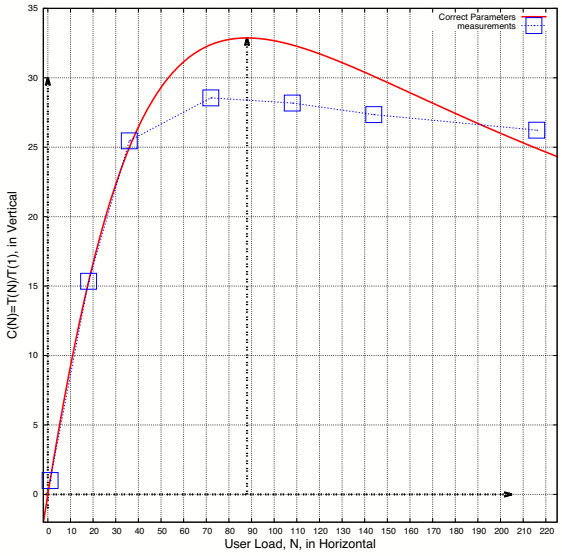

Figure 3: Comparing the Prediction Graphs with Benchmark Data of [4]

Figure 2 contains a graphical comparison of the Super-serial Scalability model function and benchmark data of Table 1. The data in Table 1 indicates that $C(108) > C(36)$ and $C(72)$ is larger than either of $C(36)$ and $C(108)$. The proposed new method predicts that $N_{max} \approx 84$. Note that $72 < N_{max} = 84 < 108$ complies with the expectation based on the predicted performance, $C(84)$, which is larger than $C(36)$,$C(72)$ and $C(108)$.

Figure 3 contains a graphical comparison of the Super-serial Scalability model function and benchmark data of Table 3. The vertical line at $N = 88$ in Figure 3 indicates the point at which maximum scalability is attained. Again $72 < N_{max} = 88 < 108$ and $C(N_{max}) \geq C(72)$.

A very important aspect of the proposed new method is that for both sets of data the predicted N_{max} is between the measured points at $N = 72$ and $N = 108$. The blue colored rows in Table 1 and Table

3 suggest such behavior. The method in [1] predicted N_{max} outside that interval in each case. So, for both sets of example data, the improved method predicts the maximum user load within the expected range suggested by the example data, which is better than the method stated in [1].

Conclusion

Comparison of some predicted values in the results with measured data and subsequent analysis have shown that the method in [1] is not quite accurate and rather qualitatively awkward. A theoretical analysis was used to identify the problem and a process for estimating the parameters was described. A correct method was explained and the results compared with data and graphs. This white paper shows that the proposed new method is much better for predicting the maximum user load and maximum throughput.

References

[1] N. J. Gunther. (2001, Oct). “Evaluating Scalability Parameters: A Fitting End.” [Online] Available:http://www.teamquest.com/pdfs/whitepaper/fitting.pdf

[2] A. O. Allen, “Probability, Statistics and Queueing Theory with Computer Science Ap- plications,” 2nd Edition, San Diego, CA:Academic Press, 1978, ISBN 0120510510.

[3] “SPEC SDM Results: June ’95,” Available:http://www.spec.org/osg/sdm91/ results/res9506/

[4] “SPEC SDM Results: Oct ’94,” Available:http://www.spec.org/osg/sdm91/results/ res9412/cp127.ps

[5] J. Choudhury. “Novel Regression Approach to Estimate the Parameters of ‘Universal Scalability Law’,” The proceedings of IEEE eit2011-2011 IEEE INTERNATIONAL CON- FERENCE on ELECTRO/INFORMATION TECHNOLOGY, Pages: 1-5, ISSN: 2154- 0357, Print ISBN: 978-1-61284-465-7, May 15 - 17, 2011, Mankato, MN.

TeamQuest Corporation

www.teamquest.com

Follow the TeamQuest Community at:

Worldwide

One TeamQuest Way
Clear Lake, IA 50428
USA
+1 641.357.2700
+1 800.551.8326
info@teamquest.com

Europe, Middle East and Africa

Box 1125
405 23 Gothenburg
Sweden
+46 (0)31 80 95 00
United Kingdom
+44 (0)1865 338031
Germany
+49 (0)69 6 77 33 466
emea@teamquest.com