What Do You 'Mean'?
Revisiting Statistics for Web Response Time Measurements

David M. Ciemiewicz
(ciemo@pacbell.net)

Using empirical data from an Internet/WAN distributed Web response time measurement system, this paper explores the relative applicability and usefulness of the geometric mean and geometric standard deviation, and introduces the lognormal distribution for quantifying the response time measurements. These statistics are particularly useful in the areas of Web content performance comparison and SLA monitoring of service providers such as Content Providers, CDNs, ASPs, MSPs, and Web Hosting companies. Some surprisingly counter-intuitive results are identified and discussed. Warning: improper use of traditional averages and standard deviations can financially hurt you, if not simply mislead you.

Overview

There is a push in the Internet industry and within corporate enterprise environments to measure end-user response time experience to ensure satisfaction of services delivered as part of a proactive, Service Level Management (SLM) strategy [STUR00]. These measurements are made using synthetic transaction probes that are placed, not at the data center, but rather at locations distributed around the network, so as to better guage the end-user's true experience.

These measurements may be made to validate conformance with contractual Service Level Agreements (SLAs). Sometimes there are financial penalties for non-compliance. These SLAs are often based on the arithmetic mean and arithmetic standard deviation of periodic samples of service transactions over time.

Yet it is common knowledge among practitioners of statistics that the average or arithmetic mean is often a misleading summary statistic for describing groups of measurements. [HUFF54] This is especially true in the presence of occasional, anomalous network events that cause significant outliers in the response time measurements for the target services.

Sometimes content or service developers attempt to make performance improvements using the arithmetic mean. It would be unfortunate if an assessment of improvement or no improvement were made based on data that was skewed upward by several seconds more than typical due to a temporary slowdown on an Internet backbone that is out of their control. This would certainly be a nuisance.

However, it might be worse than simply "unpleasant" if a single network event sends the measured response time average value into SLA violation territory, even though the responses to the vast majority of your end-users are within SLA tolerances.

To make better decisions, a better understanding of the fundamental statistics is needed to avoid being led astray by the occasional anomalous Internet or WAN network event.

Just How Bad is the Problem?

In the sets of samples used in this paper, **the arithmetic means varied between the 52nd percentile and the 77th percentile of their respective sample sets**. Obviously the relationship between what we think of as "average" and what we intend to present as "representative" or "typical" (or even "repeatable", "consistent", or "comparable") is tenuous at best.

This problem is compounded by the fact that not only are the response time distributions right-tailed but that sporadic Internet network events that affect subsets of the user populations can create long response time outliers that significantly skew the average even more significantly to the right. **For small sets of samples, the arithmetic mean sometimes represented the 95th percentile**.

Many papers and texts suggest that the geometric mean, median, or percentiles should be used as the summary descriptive statistics in preference to the arithmetic mean [BROW01]. Unfortunately, these admonitions assume the reader has a sufficient level of understanding of the underlying statistics to not only appreciate the advice, but also actually act

upon it.
In reality, most of us have forgotten the details of our introductory statistics courses and continue to fall back on the old stand-bys – the average (arithmetic mean) and the arithmetic standard deviation.

This paper focuses on how the geometric mean and geometric standard deviation, as suggested by Overton in [OVER00], really do have the properties of insensitivity to outliers that we desire for accurately comparing sets of samples. It attempts to take the reader a little deeper into the statistics backed with real world example data and even introduces the concept of the lognormal distribution, which is related to the geometric mean and geometric standard deviation.

But first, we need some background on what service response times we measured and how the measurements were made.

Response Time Measurement System Description

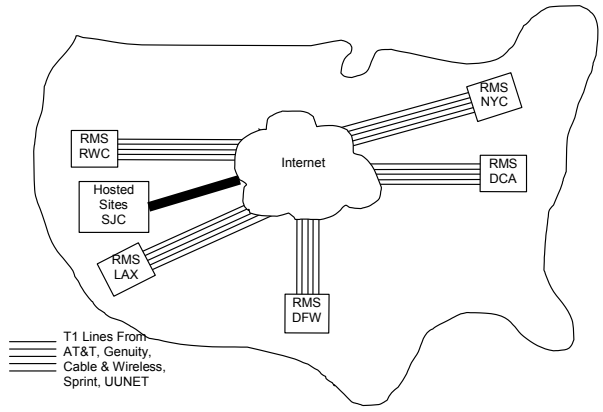
Logictier was a Managed Service Provider (MSP) providing complete outsourcing of all operational aspects of running large Web sites. As part of the service, Logictier offered performance SLAs with a financial penalty payable to the customer for missing the SLA. (Logictier shut down operations in May 2001 due to the big Internet investment retreat of that year).

To measure SLA compliance, and for other operational and research & development purposes, Logictier deployed a distributed Response Time Measurement System (RMS). This system consisted of a set of remotely distributed monitoring stations that periodically issued sets of synthetic HTTP requests across multiple backbone providers to measure the time to download a complete HTML Web page.

Each synthetic transaction included DNS hostname resolution and requests for all attendant objects including assets from CDNs and ad services. Four simultaneous HTTP 1.1 requests were issued just as most Web browsers will do.

Figure 1 shows a diagram of the RMS monitoring station network as of April 2001. By this time, Logictier had five RMS monitoring stations around the United States with a total of 24 T1 network lines from 5 major backbone providers connecting these stations to the Internet. (The location designations in Figure 1 are based on the closest airport designation or a similar abbreviation for the nearest city.)

Figure 1 - RMS Network Overview



Now we're ready to begin interpreting the data

How Measurement Outliers Skew the Mean

When attempting to compare measurements of web site response time performance, we want to be certain that our summaries of the measurements consider, but do not overly weight, the occasional Internet event that is outside of our control. In other words, we want to reflect the experience of the Internet at large without being significantly swayed due to a brief (1-15 minute) event along only one of any number of backbone providers on the Internet. Remember that a few samples of any synthetic transaction methodology are only a small representative set of potentially tens of thousands, hundreds of thousands, or even millions of real end-users across of thousands of major Internet routes.

Tables 1 and 2 below illustrate the problem with using the arithmetic mean (or average) as the sole summary statistic.

Table 1 contains fourteen measurements of a particular Web site's page download times. Most measurements fall in a range of 2-8 seconds except for one measurement from location RWC via the ATT backbone.

This single, anomalous measurement is 15 times longer than the next longest measurement. To illustrate the sensitivity of the means to this one outlier, Table 2 represents exactly the same data with one difference: a more typical response time from location RWC via ATT has been substituted.

Note that the arithmetic mean, of all the descriptive statistics calculated, showed the most sensitivity to this anomalous, outlier measurement: a change from 12,685 milliseconds to 5,037 milliseconds! The median showed the least sensitivity and the geometric mean was next best.

Date Time	Monitoring Location	Backbone Provider	Page Total Time (millisecs)
01/23/01 10:31	**RWC**	**ATT**	**109,810**
01/23/01 10:31	RWC	GTE	7,101
01/23/01 10:31	NYC	CW	6,768
01/23/01 10:31	DFW	CW	6,501
01/23/01 10:31	RWC	SPRINT	6,468
01/23/01 10:31	NYC	GTE	6,294
01/23/01 10:31	NYC	ATT	5,842
01/23/01 10:31	DFW	UUNET	5,535
01/23/01 10:31	LAX	GTE	4,696
01/23/01 10:31	RWC	CW	4,358
01/23/01 10:31	DFW	GTE	4,342
01/23/01 10:31	LAX	SPRINT	3,869
01/23/01 10:31	NYC	SPRINT	3,122
01/23/01 10:31	DCA	GTE	2,890
Descriptive Statistics		**Arith Mean**	**12,685**
		Geo Mean	**6,254**
		Median	**5,689**

Table 1 - Page Total Time Sample Set with Outlier

Date Time	Monitoring Location	Backbone Provider	Page Total Time (millisecs)
01/23/01 10:31	RWC	GTE	7,101
01/23/01 10:31	NYC	CW	6,768
01/23/01 10:31	DFW	CW	6,501
01/23/01 10:31	RWC	SPRINT	6,468
01/23/01 10:31	NYC	GTE	6,294
01/23/01 10:31	NYC	ATT	5,842
01/23/01 10:31	DFW	UUNET	5,535
01/23/01 10:31	LAX	GTE	4,696
01/23/01 10:31	RWC	CW	4,358
01/23/01 10:31	DFW	GTE	4,342
01/23/01 10:31	LAX	SPRINT	3,869
01/23/01 10:31	NYC	SPRINT	3,122
01/23/01 10:31	DCA	GTE	2,890
01/23/01 10:31	**RWC**	**ATT**	**2,726**
Descriptive Statistics		**Arith Mean**	**5,037**
		Geo Mean	**4,803**
		Median	**5,116**

Table 2- Page Total Time Sample Set with Outlier Replaced

How a Single Provider "Outage" Affects the Mean

Sometimes an external event, such as network congestion or failure of a line or router, will affect measurements from many routes. While this event affects a significant portion of your user population, it is out of your control and for all intents and purposes impossible to provide an SLA against. If you are measuring for performance tuning, this event significantly skews the mean, hiding the true improvements.

For example, Table 3 shows an event for service provider GTE that impaired traffic for all measured GTE paths to the monitored Web site. In this case, if the arithmetic mean were used as a basis of SLA measurements, a response time of over 51 seconds would most likely provoke an investigation of an SLA violation. Yet if GTE is not the bandwidth provider for the site, it is quite likely that no violation occurred.

This truly emphasizes why relying on the arithmetic mean alone, without looking at any other data or statistics, is very dangerous when it comes to comparing the results of this sample group with another.

Date Time	Monitoring Location	Backbone Provider	Page Total Time (millisecs)
01/22/01 06:16	DCA	GTE	128,671
01/22/01 06:16	RWC	GTE	109,568
01/22/01 06:16	DFW	GTE	169,363
01/22/01 06:16	LAX	GTE	155,550
01/22/01 06:16	NYC	GTE	106,093
01/22/01 06:16	NYC	CW	11,777
01/22/01 06:16	RWC	SPRINT	9,259
01/22/01 06:16	DFW	UUNET	5,732
01/22/01 06:16	NYC	ATT	5,661
01/22/01 06:16	LAX	SPRINT	4,960
01/22/01 06:16	NYC	SPRINT	3,758
01/22/01 06:16	RWC	CW	2,412
01/22/01 06:16	RWC	ATT	2,225
01/22/01 06:16	DFW	CW	2,180
Descriptive Statistics		**Arith Mean**	**51,229**
		Geo Mean	**15,044**
		Median	**7,496**

Table 3 - Single Provider Event Affecting Multiple Paths

(Re)introducing Some Statistical Concepts

Now it is time to broaden our statistical toolbox beyond statistics 101 by gaining a deeper understanding of the geometric mean and by introducing the concepts of the geometric standard deviation, geometric confidence intervals, and the lognormal distribution.

Another Perspective on the Geometric Mean

In its basic interpretation, the geometric mean is the "central" value of a geometric sequence of values. However, there is another interpretation, as we will see, that is quite useful.

First, in a geometric sequence, each value represents a constant multiple of the previous value in the sequence. The geometric mean is the central value that balances the ratio of the central value to the lowest value and the ratio of the highest value to the central value. For example, given two values, a and c, the geometric mean value b is between a and c such that:

$$\frac{b}{a} = \frac{c}{b}$$

Note that for a geometric sequence containing an odd number of values, the geometric mean value of the sequence is the middle or median value.

This property of balancing the ratios between values proves to be quite useful and underlies the proof by Fleming and Wallace in [FLEM86] that the geometric mean is only way to appropriately summarize a set of normalized values such as those measured in benchmarks.

Most texts and papers discussing the geometric mean (GM), including [ALLE90] and [BERK00], simply present the general equation for the geometric mean (GM) as:

$$\mathrm{GM} = \sqrt[n]{x_1 \times x_2 \times \cdots \times x_n}$$

However, this form of the equation hides another useful interpretation -- through a series of transformations, we discover that the geometric mean is also the antilog of the arithmetic mean of the natural log transformed (log-space) values of X or:

$$\mathrm{GM} = \exp(\mathrm{mean}(\ln(X)))$$

Where X represents all of the values x_i and $\ln(X)$ represents all of the natural transformed values of X.

Transformations of Geometric Mean Equations

To demonstrate that the two forms of the geometric mean equation are equivalent, start with the classic representation of the geometric mean equation:

$$\mathrm{GM} = \sqrt[n]{x_1 \times x_2 \times \cdots \times x_n}$$

This same equation may be expressed in product form as:

$$\mathrm{GM} = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}}$$

$$\mathrm{GM} = \mathrm{power}(\mathrm{product}(X), 1/n)$$

Remember that log-space arithmetic has the useful properties of transforming multiplication into summation and exponentiation into multiplication:

$$a \times b = \exp(\ln(a) + \ln(b))$$

$$a^b = \exp(b \times \ln(a))$$

Thus we can transform the product form equations of the geometric mean into summation form like so:

$$\mathrm{GM} = \exp\left(\frac{1}{n} \sum_{i=1}^{n} \ln x_i \right)$$

$$\mathrm{GM} = \exp(\mathrm{sum}(\ln(X)) / \mathrm{count}(X))$$

These summation form equations look very similar to the equations for computing the arithmetic mean. By substituting the relationship $Y = \ln(X)$ and using the equations for calculating the arithmetic mean (AM) or average:

$$\mathrm{AM} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

$$\mathrm{AM} = \mathrm{sum}(Y) / \mathrm{count}(Y)$$

We see that the geometric mean is really just the antilog of the arithmetic mean of the natural log transformed values of X:

$$\mathrm{GM} = \exp(\mathrm{mean}(\ln(X)))$$

This summation form of the geometric mean equation has some useful computational properties. For large numbers of samples, the sum of log-transformed values is much less likely to overflow or underflow the precision and range capabilities of the native floating-point processor than is the product form of the equation.

Also, the equation is easily expressed in programming languages that support operating on lists of numbers. For instance, a performance database built on a relational database would support an SQL expression such as:

```
SELECT
        EXP(AVG(LN(RESPONSE_TIME)))
                AS GEOMEAN
FROM
```

```
        PERFORMANCE_TABLE
WHERE
        RESPONSE_TIME > 0
```

Introducing the Geometric Standard Deviation

Analogous to the geometric mean, we can compute a geometric standard deviation (GSd) as the antilog of the standard deviation of the natural log transformed values of X or:

$$\mathrm{GSd} = \exp(\mathrm{stdev}(\ln(X)))$$

The basic summation forms of the arithmetic standard deviation may be expressed as:

$$\mathrm{ASd} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2}$$

$$\mathrm{ASd} = \mathrm{sqrt}(\mathrm{sum}(X\text{^}2)/\mathrm{count}(X) - (\mathrm{sum}(X)/\mathrm{count}(X))\text{^}2)$$

$$\mathrm{ASd} = \mathrm{sqrt}(\mathrm{mean}(X\text{^}2) - \mathrm{mean}(X)\text{^}2)$$

So the GSd may also be computed as:

$$\mathrm{GSd} = \exp\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n} \ln(x_i)^2 - \left(\frac{1}{n}\sum_{i=1}^{n} \ln(x_i)\right)^2}\right)$$

$$\mathrm{GSd} = \exp(\mathrm{sqrt}(\mathrm{sum}(\ln(X)\text{^}2)/\mathrm{count}(X) - (\mathrm{sum}(\ln(X))/\mathrm{count}(X))\text{^}2))$$

$$\mathrm{GSd} = \exp(\mathrm{sqrt}(\mathrm{mean}(\ln(X)\text{^}2) - \mathrm{mean}(\ln(X))\text{^}2)))$$

As before with the geometric mean, it is easy to calculate the geometric standard deviation in list oriented languages such as SQL:

```
SELECT
        EXP(STDDEV(LN(RESPONSE_TIME)))
                AS GEOSTDEV
FROM
        PERFORMANCE_TABLE
WHERE
        RESPONSE_TIME > 0
```

Introducing The Lognormal Distribution

Given the relationships we've seen for the geometric mean and geometric standard deviation to log transformed values, one might wonder if there are any applicable distributions of log transformed values.

In statistics, there is a well-known right-tailed distribution called the lognormal distribution that is frequently used in describing phenomena.

The lognormal distribution is the normal distribution of the log transformed (log-space) values. Figure 2 illustrates a lognormal versus normal distribution.

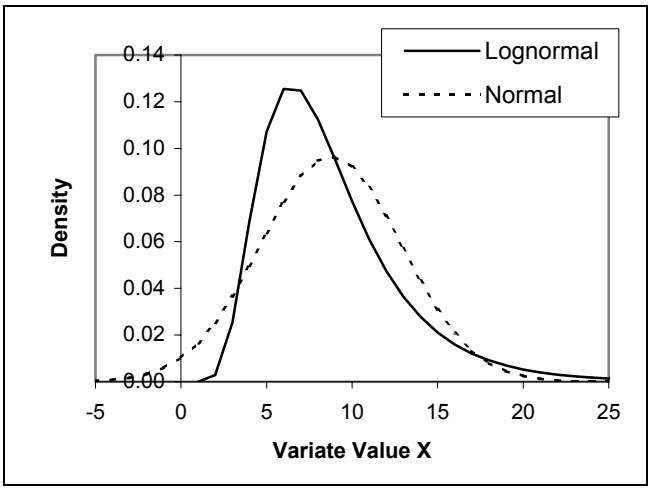


Figure 2 - Lognormal vs. Normal Distribution

Figure 3 shows the same data with log transformed values of X. Here the symmetric properties of a normal distribution in log-space – the lognormal distribution – really stand out.

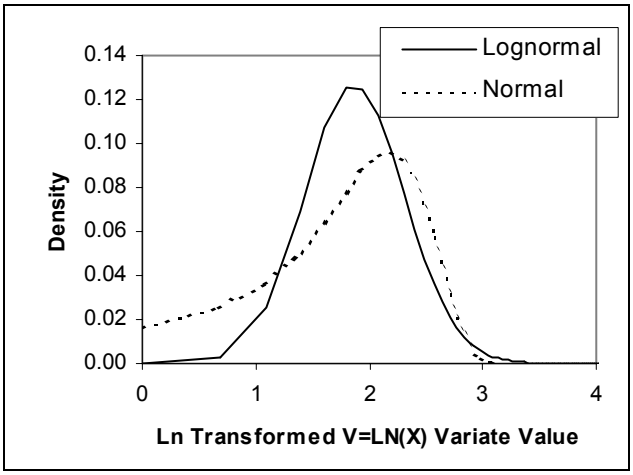


Figure 3 - Lognormal vs. Normal Distribution in Log-space

While the lognormal distribution is a tried and true tool in other disciplines (geneticists use it for analyzing the fluorescence of bacteria with certain genetic markers), it appears to be largely ignored in computing and networking disciplines.

Unfortunately, queuing theory and performance modeling texts such as Kleinrock [KLEI75], Allen [ALLE90], Gross [GROS98], and Gunther [GUNT00] make no mention of the lognormal distribution and rather focus on the markovian (memory-less) exponential distribution and related non-markovian Erlang (gamma) distribution for service time modeling. This emphasis on modeling with these distributions may bias practitioners away from looking at other distributions, such as the lognormal.

In a discussion of the lognormal and Weibull distributions [TART00], Tarter suggests that, "despite the major differences between the actual formulas for these two curves, these substantially different expressions can yield nearly indistinguishable curves." Similarly, Figure 4 illustrates that it is possible to select parameters for the lognormal and Erlang (gamma) distributions whereby the two are very close indeed.

As we will see, the lognormal distribution has additional useful properties for analyzing and modeling Web response times and other network service times.

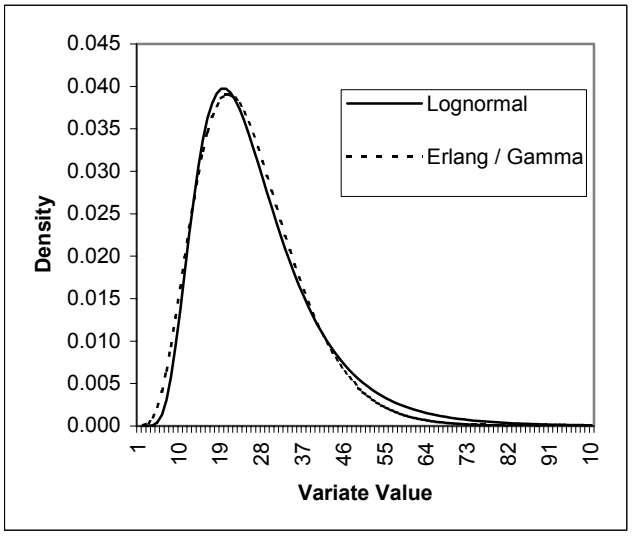


Figure 4 - Lognormal vs. Erlang / Gamma Distribution

Geometric Statistics and the Lognormal

Although not widely mentioned in the literature, the geometric descriptive statistics have much in common with the lognormal distribution. The geometric mean (GM) is the antilog of the log-space mean of the lognormal distribution (μ_{ln}). Similarly, the geometric standard deviation (GSd) is the antilog of the log-space standard deviation of the lognormal distribution (σ_{ln}).

Given these relationships, we can use the geometric mean and the geometric standard deviation or the lognormal mean and lognormal standard deviation. We can estimate lognormal or geometric confidence intervals for lognormally distributed data as we would do for a normal distribution.

If the data is lognormally distributed, these geometric confidence intervals will much more accurately and appropriately describe the data than will using the arithmetic mean and arithmetic standard deviation based confidence intervals.

Table 4 summarizes the relationships of the geometric statistics to the corresponding lognormal statistics.

Descriptive Statistic	Geometric (real-space)	Lognormal (log-space)	Relationship
Geometric Mean	GM	μ_{ln}	GM = $\exp(\mu_{ln})$
Geometric Standard Deviation	GSd	σ_{ln}	GSd = $\exp(\sigma_{ln})$
First Confidence Interval ~68%	[GM ÷ GSd, GM × GSd]	[$\mu_{ln} - \sigma_{ln}$, $\mu_{ln} + \sigma_{ln}$]	
Second Confidence Interval ~95%	[GM ÷ GSd^2, GM × GSd^2]	[$\mu_{ln} - 2\sigma_{ln}$, $\mu_{ln} + 2\sigma_{ln}$]	
Third Confidence Interval ~99%	[GM ÷ GSd^3, GM × GSd^3]	[$\mu_{ln} - 3\sigma_{ln}$, $\mu_{ln} + 3\sigma_{ln}$]	

Table 4 - Relationships of Geometric and Lognormal Statistics

At first, it may seem strange to use multiplication, division, and exponentiation to calculate the geometric confidence intervals. However, remember that the geometric mean is about ratios and geometric statistics abide by the rules of logarithmic arithmetic for their log-space equivalents.

Real World Geometric Confidence Intervals

At this point, this may all seem statistical and mathematical mumbo-jumbo. However, this background is required to appropriately interpret the results from the response time measurement system.

We had monitored a particular Web site, Site A, for 30 days. Synthetic page transactions against the home page were performed every 15 minutes from 5 locations via up to a total of 15 T1 network connections. (The system was under construction during this measurement period).

The web page was updated on an almost daily basis. It consisted of 23 HTML and image objects and page weight varied in size between 36,027 and 40,232 bytes.

Table 5 contains the results of the confidence interval calculations for this long-term sample set. The geometric confidence intervals yielded almost textbook values for the confidence intervals: 67.0%, 95.3%, and 99.9% whereas the values for an ideal normal distribution are 68.3%, 95.4%, and 99.7%. Based on this alone, it certainly appears that the data might be normally distributed in log-space.

Also it is clear that the geometric mean more closely estimates the median of the page times than does the arithmetic mean.

Table 5 – Site A: Arithmetic and Geometric Confidence Intervals

Arithmetic Confidence Intervals			
Bound	**Page Time (millisecs)**	**Percent Rank**	**Confidence Interval**
AM - 3 * ASd	*-9,937.5*	*0.00%*	99.65%
AM - 2 * ASd	*-4,743.5*	*0.00%*	98.18%
AM - 1 * ASd	*450.5*	*0.00%*	92.19%
Minimum	511.0	0.00%	
Median	5,027.0	50.00%	
AM	5,644.6	60.78%	
AM + 1 * ASd	10,838.6	92.19%	92.19%
AM + 2 * ASd	16,032.6	98.18%	98.18%
AM + 3 * ASd	21,226.7	99.65%	99.65%
Maximum	458,289.0	100.00%	

Geometric Confidence Intervals			
Bound	**Page Time (millisecs)**	**Percent Rank**	**Confidence Interval**
Minimum	511.0	0.00%	
GM / GSd^3	712.0	0.04%	99.88%
GM / GSd^2	1,331.0	2.95%	95.31%
GM / GSd^1	2,487.9	18.51%	67.03%
GM	4,650.5	43.67%	
Median	5,027.0	50.00%	
GM * GSd^1	8,692.9	85.53%	67.03%
GM * GSd^2	16,249.2	98.26%	95.31%
GM * GSd^3	30,373.8	99.92%	99.88%
Maximum	458,289.0	100.00%	

Calculating standard arithmetic confidence intervals based on the arithmetic mean and arithmetic standard deviation yielded results that were so far off what would be expected if the data were normally distributed that their use seems inappropriate for describing the data. All of the arithmetic confidence interval lower bound values computed are less than the minimum value measured and 2 of the values are negative -- negative Web page download times certainly would be a wonderful thing. Also the first confidence interval bounds more than 92% of the values. This indicates that the arithmetic variance and arithmetic standard deviation are so high that the confidence interval is casting a very wide net and not localizing the arithmetic mean very well.

Figure 5 illustrates just how well the lognormal distribution fits the Page Time data compared to the normal distribution.

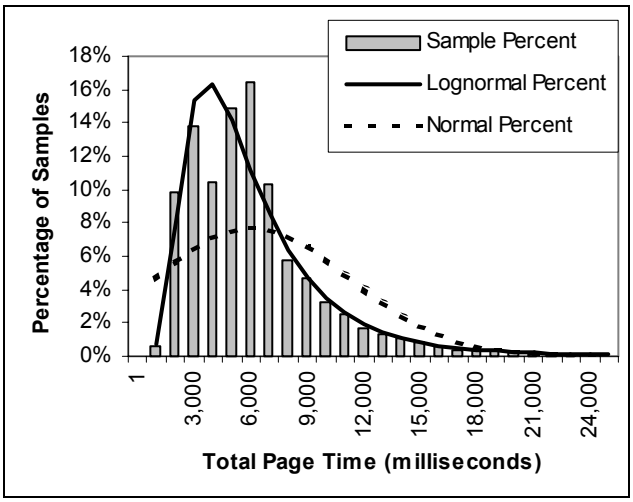


Figure 5 - Site A Histogram with Lognormal and Normal Curves

Calculating a goodness-of-fit using summed squared errors (SSE):

$$\mathrm{SSE} = \sum_{i=1}^{n} (observed_i - expected_i)^2$$

The lognormal distribution has a relatively better SSE value of only 0.007 compared to the normal distribution that has an SSE value of 0.027. (The SSE was chosen over other goodness-of-fit tests such as the Chi-Square, Pearson, or Neyman tests as a practical matter of choice - the normalizing denominators in these tests often resulted in division by zero in the case of sparse tails or floating-point precision underflow when evaluating the tail.)

Clearly, in this case, the geometric descriptive

statistics and the lognormal distribution much more accurately describe and fit the page time data than do the normal distribution based Arithmetic descriptive statistics.

Another Example with Counterintuitive Results

Now we'll look at the case of another subject Web site page – the static home page of Site B. Site B is a very lightly loaded site that significantly changed its home page three times in four days with some rather counterintuitive results:

- Doubling the total page weight bytes *reduced* the total page time by approximately 40%
- A 14% increase in total page weight bytes had no significant increase in total page times.

	Site B - Home Page 1			Site B - Home Page 2			Site B - Home Page 3			Site B - Aggregate		
First time stamp	2001/04/29 20:02			2001/05/01 00:02			2001/05/01 23:02			2001-04-29 20:02		
Last time stamp	2001/04/30 23:02			2001/05/01 22:02			2001/05/03 21:02			2001-05-03 21:02		
Total samples	644			529			1,081			2,254		
Sample groups	27			22			46			97		
Samples / group	24			24			24			23		
Total Errors Ignored	0			2			0			2		
Max Objects / Page	62			44			44			62		
Max Page Weight Bytes	37,011			69,795			75,015			75,015		
Mean Bytes / Object	597			1,586			1,705					
Total Page Time Stats (milliseconds)	**Page Time (ms)**			**Page Time (ms)**			**Page Time (ms)**			**Page Time (ms)**		
Minimum	1,602			1,920			1,645			1,602		
Maximum	37,408			25,660			174,037			174,037		
Arith Mean (AM)	12,812.6			9,659.9			9,811.2			10,634.1		
Arith Stdev (ASd)	4,920.7			3,395.5			6,274.5			5,514.0		
Geo Mean (GM)	11,805.4			9,053.3			9,013.6			9,746.6		
Geo Stdev (GSd)	1.534			1.452			1.486			1.519		
Median	12,714.5			9,364.0			9,171.0			9,870.5		
Arith Conf Intrvl	**Page Time (ms)**	**Percent Rank**	**Interval**	**Page Time (ms)**	**Percent Rank**	**Interval**	**Page Time (ms)**	**Percent Rank**	**Interval**	**Page Time (ms)**	**Percent Rank**	**Interval**
AM - 3 * ASd	*- 1,949.6*	*0.00%*	99.22%	*- 526.5*	*0.00%*	99.19%	*- 9,012.2*	*0.00%*	99.82%	*- 5,908.0*	*0.00%*	99.63%
AM - 2 * ASd	2,971.1	0.80%	96.42%	2,869.0	0.54%	96.62%	*- 2,737.7*	*0.00%*	99.39%	*- 394.0*	*0.00%*	98.42%
AM - 1 * ASd	7,891.9	15.47%	71.12%	6,264.5	16.72%	69.88%	3,536.8	1.57%	94.28%	5,120.1	6.29%	84.59%
Median	12,714.5	50.00%		9,364.0	50.00%		9,171.0	50.00%		9,870.5	50.00%	
AM	12,812.6	51.13%		9,659.9	55.19%		9,811.2	57.63%		10,634.1	56.07%	
AM + 1 * ASd	17,733.4	86.59%	71.12%	13,055.4	86.60%	69.88%	16,085.7	95.85%	94.28%	16,148.2	90.88%	84.59%
AM + 2 * ASd	22,654.1	97.22%	96.42%	16,450.9	97.16%	96.62%	22,360.2	99.39%	99.39%	21,662.2	98.42%	98.42%
AM + 3 * ASd	27,574.8	99.22%	99.22%	19,846.4	99.19%	99.19%	28,634.6	99.82%	99.82%	27,176.2	99.63%	99.63%
Geo Conf Intrvl	**Page Time (ms)**	**Percent Rank**	**Interval**	**Page Time (ms)**	**Percent Rank**	**Interval**	**Page Time (ms)**	**Percent Rank**	**Interval**	**Page Time (ms)**	**Percent Rank**	**Interval**
GM / GSd ** 3	3,269.2	0.85%	99.15%	2,957.6	0.56%	99.44%	2,746.2	0.34%	99.48%	2,780.1	0.53%	99.32%
GM / GSd ** 2	5,015.6	3.13%	96.11%	4,294.3	2.95%	95.79%	4,081.2	2.50%	96.51%	4,223.4	2.37%	96.53%
GM / GSd ** 1	7,694.9	14.82%	72.70%	6,235.2	16.28%	70.64%	6,065.2	16.27%	69.99%	6,415.9	16.48%	68.81%
GM	11,805.4	43.39%		9,053.3	46.83%		9,013.6	48.23%		9,746.6	48.86%	
Median	12,714.5	50.00%		9,364.0	50.00%		9,171.0	50.00%		9,870.5	50.00%	
GM * GSd ** 1	18,111.6	87.52%	72.70%	13,145.0	86.91%	70.64%	13,395.3	86.26%	69.99%	14,806.5	85.29%	68.81%
GM * GSd ** 2	27,786.5	99.24%	96.11%	19,086.1	98.74%	95.79%	19,907.1	99.01%	96.51%	22,493.0	98.90%	96.53%
GM * GSd ** 3	42,629.6	100.00%	99.15%	27,712.3	100.00%	99.44%	29,584.4	99.82%	99.48%	34,170.0	99.85%	99.32%

Table 6 – Site B Response Time Results

Armed with the geometric descriptive statistics and the lognormal distribution, we can validate that these results are indeed the case and then we can go hunting for explanations.

Table 6 summarizes the descriptive statistics for Web Site B looking at the three different home pages one at a time and in aggregate for the whole sampling period.

First note that even though the total page weights of Home Page 2 and Home Page 3 are 69,795 bytes and 75,015 bytes, respectively, versus the 37,011 bytes of Home Page 1 (roughly twice the bytes), the Home Page 2 and Home Page 3 arithmetic mean, geometric mean, and median total page times are all 23% to 28% *less* than the corresponding total page time statistics for Home Page 1. The most likely explanation is the 29% reduction in object requests made per page from 62 down to 44.

Second, we noted that the 14% page weight bytes increase (from Home Page 2 to Home Page 3) had no significant increase in the total page times. Relative to Home Page 1 total page times, Home Page 2 showed a 1.2% decrease in arithmetic mean, a 0.3% increase in geometric mean, and a 1.5% increase in median value. Value discrepancies appear to be just noise.

Figures 6 and 7 show the data histograms and corresponding lognormal distributions for Site B Home Pages 1, 2, and 3. These charts confirm what we saw in the descriptive statistics:

- Home Pages 2 and 3 with fewer objects typically took less total time than Home Page 1 with more objects, despite the fact that the total page weight of Home Page 1 was approximately have the weight of the other two pages.
- The total page times for Home Pages 2 and 3 are, for all intents and purposes, indistinguishable.

The added advantage of using these charts for characterizing the total page time is that they illustrate timing relationships as whole better than relying on the descriptive statistics values alone. They confirm that the tiny differences observed are relatively insignificant.

Zhi's recent paper on web page design and download performance [ZHI01] provides a model for web page times based on number of packets, rather than total page weight. His paper demonstrates how number of packets is correlated to number of page objects and his work demonstrates similar counterintuitive results as measured for Site B.

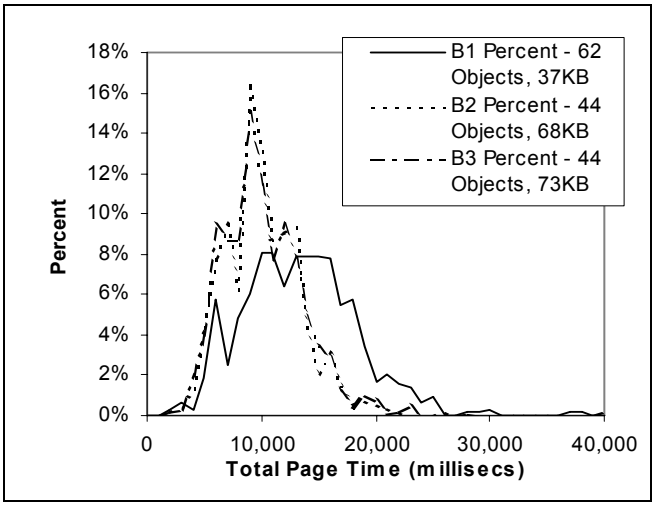


Figure 6 - Site B Home Page Histograms

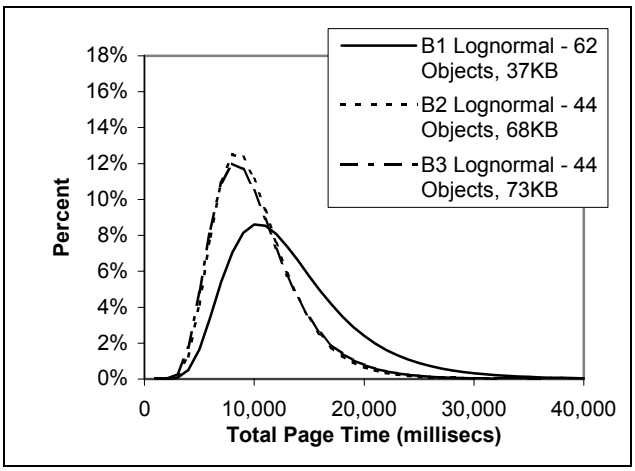


Figure 7 - Site B Home Page Lognormal Distributions

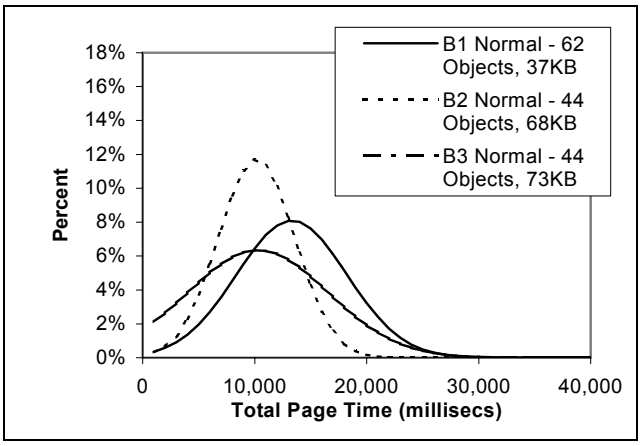


Figure 8 - Site B Normal Distributions

How Outliers Skew the Variance

Figure 8 shows the normal distributions that correspond to the data in Figure 6. Clearly some measurements are heavily influencing the computed

arithmetic variance and standard deviation and thus are affecting the shape of the distribution. In this case, only 2 measurements out of 1,081, less than 0.2%, were all it took to disturb the curves. As seen in "How Outliers Skew the Mean", these two measurements could simply be two ephemeral events on the Internet at large.

If we had looked at Figure 8 without looking at Figures 6 and 7 we might be mislead into believing that the total page times for Site B Home Pages 2 and 3 were significantly different when they are, in fact, almost identical.

We see the same problem in the Arithmetic confidence intervals for Site B Home Page 3 and for the Site B Home Page Aggregate. The outliers broadly widen the corresponding arithmetic confidence intervals whereas the corresponding geometric confidence intervals are relatively uninfluenced by the outliers.

Table 7 illustrates the affect of trimming (removing) the outliers from data set of measurements for Site B Home Page 3.

Clearly one or two anomalous events on the Internet are sufficient to wreak havoc on any SLA or performance comparison decisions based the arithmetic standard deviation – the results can be swayed by multiple seconds of time. In the case of the upper bound of the first confidence interval (mean + 1 stdev), the swing was approximately almost 3 seconds or 18% (Trim 2 relative to Trim 0).

On the other hand, decisions based on the geometric mean and the geometric standard deviation show much more stability in the presence of anomalous events. In the corresponding case of the high bound of the first geometric confidence interval (GM * GSd ^ 1), the swing was approximately 0.25 seconds or under 2%, which is not significant.

Table 7- The Effects Trimming on Site B Home Page 3 Descriptive Statistics

	Trim 0	Trim 1	Trim 2
3 Longest Measurements (milliseconds)			
5/2/2001 6:02	174,037	~~174,037~~	~~174,037~~
5/2/2001 9:02	55,672	55,672	~~55,672~~
5/2/2001 11:02	27,742	27,742	27,742
Total Measurements	1,081	1,080	1,079
Median	9,171	9,170	9,168
Mean	9,811	9,659	9,617
Variance	39,368,975	14,386,853	12,434,389
Stdev	6,274	3,793	3,526
Mean + 1 Stdev	16,086	13,452	13,143
Mean - 1 Stdev	3,537	5,866	6,090
Geometric Mean	9,014	8,989	8,974
Geometric Stdev	1.486	1.471	1.465
GM * GSd ^ 1	13,393	13,221	13,148
GM / GSd ^ 1	6,066	6,112	6,125
Lognormal E(X)	9,749	9,683	9,653
Lognormal Var(X)	16,134,463	15,047,168	14,634,134
Lognormal Stdev(X)	4,017	3,879	3,825

Why Doctor the Data If You Don't Have To?

An all too popular approach to preventing outliers from significantly skewing the mean and variance is to throw out values that are at either extreme of the distribution. There is an argument to be made that manipulating the data to make it fit a desired set of descriptive statistics rather than using appropriate descriptive statistics to describe the data, no matter the intent, is simply bad statistics.

In particular, after examining many papers and texts on statistics and performance measurements, it is finally become clear to me, just in the few months leading up to the writing of this paper, that many data sets may be better fit, modeled, or described with the lognormal distribution and geometric statistics than with the more commonplace normal distribution and arithmetic statistics. By using the correct statistical tools, the whole messy business of trimming or otherwise doctoring the data might be avoided.

Remember the stories of how scientists had for years thrown out data values that indicated a hole in the ozone layer existed above Antarctica. Trimming data to meet one's purposes is somewhat dubious at best and misleading at worst.

Final Example: Single Object Request Measurements

In trying to establish a baseline measurement for the network, Logictier set up a reference server which would support synthetic transactions from the RMS monitoring stations requesting single files of 1000, 10,000, and 100,000 bytes. The 1000 byte file was chosen in particular to represent HTTP requests for tiny objects requests that were smaller than the Ethernet Maximum Transmission Unit (MTU) of 1500 bytes.

While examining total object request times from a single location to the reference target server, I was initially disheartened that not even the geometric statistics yielded decent confidence intervals. See Table 8 – Total Page Time DNS + HTTP.

However, upon closer examination, the total object request times included DNS lookups that added significant variance to the data. Since DNS lookups are usually cached by the Web browser and are not performed for every object request, using the measurement times for the HTTP request / response only, without the DNS lookup time was justifiable. To make matters worse, since DNS lookups were obtained from geographically distributed authoritative servers, if the intent was to study the effects of physical distance from the server, eliminating the DNS times is doubly justified.

Table 8 - Single 1000 Byte Object Request Times, with and without DNS Lookup Times, DCA Location Only

	1000 Byte Object from DCA			
	Total Page Time DNS + HTTP		Total Page Time HTTP Only	
First time stamp	2001/04/29 20:00		2001/04/29 20:00	
Last time stamp	2001/05/03 22:00		2001/05/03 22:00	
Total samples	2,277		2,277	
Total Errors Ignored	1		1	
Maximum objects	1		1	
Maximum bytes	1,000		1,000	
Total Page Time Stats (milliseconds)	**Page Time (ms)**		**Page Time (ms)**	
Minimum	281		191	
Maximum	55,360		3,618	
Arith Mean (AM)	528.1		267.8	
Arith Stdev (ASd)	2,778.8		245.3	
Geo Mean (GM)	369.2		248.7	
Geo Stdev (GSd)	1.442		1.320	
Median	335.5		223.0	
Arith Conf Intrvl	**Page Time (ms)**	**Interval**	**Page Time (ms)**	**Interval**
AM - 3 * ASd	*- 7,808.2*	99.77%	*- 468.3*	99.53%
AM - 2 * ASd	*- 5,029.4*	99.76%	*- 222.9*	99.51%
AM - 1 * ASd	*- 2,250.7*	99.20%	*22.4*	99.39%
Median	335.5	50.00%	223.0	50.00%
AM	528.1	97.68%	267.8	67.78%
AM + 1 * ASd	3,306.9	99.20%	513.1	99.39%
AM + 2 * ASd	6,085.6	99.76%	758.4	99.51%
AM + 3 * ASd	8,864.4	99.77%	1,003.8	99.53%
Geo Conf Intrvl				
GM / GSd ** 3	123.1	99.02%	108.1	99.50%
GM / GSd ** 2	177.5	98.99%	142.7	99.07%
GM / GSd ** 1	256.0	97.81%	188.4	87.95%
Median	335.5	50.00%	223.0	50.00%
GM	369.2	63.19%	248.7	61.45%
GM * GSd ** 1	532.4	97.81%	328.4	87.95%
GM * GSd ** 2	767.8	98.99%	433.5	99.07%
GM * GSd ** 3	1,107.2	99.02%	572.2	99.50%

Table 8 – Total Page Time, HTTP Only, shows that removing the DNS times and focusing solely on the HTTP request / response does yield tighter geometric confidence intervals indicating that the DNS times might be better studied independently from the single object request measurements and that the lognormal distribution may be a better fit for the data than the normal distribution.

Summary

As many before me have pointed out, relying on the arithmetic mean and arithmetic standard deviation can prove unreliable, especially when comparisons of Web Site page times for performance and for financially binding SLAs are in effect.

Also note that the literature is loaded with Web response time distributions that appear to be lognormally distributed just like the examples presented in this paper. This gives cause to believe that the statistics in this paper are likely applicable. For just a few examples, see [MILL01], [TU01], and [LOOS00].

The key points to take away are:

- In small sets of samples, the arithmetic mean is highly sensitive to outliers whereas the median and geometric mean are relatively insensitive to outliers.
- The sensitivity of the arithmetic mean to a single anomalous Internet event can be on the order of seconds or even tens of seconds, thus causing a false-positive assessment of a performance SLA violation.
- In both large and small sets of samples, the arithmetic standard deviation and variance are highly sensitive to single anomalous Internet events that cause poor performance outliers.
- The geometric mean, geometric standard deviation, and geometric confidence intervals are superior to their arithmetic versions because they are less sensitive to single outliers.
- The lognormal distribution is the normal distribution of the log-transformed service or response times.
- The lognormal distribution very often fits Web site response time data and this reinforces the guidance of using the geometric mean and geometric standard deviation in characterizing Web site response times.

- The lognormal distribution, similar to it's geometric cousins, is relatively insensitive to anomalous Internet events and thus should be used for comparing distributions of response times.
- And lastly, really look at your data, the results can be surprising counter-intuitive. You never know when doubling you page weight might result in a 25% reduction in page download times.

In short, I hope I have successfully demonstrated the value of going beyond the average and statistics 101 and have inspired you to add the geometric statistics and lognormal distribution to your toolkit for evaluating Web, network, and application service times.

References

[ALLE90] A. Allen, "Probability, Statistics, and Queueing Theory with Computer Science Applications, Second Edition", Academic Press, pp 471-472, (1990).

[BERK00] K. Berk and P. Carey, "Data Analysis with Microsoft Excel", Duxbury Thomson Learning, (2000).

[BROW01] N. Brownlee and C. Loosley, "Fundamentals of Internet Measurement: A Tutorial", CMG Journal of Computer Resource Management Issue 102, p12, (2001).

[FLEM86] P. Fleming and J. Wallace, "How Not to Lie with Statistics: the Correct Way to Summarize Benchmark Results", Communications of the ACM, Volume 29, Number 3, (March 1986).

[GROS98] D. Gross and C. Harris, "Fundamentals of Queueing Theory, Third Edition", John Wiley & Sons, pp 127-130, (1998).

[GUNT00] N. Gunther and R. Jain, "The Practical Performance Analyst", iUniverse.com, (2000).

[HUFF54] D. Huff, "How to Lie with Statistics", W.W. Norton & Co., (1954).

[KLEI75] L. Kleinrock, "Theory, Volume 1, Queueing Systems", John Wiley & Sons, (1975).

[LOOS00] C. Loosley, R. Gimarc, and A. Spellmann, "E-Commerce Response Time: A Reference Model", CMG 2000 Conference Proceedings, (2000).

[MILL01] P. Mills and C. Loosley, "A Performance Analysis of 40 E-Business Web Sites", CMG Journal of Computer Resource Management Issue 102, pp28-33, (2001).

[OVER00] C. Overton, "Service Level Agreements (SLA's) Between Content Distributors and Content Creators", Keynote Systems, pp20-21, (2000).

[SEIG01] K. Siegrist, "Virtual Laboratories in Probability and Statistics", University of Alabama, Huntsville, (2001), http://www.math.uah.edu/stat/index.html

[STUR00] R. Sturm, W. Morris, and M. Lander, "Foundations of Service Level Management", SAMS, (2000).

[TART00] M. Tarter, "Statistical Curves and Parameters", A K Peters Ltd, pp 336-339, (2000).

[TU01] T. Tu, "Analysis of Broadband Performance", CMG Journal of Computer Resource Management Issue 102, p38, (2001).

[ZHI01] J. Zhi, "Web Page Design and Download Time", CMG Journal of Computer Resource Management Issue 101, pp40-41, (2001).

Acknowledgements

I would like to thank the people of Logictier who instigated and built the RMS system used to gather the data used for this paper, including, Kiran Patel, Eileen Li, Jay Molteni, Leslie Gaillard, Bill Wohnetka, Bruce Bruning, Michael Gilliam, Lee Eddy, Alex Nguyen, Erik Takaoka, Semele Halkedis. I'd like to thank Dr. Neil J. Gunther for useful feedback on early drafts, Mark Scarr who caught a significant mathematical typo, and my editor, Bill Jouris.

I would also like to thank Omar Ahmad, my father Jon Ciemiewicz, my wife, Beth, and other family and friends who have all suffered my ramblings about my personal learnings about the geometric statistics and the lognormal distribution.

About the Author

David Ciemiewicz was Sr. Principle Engineer and Director of Performance Metrics and Capacity Planning at Logictier, at the time these measurements were taken. His career included 10 years at Silicon Graphics where he participated in the development of SGI's ProDev WorkShop software development and performance tools, SGI's WebFORCE web product line, Performance Co-Pilot / Web, the WebStone benchmark, as well as having had many other roles. His career also included 4 years with Excite@Home, architecting, developing, and capacity planning servers and network services for millions of broadband consumers.