

# **Keynote Systems**

**Data Accuracy and Statistical Methodology**

**Keynote Systems, Inc.  
2855 Campus Drive  
San Mateo, CA 94403  
(650) 522-1000  
[www.keynote.com](http://www.keynote.com)**

**TABLE of CONTENTS**

**INTRODUCTION .....3**

**MEASURING THE INTERNET .....3**

    DATA VARIABILITY ..... 3

    WEIGHTED AVERAGE ..... 3

**INSURING DATA ACCURACY.....4**

    POPULATION AND BACKBONE BIAS ..... 4

    POINT OF PRESENCE BIAS ..... 4

    PC BIAS ..... 5

    BROWSER BIAS ..... 5

**USING CONFIDENCE INTERVALS.....5**

    COMPARING TWO SITES..... 5

    MINIMUM SAMPLE SIZE ..... 5

**ERROR RATES .....6**

**REFERENCES .....6**

**APPENDIX—KEYNOTE AGENT REPRESENTATION AND WEIGHTING .....7**

    WHAT KEYNOTE AGENTS REPRESENT ..... 7

    AGENT PLACEMENT ..... 7

    MULTIHOME VS. SINGLE HOME AGENTS..... 7

    WEIGHTING FACTORS..... 7

## Introduction

This paper explains the key statistical issues involved in presenting summary data about Internet performance and describes how Keynote Systems ensures that statistics are accurate and reliable. Measuring the Internet is a complex task both in terms of how measurements are taken and how the results are presented. A complete picture of Web performance includes measurements of the following three elements:

- Average download time—represents the average download time for your Web site
- Data variability—reflects the consistency of download times
- Download errors—represents all failed download attempts

Measurements of these three elements can be incorporated into a statistical model to provide an overall assessment of your Web site performance. This statistical model can capture all aspects of performance and be used to answer questions such as the following:

- Is my Web site better than my competitors’?
- Has my Web site improved since last month?
- What is the typical experience of a user coming to my Web site?

You can, of course, still use Keynote statistics individually. For example, you do not need summary statistics to use download error measurements for debugging, to focus on a component of the data (such as DNS times), or to answer a specific question such as “Why are the agents in Tokyo having trouble getting to my site?”

## Measuring the Internet

The starting point for presenting information about Web site performance is generating a number that represents typical or average performance. The two leading metrics for such a number are the mean and median. (From a statistics point of view, the word “average” can be used to refer to either metric.) The simplest statistic calculates the grand mean or grand median of all the data for a time period such as a day. Keynote uses a grand mean to provide this measurement for the first order views of the data. The grand mean indicates when and where abnormal performance needs to be investigated. It also provides a basic comparison between two time periods or between two sites.

### *Data Variability*

The statistical nature of Internet performance data causes problems when simple means are used for more rigorous comparisons. A more definitive statement about performance needs to include information about the variability of the data. Data variability is important because two Web sites with equal average download times, but different data variability, are not equivalent. For example, two Web sites might both have an average 10-second download time. However, for one Web site, all download times might be between 9 and 11 seconds. For the other, half the download times might be 1 second and the other half 19 seconds.

Keynote calculates a confidence interval to estimate data variability. When comparing two data sets with approximately equal numbers of data points, a narrow confidence interval indicates consistency in performance, and a wider confidence interval reflects greater variability in the data and indicates less consistency in performance. For instance, the second Web site in the previous example would have a wider confidence interval. The confidence interval also tells you how confident you can be that the true mean lies within the interval. Keynote calculates a 95% confidence interval, meaning that there is a 95% probability that the true average of performance, as experienced by the population of end users that Keynote represents, falls within the calculated confidence interval.

### *Weighted Average*

Two facts about Internet data make it difficult to calculate a confidence interval. First, Internet performance data is not normally distributed. Instead, the data tends to have a heavy tail, meaning that a relatively small

number of very high download times skews a mean calculation. This characteristic is clearly confirmed in Keynote data<sup>1</sup>. Second, each measurement agent has its own unique distribution of measurement data.

To overcome these difficulties, Keynote uses a weighted mean of medians calculation to represent average response time and uses a resampling technique called the bootstrap method to calculate the confidence interval. To calculate the weighted mean of medians, a median is calculated for all of the data from each individual agent for the time range in question. If there are 50 agents, the result is 50 medians. Each median eliminates the effect of a heavy tail since no individual measurement can unfairly skew the data. The weighted mean of these 50 values is taken and presented as the overall average. The weighting for each agent eliminates bias and is discussed in the next section. Each agent is handled separately and has its own unique distribution of data. As a result, the bootstrap cannot put all the data into one big data set as if they all had the same distribution but must calculate a combined 95% confidence interval by combining the different agent sets. Keynote has a patent pending on this technique of combining agent data to produce an accurate average and confidence interval.

## Insuring Data Accuracy

The difference between the Keynote estimate of the average download time and the true average download time is the measurement error. The measurement error is a combination of bias and variance, so minimizing both provides the most accurate possible estimate of the true average response time. Variance is tied to the number of measurements and is steadily reduced as Keynote adds more agents. Bias is the result of taking measurements that over- or underrepresent particular segments of the Internet user population. There are five types of bias:

- Population bias
- Backbone bias
- Point of presence (POP) bias
- PC bias
- Browser bias

### *Population and Backbone Biases*

Keynote minimizes population and backbone biases by using weights to rank the importance of each agent. The agent weights account for both Internet population and backbone market share. Keynote considers an agent to be part of a metropolitan area if it is within the CSMA (Consolidated Metropolitan Statistical Area), PMSA (Primary Metropolitan Statistical Area), or MSA (Metropolitan Statistical Area) of the city by which it is identified. (These metropolitan areas are defined by the U.S. Census Bureau, which ensures that the definition is standard and consistent.) The Internet population weighting in the calculation is done to reflect the percentage of the population that each agent represents. For example, the metropolitan New York area has 2.8 million people who have access to the Internet at work, while Portland has 200 thousand. All of the agents in New York will be weighted, as a group, more than 10 times higher than Portland. If this weighting were not done, each agent would automatically be weighted equally. The backbone to which the agent is connected is also incorporated into the agent weight to reflect the market share of that backbone.

Weights are updated regularly to reflect changes in the Internet population and backbone market share. The appendix provides more information on weighting.

### *Point of Presence Bias*

Point of Presence (POP) bias reflects all the variables within a metropolitan area that can vary between multiple POPs owned by one backbone provider and that can affect performance. POP bias also captures differences within a POP, such as different routers, that would affect performance between users who connect to that POP. The list of POP specific variables includes

- POP bandwidth and peering

- Router differences
- Router port
- Routing table within the router

### *PC Bias*

PC bias includes anything about the PC taking the measurement that affects accuracy. Instead of attempting to eliminate bias by representing a variety of different PCs, Keynote standardizes on a particular configuration of PC hardware and the Microsoft Windows NT operating system. There are two reasons for this approach. First, Keynote is primarily concerned with measuring the impact of the Internet, server, and Web page content and does not try to exactly reproduce the effects of a particular hardware configuration. Using a standard PC configuration allows us to focus on identifying problems on the Internet and on Web servers. Second, it is not possible to effectively represent the large number of PC configurations currently in use. We use the Windows NT operating system because the Windows TCP/IP protocol stack is used by the overwhelming majority of users connecting to the Internet.

The Keynote agent software is a multithread browser that takes multiple measurements simultaneously. We have rigorously tested the agent to determine its capacity, and we keep the agent running below that capacity to minimize the bias introduced by the agent software itself.

### *Browser Bias*

Browser bias accounts for performance differences experienced by users with different types of browsers, such as Internet Explorer and Netscape Navigator. Keynote currently emulates a multithread browser that uses four parallel threads and identifies itself in the user agent string as Mozilla-compatible. Keynote is primarily concerned with measuring the impact of the Internet, server, and Web page content and does not try to exactly reproduce the effects of a particular browser. The Keynote measurement agent does represent a typical user experience and has been tested to insure that its download times are not statistically different from Internet Explorer's download times.

## **Using Confidence Intervals**

The importance of studying the distribution of data was discussed in the earlier section of this paper on data variability. With Internet data, calculating a confidence interval is not straightforward because the data is not distributed normally. The bootstrap is used to calculate confidence intervals<sup>2</sup> and takes into account the fact that each Keynote agent has its own unique distribution of data. The bootstrap method is a technique that is relatively new in the last 10 to 15 years but is, nevertheless, a well-accepted method in the statistics community. Keynote uses the S-Plus statistics server from Mathsoft Corporation to calculate the confidence interval.

Confidence intervals are only valid when calculated over time intervals in which the probability distribution is fairly constant. Therefore it is possible to calculate an interval for daytime activity or nighttime activity. However, since day and night performances have different characteristics, it is not a good idea to calculate a single confidence interval for a full 24-hour period.

### *Comparing Two Sites*

Keynote uses the confidence interval when comparing two sites to determine which one is better. If the confidence intervals do not overlap, one site can be said to be statistically different from the other. If two sites have a different number of samples, they can still be compared as long as confidence intervals are used. Sites are only compared when data from the same time period is used.

### *Minimum Sample Size*

Six samples are the minimum needed to make a 95% confidence interval feasible. A site can be evaluated and an average and confidence interval can be stated with as few as six data points from each agent over a specified time period. To get 95% confidence for the median, we have to be 95% confident that the median lies between the highest and lowest observations. If not, we cannot give a valid upper or lower confidence bound for the median.

To understand why fewer than six data points cannot give a 95% confidence interval, assume we have only three data points. With three data points, all data points will be below the median 1/8th of the time and all will be above the median 1/8th of the time. The median is between the highest and lowest observations only 75% of the time. The list below shows the number of data points required to provide specific confidence intervals.

- Three data points provide a 75% confidence interval.
- Four data points provide a 87.5% confidence interval.
- Five data points provide a 93.75% confidence interval.
- Six data points provide a 96.8% confidence interval.
- Seven data points provide a 98.4% confidence interval.
- Eight data points provide a 99.2% confidence interval.

Although a 95% confidence interval is mathematically feasible with six data points, it is only practical because Keynote has a large number of agents, each measuring data and producing six data points.

## **Error Rates**

Errors occur when a Web page fails to download completely. Errors can result from many causes, including network or server failures. The error rate is the percentage of all download attempts that result in an error, and it should always be reported along with the download time to give a complete picture of performance. Ignoring the error rate can result in erroneous conclusions about the performance of a site or a comparison between sites because a site may be fast but error prone.

When a particular image on a Web page does not download correctly, that error is defined as a component error. A component error is equivalent to a complete download failure of the page in Keynote's calculation of the error rate. The reason that a complete page failure and a component error are considered equivalent is that some Web pages are dominated by one large image while others are made of many smaller images. There is no way to know the importance of a particular image, so Keynote defines a successful page download as one in which every component of the page downloads without error.

## **References**

<sup>1</sup> Paxson, Vern. "An Introduction to Internet Modeling and Measurement." SIGCOMM '98.

<sup>2</sup> Efron, Bradley and Tibshirani, Robert, J. *An Introduction to the Bootstrap*. Chapman and Hall, 1993.

## APPENDIX <sup>3</sup>/<sub>4</sub> **Keynote Agent Representation and Weighting**

This appendix explains the current implementation and future plans for Keynote's agent network and how it will represent the population of Internet users.

### *What Keynote Agents Represent*

The aggregate performance seen by Keynote agents represents the typical Internet performance seen by business users using dedicated connections from large backbone providers in large cities. If another set of agents were placed on the Internet using the same criteria that Keynote uses, it would produce an aggregate performance number close to the number from Keynote even though it came from a different set of agents.

### *Agent Placement*

Keynote agents are a representative sample, not an attempt to audit every possible access router in every possible ISP. In order to represent the user base properly, Keynote agents must have both geographic and topological diversity. Keynote agents are placed only in the top 25 metropolitan areas and the majority are placed on major backbones in those cities. Because Keynote agents are a sample, they will not necessarily be placed in identical numbers in each city or on the same backbones within each city. When viewed as a group they correctly represent a sample of the business users, but within any individual city, they will not represent every backbone in the city.

Two important criteria are used when selecting a facility for an agent—backbone connectivity and the reliability of the local facility. Backbone connectivity is important because Keynote agents represent the online population, most of whom are connected by large backbones. The reliability of the local facility is important because it affects the reliability of the agent and because Keynote represents commercial users who tend to use the more reliable, more professional ISPs in their cities. Keynote therefore tries to locate agents in larger facilities that have 24x7 coverage and adequate bandwidth. We also try to place an agent no more than three hops away from the primary backbone.

Keynote already has sufficient agents in the U.S. to produce an acceptable statistical confidence interval. A typical range for the confidence interval is +/- a few tenths of a second for a business day of data (6am to 6pm). To further improve the confidence level, additional agents will be added during 1999 to U.S. ISPs. Keynote's target is to have 100 agents in the U.S. by the end of 1999. Outside of the U.S., our target is to add 30 new agents to our 20 existing agents, resulting in 150 agents worldwide. The additional Keynote agents added in 1999 will also increase our diagnostic capabilities. As Keynote adds more agents to more backbones in more cities, the increased diagnostic data (trace routes, pings, detailed DNS analysis, download component times, and so on) will speed the detection and repair of backbone problems.

### *Multihomed versus Single-homed Agents*

Keynote has two types of agents: single-homed and multihomed. Single-homed agents are connected by a dedicated backbone that is used to access all Web sites on the Internet. Multihomed agents are connected by multiple backbones and can potentially use any of the backbones to connect to a particular Web site. Both types of agents may traverse multiple backbones when connecting to a particular Web site, but the backbone to which an agent is connected determines the group of users that it represents. For example, if an agent is connected by a UUNET T1 to the Internet, it represents UUNET users even if data leaves the UUNET network and traverses other backbones to connect to a Web site.

Both types of agents are needed to fully represent Internet users. Currently many large corporations (especially those with significant Web sites) are multihomed, and smaller corporations tend to be single-homed. We use both types of agents so that our measurements will not be biased toward users in either small or large companies. Keynote targets a mix of 75% single-homed agents and 25% multihomed agents, with one multihomed agent per city.

### *Weighting Factors*

In order to correctly represent the user population when combining data from multiple agents, a weighting factor must be assigned to each agent so that a weighted average can be calculated. The weighting factor is used to insure that the various agents are not over- or underrepresented in the overall average. If a weighting factor were not used, equal weighting would be applied by default to each agent when an

average was calculated. Without a weighting factor, an agent on UUNET in New York would carry the same importance as an agent on UUNET in Miami even though the populations of the two cities are very different.

The weighting factor for each agent has two components—Internet population size and backbone importance. The sum of the weighting factors for all the agents in a metropolitan area equals the population of Internet users in that metropolitan area. The weighting factor for each agent equals the population of the metropolitan area it is in multiplied by a fraction indicating the importance of the backbone it is on. If an agent is multihomed, the backbone factor itself is a weighted average of all the backbones to which it is multihomed. For example, if there were four agents in New York, they could have the following weighting factors:

<i>Agent</i>	<i>Backbone</i>	<i>Internet Population Component</i>	<i>Backbone Component</i>	<i>Agent Weight</i>
NYC Agent #1	Large	28	.3	(28)(.3)=8.4
NYC Agent #2	Large	28	.3	(28)(.3)=8.4
NYC Agent #3	Medium	28	.2	(28)(.2)=5.6
NYC Agent #4	Medium	28	.2	(28)(.2)=5.6
<b>TOTAL</b>			<b>1</b>	<b>28</b>

The first number, 28, represents the Internet population of New York (2.8 million) and does not change. The second number represents the importance of the backbone the agent is connected to and varies based on the general market share of that backbone nationally and locally and whether it is multihomed. The sum of the backbone factors always adds up to 1 so that the overall weighting of New York is always 28. If a new agent is added to New York, all of the backbone factors are adjusted so that their sum still equals 1 and the overall weighting of New York is still 28.