



Whitepaper

Measurement and
Characterization of
Latency in Trading
Networks

Contents

1. Describing Latency	3
2. Position and Range	4
3. Measuring Position	6
4. Measuring Range	7
5. Timescales	9
6. Conclusion	10

1. Describing Latency

The advent of algorithmic trading has made latency one of the most important performance metrics for market data dissemination and order management systems. System providers are under pressure from users to continuously reduce latency, and those who can demonstrate the fastest systems will achieve a clear

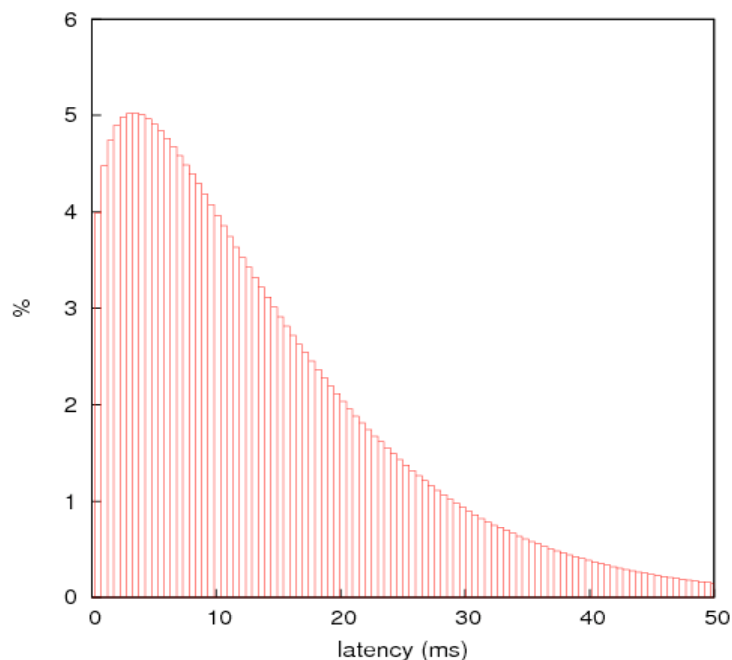
The latency that messages experience in practice can be strongly affected by system interactions and usage patterns

competitive advantage. To exploit this advantage, service providers need to measure and communicate information about latency in a manner which users find easy to understand, and easy to compare across competing services. Ideally, this information could be suitable for

inclusion in a service level agreement. But measuring latency and presenting the results in a clear intelligible fashion can prove trickier than expected.

The latency that messages experience in practice can be strongly affected by system interactions and usage patterns that are difficult to reproduce in the lab, or using synthetically-generated message streams. While lab tests allow equipment to be profiled under controlled conditions, ultimately what matters is the latency experienced by actual user messages in the live environment. Measuring end-to-end message latency with sub-millisecond precision in a live distributed system is challenging, but feasible - and in a previous Corvil whitepaper we described the type of instrumentation which can be used to do this. Once the data is available, the next step is to provide a succinct and accurate description of measured

Figure 1
Latency measurements collected over a period of time typically form a distribution spanning a broad range of values.



performance for user consumption.

Measured latency values typically vary substantially from moment to moment, due primarily to changes in system load. Fluctuations in load occur on both short and long timescales, influenced by machine-generated bursts of traffic and changes in user or market activity. These fluctuations can produce transient backlogs within processing systems and at network bottlenecks. The

When the spread of measured values is large, the average by itself does not capture the full performance picture.

minimum latency is determined by the delay needed to process and transfer the simplest type of message during periods when the system is idle and no backlogs are present. The largest latency values are measured at times

when the system is very busy, and potentially backlogged by tens or hundreds of messages.

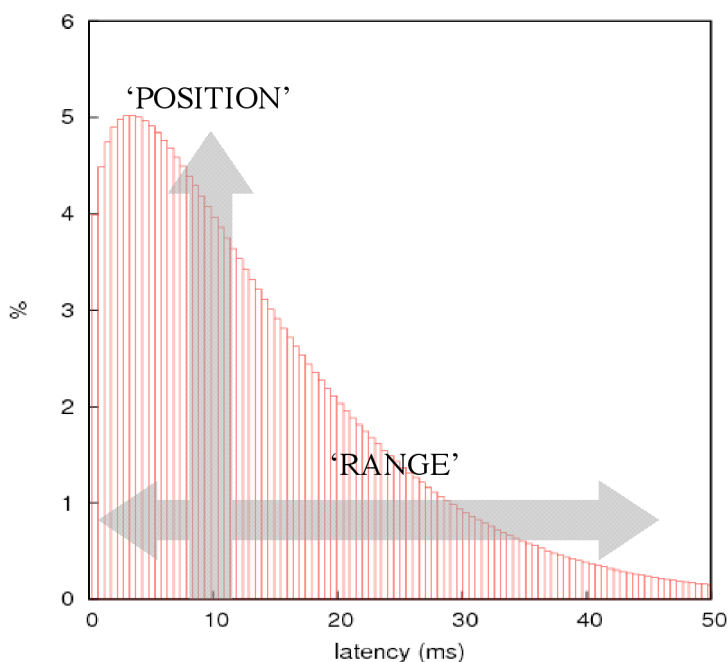
For these reasons, a set of latency measurements collected over time will typically span a broad range of values, in which the maximum value is many times larger than the minimum. This data is sometimes summarized by stating just the average latency value. But when the spread of measured values is large, the average by itself does not capture the full performance picture.

This paper describes more complete methods of summarizing latency performance, and the features needed in an analysis system intended to help automate the process.

2. Position and Range

Summarizing a set of varied data is the province of statistics, which provides a range of different

Figure 2
Some summary statistics measure the 'position' or 'centre' of a distribution, while others measure its 'range' or 'spread'.



tools for the task. For our purposes, the available tools can be separated into two camps: those which tell us about the 'position' or 'center' of the latency distribution, and those which measure its 'range' or 'spread'. For example, measures of position include the average value and the median value (i.e. the value which is larger than exactly 50% of the data). The range of a set of positive data can be measured using the value of a high percentile (such as the 95th or 99th), or the fraction of values exceeding a stated threshold.

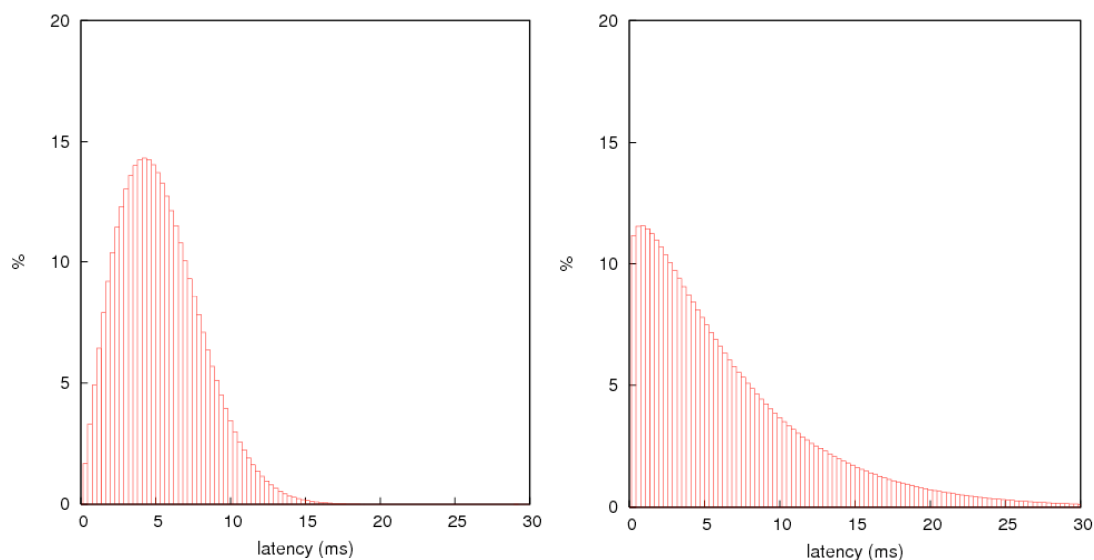
'Position' and 'range' can each be measured in a number of different ways, but there is no way of reducing these two aspects of a data set to a single figure without losing information. If we want to describe performance using just a single latency figure, then we will have to choose between describing only the position of the distribution or only its range. Using a single figure might be adequate in some cases – for example, if there is a strong preference for simplicity, or if the range of measured values turns out to be very small. In general, however, latency measurements will have a broad range, and there are good reasons to think that both position and range will matter to users.

To a financial trading application the latency of every message is potentially important.

To a financial trading application the latency of every message is potentially important. This implies that the upper range of measured latency values should not be ignored when describing performance. In addition, there is no 'minimum value' below which latency no longer matters. This suggests that the position of the bulk of measured values remains relevant, even when the upper range is already known. A few examples will help to illustrate these points.

Figure 3 shows two sets of data which have roughly the same 'position' – the average value is 5ms in both cases. But the two data sets differ markedly in their range – in the data on the left, less than 0.2% of the values exceed 15ms while

Figure 3
Two distributions of latency measurements. Both have the same average value (5ms). In the distribution on the left, less than 0.2% of the values exceed 15ms, while this level is breached by almost 10% of the values on the right.



on the right almost 10% of the values exceed this level. The latency profile on the left is likely to be preferred by users - a fact which would be hidden if only the average value were reported.

The two data sets of Figure 3 can be differentiated by reporting their respective ranges – for example, by specifying the fraction of values which exceed 15ms. But Figure 4 shows another example where reporting the range alone proves inadequate. Here we see two data sets in which the fraction of values exceeding 15ms is identical (1%). But now the positions of the two distributions, as measured by their average values, are very different (less than 3ms and more than 9ms, respectively). In the data set on the left a larger number of messages achieve relatively lower levels of latency, and this profile clearly represents a superior level of service.

These two examples show that both the position and the range of the measured latency distribution are important aspects of system performance. A complete description will in general have to include both.

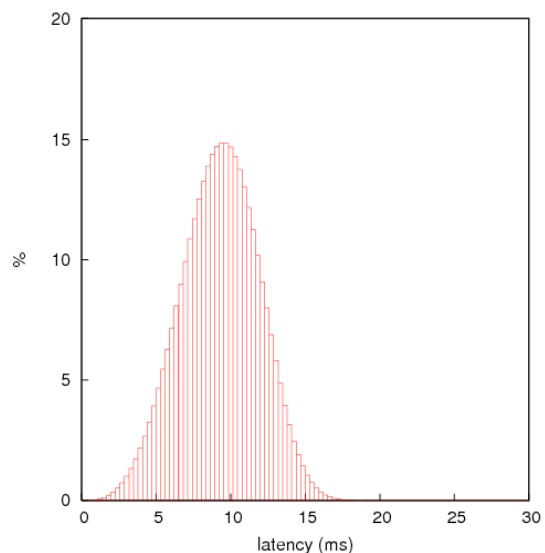
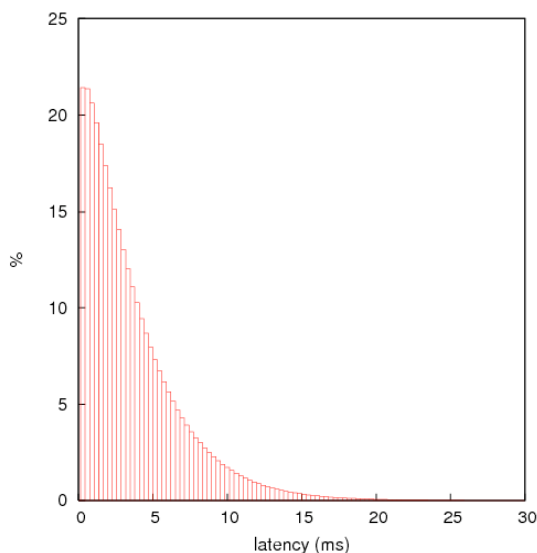
3. Measuring Position

The arithmetic average, the median value, the geometric average, and the 'mode' (the value which occurs with the highest frequency) are all measures of a distribution's 'position'.

Extreme data points can have a disproportionate impact on the value of the arithmetic average

Intuitively speaking, each of these statistics carries information about where the data set is 'centred' or 'anchored'. Of these four, the arithmetic average and the median are the most widely understood and therefore the best suited for inclusion in a user SLA.

Figure 4
A second pair of latency measurement distributions, this time with the same 'range' (as measured by the fraction of values exceeding 15ms). The average of the values on the left is less than 3ms, while on the right the average value exceeds 9ms.



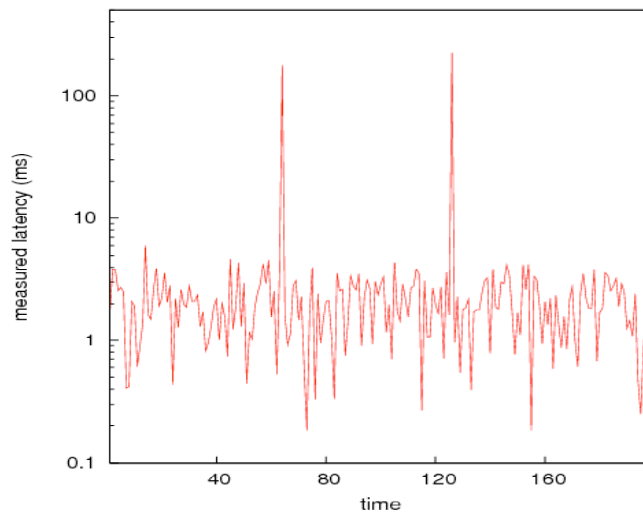


Figure 5

A series of 200 measurements containing two unusual extreme values (the latency values are shown on a log scale). The presence of these two values doubles the arithmetic average from 2000us to 4000us. They alter the median value by only 2us (from 1985us to 1987us).

The median is sometimes preferred to the arithmetic average in cases where the data is likely to contain a small number of extremely large values. Extreme data points can have a disproportionate impact on the value of the arithmetic average, but arguably should not be included when measuring a distribution's position; because they are not typical (they may of course have implications for the distribution's range). The median value provides a more robust measure of position which is insensitive to

The median value provides a more robust measure of position which is insensitive to extremes.

extremes, as illustrated in Figure 5. However, where the data does not contain sporadic extreme values, the arithmetic average remains suitable and has the advantage of being perhaps the most widely used and understood of all statistics.

4. Measuring Range

Intuitively, the 'range' of a distribution provides information about its width. For a set of positive values, possible measures of range include the variance (or standard deviation), the maximum value, the value of a high percentile such as the 95th or 99th, or the fraction of values exceeding a stated threshold.

In the context of message latency, the variance of the distribution does not have a strong interpretation in terms of performance and is not a commonly-used metric. The maximum latency value on the other hand has a very strong performance interpretation, which might lead to problems of another kind. A statement of maximum latency can be understood to mean that the system has been engineered to ensure that latency can never exceed the specified value. But in practice few systems are engineered to provide hard latency bounds

under all conditions. Occasional high latency events occur for numerous reasons, including system reconfiguration, equipment defects or failures, or just unusually high transient loads.

Given that maximum latency is not normally limited by design, its value has to be determined from measurement. This raises another set of problems: the value of the maximum can be quite unstable, and is difficult to measure in a repeatable fashion. It is often determined by the position of a single data point, and is prone to jump around from hour to hour and day to day. To reliably derive a statement of maximum latency requires a long observation period and a wide margin of error.

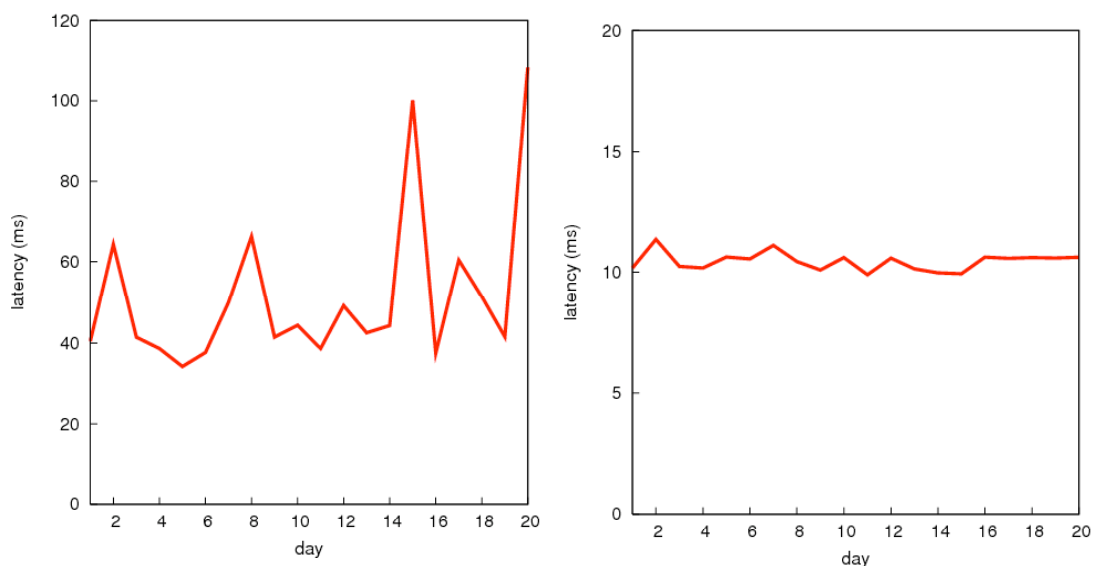
Maximum latency is not normally limited by design, its value has to be determined from measurement

The use of a high percentile, such as the 95th or 99th, can avoid these pitfalls while still providing a highly relevant performance metric.

Describing 99th percentile latency does a better job of correctly setting user expectations, because it does not create the impression that the system provides an absolute latency bound by design. And given sufficient data, the 99th percentile is a relatively stable measurement which will not shift around appreciably unless the underlying data distribution changes. See Figure 6 for a comparison of the maximum and 99th percentile measurements.

The fraction of values exceeding a stated threshold also provides a stable measure of latency range, with similar properties to those of a percentile. It may also be easier for users to interpret. Its main disadvantage is that there is no natural threshold value associated with application performance (lower latency is always better), and if system performance improves over time, then the threshold used to describe it may have to be revised downwards repeatedly.

Figure 6
Maximum (left) and 99th percentile (right) latency computed each day from the same underlying data set.



5. Timescales

Statistics such as the average, median, or percentiles, are computed from data collected over a period of time. For example we can choose to compute 99th percentile latency from measurements collected over each business day, or over each hour during the business day. The choice of which timeframe to use has implications for the strength of any performance statement based on the results.

The daily 99th percentile latency cannot exceed the maximum hourly 99th percentile, and often lies significantly below it. For this reason, to say that hourly 99th percentile latency does not exceed a certain value represents a better level of service than an equivalent statement about the daily percentile. The same is true for other statistics such as the average or the median.

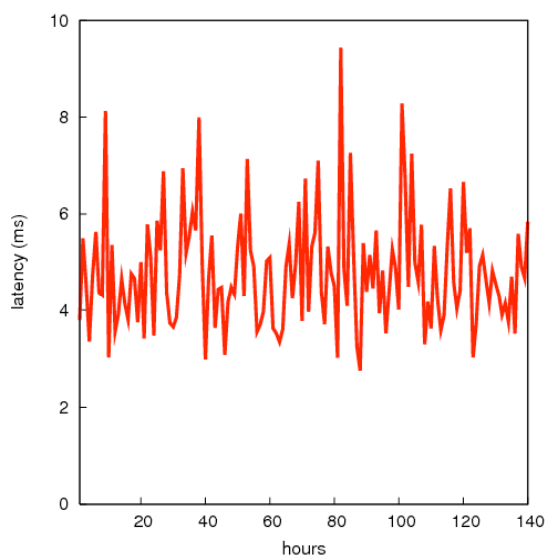
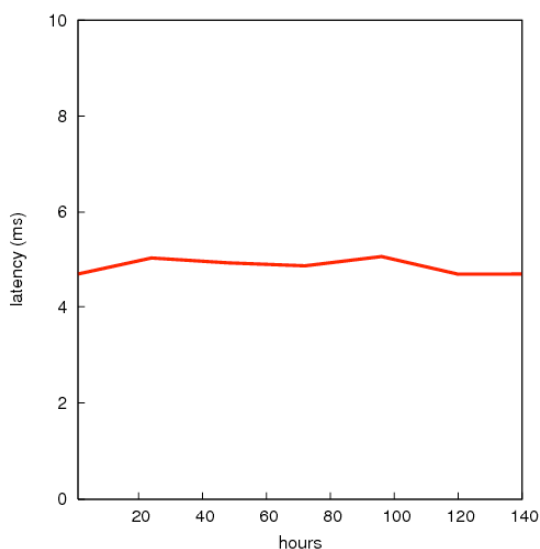
From the user's point of view, a service which keeps the percentage of messages experiencing high latency below 1% over the course of the day, might in fact allow this percentage to reach 5% or 6% during a given hour. The difference can be made up during the remaining hours. But if the high-percentage hour happens to have

particular importance (for example, it is the hour following the release of important economic data), this drop in performance could have a material impact. Clearly most users would prefer to be assured that the high-latency percentage will be below 1% in every hour.

When the spread of measured values is large, the average by itself does not capture the full performance picture.

In the same way, using an even shorter measurement timescale such as 5 minutes leads to a still stronger statement of performance, all other things being equal. In practice, however, performance statistics become more variable as

Figure 7
Daily (left) and hourly (right) 99th percentile latency, computed from the same set of underlying measurements. In this example, the 99th percentile is usually just under 5ms over the course of a day. But it can reach nearly twice this level during individual hours within each day.



the timescale is reduced, because there are fewer measurements available within each time

Performance targets based on short measurement timescales substantially more difficult to achieve.

period to compute them. This can make performance targets based on short measurement timescales substantially more difficult to achieve. The appropriate choice of timescale will ultimately be a trade-off between variability and user demands based on competing service offerings.

6. Conclusion

Performance monitoring in a financial trading environment requires a measurement system that can accurately determine message latencies in the sub-millisecond range. Measured latencies must then be collected and summarized over a period of time in order to characterize performance. Specifying just the average latency, measured over an arbitrary timeframe, does not provide a full description of performance when the measurements cover a wide range of values. The upper range of the

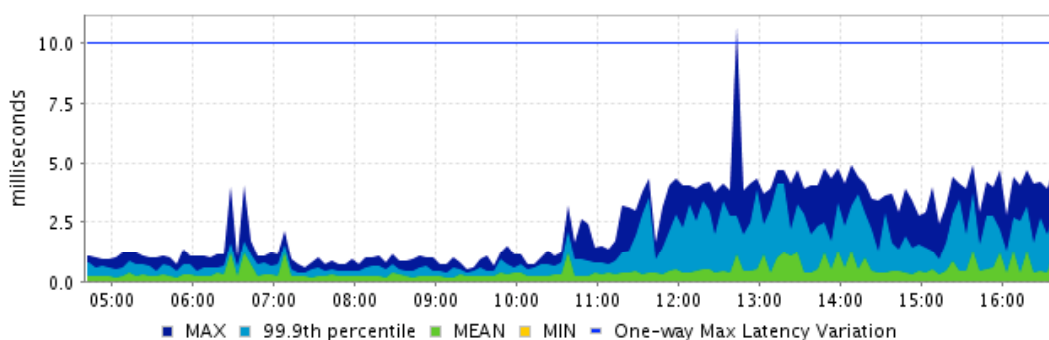
latency distribution and the timeframe used for measurement are also important information.

A monitoring system can assist in the characterization process by providing flexible summary statistics such as average, median, and percentile latency, computed over a configurable timeframe and using data collected over periods of days, weeks, or months. These performance statistics can then be used as the basis of an

Statistics such as averages and percentiles are computed on demand for longer timeframes by rolling up the stored distributions to each timeframe period ending at a five-minute boundary.

SLA, either to be shared with users or just for internal tracking purposes. Continuous monitoring of the attained performance level against the SLA requirements will provide early warning of any deviations, and assist with trouble-shooting.

Figure 8
Monitoring latency with CorvilNet.



1.102 ms	average latency
5.294 ms	99.9th percentile latency
5 minutes	busy period
1 week	measurement duration

During every 5-minute period over the past week, average latency was below 1.102ms and 99.9th percentile latency was below 5.294ms

In CorvilNet 4.0, these capabilities are delivered by measuring latencies with an accuracy of a few tens of microseconds, and then storing the full distribution of measured values for each five-minute period in a database. From this data, statistics such as averages and percentiles are computed on demand for longer timeframes by rolling up the stored distributions to each timeframe period ending at a five-minute boundary. For example, the system can be configured to compute and monitor 99th percentile latency over all one-hour periods ending at each five-minute boundary. The timeframe for computing each performance

Performance characteristics can also be programmed back into CorvilNet as a specification of performance requirements.

statistic – called the 'busy period' in CorvilNet – can be as short as five minutes or as long as one week. Data is stored for up to 60 days on the measurement appliance, or longer on external disks.

Table 1
Latency characteristics as determined by CorvilNet.

A monitoring system of this type allows latency performance to be fully characterized in a form that indicates the position and range of the measured distribution, the performance timeframe ('busy period') and the duration of measurement.

Data is stored for up to 60 days on the measurement appliance, or longer on external disks.

Performance characteristics can also be programmed back into CorvilNet as a specification of performance requirements. The system will then continuously compare the measured performance level against the specified requirements, providing email or SNMP alerts whenever a breach is detected. Trouble-shooting is supported by capturing all packets to disk during any violating event, and providing a graphical drill-down event inspection facility.



Corvil Ltd., 2nd Floor, 6 George's Dock, IFSC, Dublin 1, Ireland

T +353 1 859 1000 **Tech Support** +353 1 859 1010 E info@corvil.com W www.corvil.com

Copyright © 2008 Corvil Ltd. Corvil is a registered trademark of Corvil Ltd.
All other brand or product names are trademarks of their respective holders.