Google Cloud

VPC Flow Logs: Understanding Byte and Packet Counts

Accuracy and Sampling Explained

Ephi Sachs, Danny Raz

Based on prior work by Leonid Raskin

December 2025

TL;DR

Individual VPC Flow Logs records provide estimates for byte and packet counts due to sampling. However, when logs are aggregated over longer time windows or across multiple resources, these estimates are highly accurate.

Example use cases:

- **Broad Trends:** Analyzing ~100k packets is sufficient to accurately identify major traffic drivers and troubleshoot issues.
- **Precise Accounting:** Aggregating larger volumes (millions of packets) ensures **>99%** accuracy, making the logs suitable for detailed cost allocation.

Overview

VPC Flow Logs provides valuable insights into network traffic patterns within your Virtual Private Cloud (VPC). To manage the massive volume of network data, it uses **sampling**.

While the logs represent a subset of actual packets, the reported byte and packet counts are **extrapolated to estimate the total traffic**, compensating for the sampling rate. This means the numbers are designed to reflect the full volume of your traffic, not just the sampled subset. These reported counts provide a highly accurate estimate of the total traffic, assuming the number of sampled packets in your analysis is high.

However, for small or short-lived flows where the number of sampled packets is low, the margin of error is higher. To achieve high accuracy, the most effective strategy is to **increase the effective sample size** of your analysis. By aggregating logs—for example, summing traffic over a longer time window or grouping multiple connections—you analyze a much larger volume of packets. As this volume increases (e.g., into the millions of packets), the statistical accuracy of the reported counts approaches 100%.

See the Example section for a detailed calculation of how aggregation improves accuracy.

How Sampling Works: Primary and Secondary Sampling Rates

VPC Flow Log sampling occurs in two stages:

---

1. **Primary Sampling (system determined):** The underlying infrastructure performs an initial sampling of network packets, which are then aggregated into flow logs. The initial sampling rate is not configurable by the user and varies based on the traffic volume handled by the physical host running your virtual machine (VM) instances. Higher traffic volume often leads to a lower sampling rate.
    - *Example*: On a quiet host, the primary sampling rate might be 100% (1 in 1). On a host processing massive traffic, the system might drop to sampling 0.1% (1 in 1000) of packets. This primary rate sets the maximum possible resolution for your logs.
2. **Secondary Sampling (user configurable):** You can further reduce the volume of logs by configuring a secondary sampling rate. This rate applies to the flow logs generated by the primary sampling process.
    - *Example:* If you set the secondary rate to 0.5 (50%), the system will discard half of the flow logs.

The Effective Sampling Rate

When a flow log is generated, the effective sampling rate is determined by multiplying the two sequential stages:

$$Effective\ Rate = Primary\ Rate \times Secondary\ Rate$$

- *Scenario:* If the underlying system's primary sampling rate is 0.1 (10%) and the user-configured secondary sampling rate is 0.5 (50%), the resultant effective rate is 0.05 (5%).

This final effective rate is then used to statistically extrapolate the original traffic volume. For example, if 5 packets are observed at a 5% effective rate, the system estimates the total traffic was 100 packets ($5 \div 0.05$).

Computing Accuracy from Reported Packet Volume and Primary Sampling Rate

The accuracy of the byte and packet counts reported in flow logs is primarily influenced by the total number of packets and the effective sampling rate (primary × secondary). The time interval in which these packets were sent does not affect the analysis.

The following table shows the estimated accuracy with a 95% confidence level.

Number of packets sent (as reported in logs)	Accuracy (%) for Cloud hosts, by host percentile[1]			
	Median host	90th percentile	99th percentile	99.9th percentile
10K	93	81	66	46
100K	97	94	89	83
1M	99	98	97	95
10M	100	99	99	98
100M	100	100	100	99
1B	100	100	100	100

[1] The accuracy is based on the average primary sampling rate for these hosts and 100% secondary sampling.

Interpreting the Table

- Rows (packet volume): The rows represent the total number of packets reported in the analyzed logs. As you move down the rows (increasing the volume of analyzed packets), the accuracy percentage generally increases.
- Columns (host percentiles): The columns represent percentiles of physical hosts running virtual machine instances.
    - *Example*: The median host column reflects the primary sampling rate for the median Google Cloud host. This means 50% of hosts provide this level of accuracy or better.
    - *Trend*: Moving to the right represents hosts with lower primary sampling rates (typically due to higher traffic load), which results in lower resolution for small packet counts.
- Values: The percentages represent the estimated accuracy with 95% confidence.
    - *Example*: An accuracy of 93% means the reported count is expected to be within roughly ±7% of the true value, with 95% confidence.
    - *Note*: These figures assume the secondary sampling rate is set to 100%. If your configured secondary rate is lower, refer to the adjustment instructions.
    - *Note*: These figures apply to byte counts under the assumption that the variation in packet size is small (see Formulas for Statistical Error Analysis).
    - *Note*: The values for each percentile are accurate as of April 2025.

Adjusting for Secondary Sampling Rate < 100%

The accuracy figures presented in the table assume a secondary sampling rate of 100% (1.0). If you have configured a lower rate, you are collecting fewer samples, which effectively lowers the resolution of your data.

To estimate accuracy in this case, calculate the effective packet count before using the table:

$$Effective\ Count = Reported\ Packet\ Count \times Secondary\ Sampling\ Rate$$

- **Example:** You are analyzing logs that report 200,000 packets, and your secondary sampling rate is 50% (0.5).
    - *Calculation*: $200,000 \times 0.5 = 100,000$ effective packets.
    - *Lookup*: Use the row for 100,000 packets in the table (instead of 200,000) to find your accuracy.

Example Scenario

Consider a scenario where you are analyzing an aggregation of flows that sums to 100 million reported packets (with a 100% secondary sampling rate). If your instance is running on a

heavily loaded host (99.9th percentile):

- **Accuracy:** The table shows an estimated packet count accuracy of 99%. This means you can be 95% confident that the reported count is within approximately ±1-2% of the true value.
- **Byte Counts:** Byte count accuracy is similar, provided the packet sizes are consistent (see Formulas for Statistical Error Analysis).
- **Secondary Sampling Impact:** If your secondary sampling rate is 50% (0.5) instead of 100%, the effective packet count is reduced to 50 million packets. You would then check the accuracy for 50 million packets (interpolating between the 10M and 100M rows).

Improving Accuracy

If you need higher accuracy from your flow log data, analyze data derived from a larger number of logged flow records. Since you cannot control the primary sampling rate, your main levers are:

1. **Increase the Secondary Sampling Rate:** Set the user-configurable secondary sampling rate closer to 1.0 (100%). This logs more of the flows that pass the primary sampling stage.
    - *Note:* This increases the volume of logs generated and associated costs.
2. **Increase the Time Window:** Analyzing traffic over longer periods (e.g., hours instead of minutes) naturally includes more sampled packets, significantly improving the accuracy of the aggregated counts.
3. **Aggregate Flows (broaden the scope):** Instead of looking at highly specific flows (e.g., a unique 5-tuple connection), aggregate traffic across broader categories (e.g., all traffic between two subnets). Summing these records increases the total packet count in your analysis, which directly improves statistical significance.
    - *Rule of Thumb:* The fewer filters you apply (i.e., the larger the set of logs you analyze), the higher the accuracy of the result.

Formulas for Statistical Error Analysis

For users interested in the statistical basis for the accuracy estimations, the approximate relative error percentages can be calculated using the following formulas, based on a 95% confidence level (Z-score ≈ 1.96). The accuracy percentages shown in the table relate to these error percentages via:

$$Accuracy\,\% = 100\% - Error\,\%.$$

These formulas are based on standard statistical methods for confidence intervals, discussed

in resources such as:

- Choi, Baek-Young et al. “Adaptive random sampling for traffic load measurement.” IEEE International Conference on Communications, 2003. ICC '03. 3 (2003): 1552-1556 vol. 3
- https://sflow.org/packetSamplingBasics

Packet Count Error

The relative error for the packet count estimation primarily depends on the number of sampled packets.

$$Packet\ Count\ Error\ \% \approx 196\ \times \sqrt{\frac{1}{c}}$$

Where:

- c = Number of sampled packets (i.e., number of packets * effective sampling rate).

Byte Count Error

The relative error for the byte count estimation depends on the number of sampled packets (c) and the distribution of packet sizes, specifically the coefficient of variation (CV).

$$Byte\ Count\ Error\ \%\ \approx\ 196\ \times \sqrt{CV^2 + \frac{1}{c}}$$

Where:

- c = Number of sampled packets.
- CV = Coefficient of variation for the packet sizes ($\frac{\sigma}{\mu}$).
- σ = Standard deviation of the packet sizes.
- μ = Mean packet size.

Table Assumption

The accuracy percentages shown in the table are derived from these formulas, making the simplifying assumption that the CV for the packet sizes is small compared to the inverse of the number of sampled packets. This occurs in traffic patterns where the standard deviation of the packet sizes is significantly smaller than the mean packet size, which is a common case for "well behaved" flows. Under this assumption, the formula for byte count error becomes similar to the packet count error formula.

If the actual traffic has a significantly different packet size distribution (e.g., mostly very small packets with a few very large ones, leading to a large CV), then the accuracy for the byte count

will differ from the accuracy derived purely from the packet count error. Specifically, a higher CV leads to lower byte count accuracy for the same number of sampled packets.

Conclusion

VPC Flow Logs provides important monitoring data. However, the inherent two-stage sampling means byte and packet counts are estimates. The non-configurable primary sampling sets a baseline based on system load, and the user-configurable secondary sampling further reduces data volume. Accuracy depends on the number of captured samples, which is influenced by total traffic volume and the effective sampling rate. The provided table illustrates estimated accuracy percentages (with 95% confidence) based on the primary rate percentiles, assuming 100% secondary sampling and a small coefficient of variation compared to the inverse of the number of sampled packets. Adjust your interpretation based on your configured secondary rate and knowledge of your traffic's packet size distribution. To improve accuracy, you can increase the secondary sampling rate, analyze data over longer time windows, or aggregate flows. Keep in mind the accuracy limitations discussed in the preceding sections, especially with low-volume flows or low effective sampling rates.