Подсчет трафика

Подсчет трафика простым суммированием значений поля Packet_size для ip адресов не является корректным, так как сбор статистики трафика в формате sFlow предполагает отправку на анализ одного из каждых n (Sampling rate) пакетов, проходящих по сети. Т.е. просуммировав размеры пакетов, можно получить сумму пакетов из выборки.

Согласно теоретическому материалу, представленному в http://www.sflow.org/packetSamplingBasics/index.htm, можно сделать вывод о том, что процентное соотношение трафика для данного ip относительно всего проанализированного будет таким же, как и процентное соотношение трафика для данного ip относительно всего, прошедшего по сети.

Пусть N – общее число пакетов, переданных через маршрутизатор, n – размер выборки, c_{ip} – число пакетов в выборке, переданных через данный ip, тогда общее число пакетов Nc, переданных через данных ip, будет равно $N_c = \frac{c_{ip}}{n} N$. Исходя из структуры sFlow $N = n\,sampling\_rate.$

Средний размер пакета для данного ip S_{ip} будет равен $\bar{S}_{ip} = \frac{\sum_{ip} S_{ip}}{c_{ip}}$. Таким образом, сумма трафика для данного ip будет равна $S = \bar{S}_{ip} N_c = \frac{\sum_{ip} S_{ip}}{c_{ip}} \frac{c_{ip}}{n} n\,sampling\_rate = \sum_{ip} S_{ip}\,sampling\_rate$.

Кроме предоставленного тестового набора данных, проводилось тестирование на в 2 раза большем наборе, превышающим по объему размер оперативной памяти.

Эффективность параллельной реализации

Тестирование проводилось на ПК с процессором Intel i7 (4 ядра, 8 потоков) и 4 гб оперативной памяти.

Небольшое исследование эффективности распараллеливания показало, что эффективность параллельной версии при распараллеливании на 2 потока около 0.9, на 4 – около 0.7. Подобное поведение программы, вероятно, связано с тем, что читается значительных объем данных, который затем фильтруется, и арифметические операции проводятся в небольшом количестве и над значительно меньшим объемом данных.