Network Traffic Capture & Aggregation: Why buffer size is crucial

In relation to network traffic capture, buffering is the process of an aggregation tap storing incoming network packets, usually in memory, mainly for the purpose of absorbing short-term bandwidth oversubscription of the output port(s) by the input port(s), before egressing them.

Three main ways to capture Ethernet traffic

There are three ways in which you can capture Ethernet traffic:he test was repeated for a total of four times.

1. **Tapping the link:** Optical taps are widely used, however, due to their enforcement of fibre over copper media, the limitation of only being able to tap the link they are physically plugged in to and the additional rack space they take up, Layer 1 switches are increasingly becoming the enterprise standard for tapping. This is due to their higher tap port density and their ability to be dynamically re-configured to tap any port connected to/through them. Arista devices even offer the ability to provide Ethernet statistics on every port.

2. **Mirroring on the switch:** Switch ports can often be configured to mirror traffic to a dedicated egress port for connection to a capture device; some switches can be configured to mirror a single or multiple incoming or outgoing link(s) to a dedicated egress port for off-box capture/analysis. This is typically referred to as a SPAN port.

3. **Packet capture on a host:** On a Linux host, the libpcap interface provides the ability to request that the kernel replicate and deliver raw frames entering or leaving the host to an application running locally.

Each of the above capture methods has its pros and cons , however in all cases, the ultimate goal is to to capture traffic without having to drop frames. One key point above worth reinforcing is that a 10GbE link is bidirectional and can carry 20 Gbps of traffic. Capture ports are inherently unidirectional, so for complete bidirectional capture two capture ports are required per link.

Why buffering is useful

The vast majority of network links are not running at anything near their capacity. Less than 10% average link utilisation is very common. Given that ports on capture devices can cost tens of thousands of dollars each, it really does not make sense to have to dedicate them to links running at a fraction of their capacity. Independently of the type of capture device ultimately used they can be combined, or aggregated: multiple links with low utilisation feeding into a smaller number of capture ports.

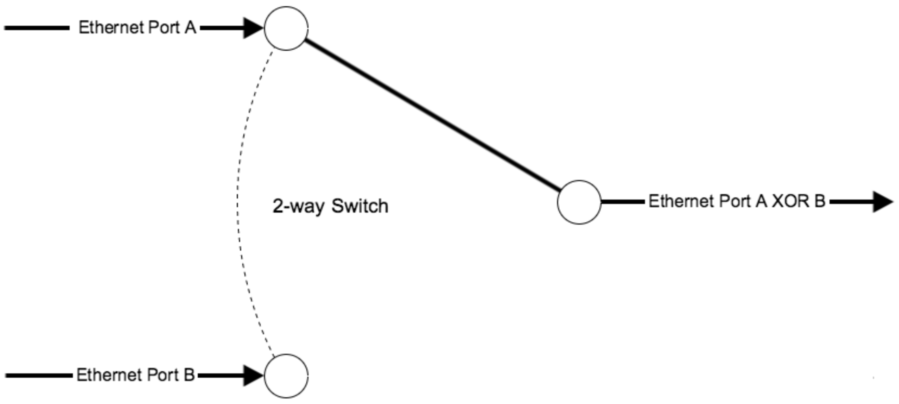
This is the role of the aggregation tap. Without buffering the instant aggregate ingress bandwidth cannot exceed the bandwidth of the egress port or packets will be dropped. This is because there is nowhere to store even a single packet if two packets arrive simultaneously.

By providing the ability to buffer incoming packets arriving on multiple ports simultaneously all packets are queued and available for capture as long as the buffer does not overflow (the aggregate ingress data exceeds the egress buffer queue for too long).

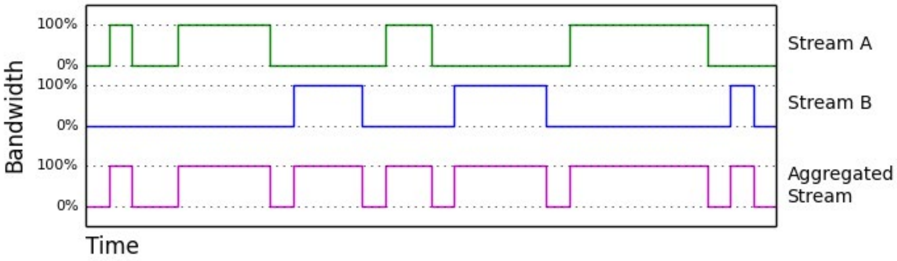
Aggregating ports without a buffer

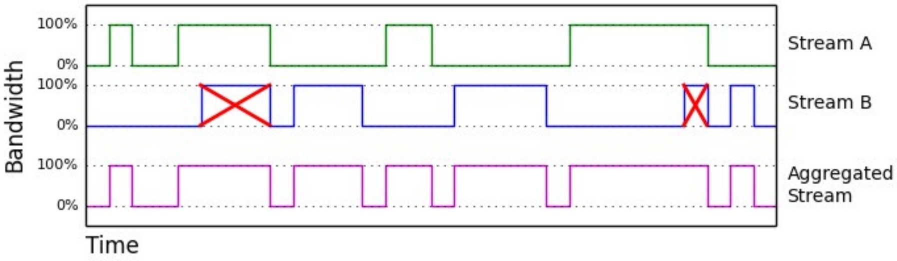
The following diagram represents a situation where two Ethernet input streams are aggregated into a single stream with no buffering. The aggregated stream can only be switched between one or the other input stream at any time.

Any time packets arrive on both input streams simultaneously, one of the packets will have to be dropped.



The below two diagrams show a contrived scenario when packets are interleaved successfully as packets never arrive simultaneously and the real-world scenario where they do and the resulting loss. In the second diagram two of Stream B's packets coincide with Stream A's packets and so must be dropped:





Managing oversubscription ratios

When aggregating captured traffic using an aggregation switch, it is worthwhile taking the time to understand the trade-off between the ratio of capture ports per egress port. If too many capture ports are being aggregated into a single egress port and both capture and egress ports are the same speed e.g. 1 GbE or 10 GbE, and if the aggregate instantaneous bandwidth of the capture ports exceeds the bandwidth of the egress port, the delta is either buffered or dropped.

To mitigate the problem of dropped packets, either larger buffers can be used or the ratio of capture ports per egress needs to be adjusted downward. Even a 2:1 aggregation ratio can cause a problem if the sum of the instantaneous traffic bandwidth on each capture port exceeds the bandwidth of the output port.

Buffering and line rate

The term line rate is readily understood to mean that a given link is saturated with data. In the case of Ethernet this means frames of any valid sizes separated by an average of 12 bytes of Inter-Frame Gap (IFG). It is a term that relates to bandwidth. It is important to keep in mind that bandwidth is generally expressed in units of bits per second or bps. So if a 10 GbE link is at 50% bandwidth utilisation over a second, 625 MB is being transferred in that second (actually somewhat less when the IFG is taken into account).

What is not apparent however is the distribution of that data over the second:

- Was all the data sent in the first half of the second and the line was idle for the second half?
- Were the frames spread out so the average utilisation was consistently 50% throughout the second?
- Was the frame distribution throughout the second random but with an average bandwidth utilisation of 50%?

In fact, the instantaneous bandwidth utilisation of an Ethernet link can really either be 100% (data) or 0% (idle). Though Ethernet bandwidth is usually quoted over a second, when it comes to buffering, the traffic patterns within that second can be extremely important, especially the length of the periods where the link is at line rate (100%) i.e. back-to-back frames.

The importance of buffer size

At this point, it is worth looking at what constitutes instantaneous. Network bandwidth is usually expressed in kilo-, mega- or giga- bits per second (Kbps, Mbps, Gbps) and memory in Bytes with the same prefixes being applied. For any given amount of buffer memory, bandwidth can be converted into the amount of time it takes to fill it. For example, assuming we have a typical aggregation switch with 10 MB of buffer and two 10 GbE capture ports into a single 10GbE egress port, what is the longest line rate burst that can be absorbed without dropping any packets?

The formula is:

$$\text{Time to fill Buffer} = \frac{\text{Buffer Size (MB)}}{\left(\text{Number of Capture Ports} \times \frac{\text{Capture Port Bandwidth (Gbps)}}{8\text{ (bits)}}\right) - \frac{\text{Egress Port Bandwidth (Gbps)}}{8\text{ (bits)}}}$$

N.B. As the buffer can drain to the egress port at 10 Gbps, the buffer will only grow at 10 Gbps rather than the incoming 20 Gbps.

Filling it in:

$$\frac{10\text{ (MB)}}{\left(2 \times \frac{10\text{ (Gbps)}}{8\text{ (bits)}}\right) - \frac{10\text{ (Gbps)}}{8\text{ (bits)}}} = \frac{10\text{ (MB)}}{1.25\text{ (GB/s)}} = 0.008\text{ s or }8\text{ ms}$$

If we increase the aggregation ratio to a more realistic 10:1, how does this affect our ability to absorb bursts?

$$\frac{10\ (\mathrm{MB})}{\left(10 \times \frac{10\ (\mathrm{Gbps})}{8\ (\mathrm{bits})}\right) - \frac{10\ (\mathrm{Gbps})}{8\ (\mathrm{bits})}} = \frac{10\ (\mathrm{MB})}{11.25\ (\mathrm{GB/s})} = 0.0009\ \mathrm{s\ or}\ 900\mu\mathrm{s}$$

In summary, with 2:1 oversubscription, a line rate burst on both capture ports longer than 8 ms would cause the buffer to overflow and almost certainly result in lost capture packets. At a 10:1 oversubscription, a line rate burst on all ten capture ports longer than just under one ms would have the same impact. It is important to keep in mind that in these worked examples, the aggregate per second line bandwidth utilisation could be as low as 1% and the buffer would still overflow. All it takes for a short burst of traffic (a microburst) on multiple otherwise idle aggregated Ethernet links at the same time.

The Solution: Using Deep Buffers

Deep buffering allows many more ports to be aggregated into a single link per device and hence fewer aggregation taps. These deep buffers permit aggregation at significantly higher link utilisation and/or higher aggregation ratios. The buffers can hold multiple seconds of captured packets at 10 GbE coming in across multiple links. In short, deep buffers can smooth out much longer bursts and sustain far longer periods of oversubscription without dropping packets.

An Overview of Arista Deep Buffering

Arista's MetaWatch application turns Arista's FPGA enabled K and L series devices into extremely powerful aggregation taps. The devices have 32, 48, or 96 ports and capture up to 48 ports into either a 8GB or 32GB of buffer. To put this into perspective, this buffer is 1000 times larger than that of most ASIC-based aggregation taps. Taking our earlier worked examples and adjusting them for 8 GB of buffer:

2 x 10 GbE capture:

$$\frac{8\ 000\ (\mathrm{MB})}{\left(2 \times \frac{10\ (\mathrm{Gbps})}{8\ (\mathrm{bits})}\right) - \frac{10\ (\mathrm{Gbps})}{8\ (\mathrm{bits})}} = \frac{8\ 000\ (\mathrm{MB})}{1.25\ (\mathrm{GB/s})} = 6.4\ \mathrm{s}$$

10 x 10 GbE capture:

$$\frac{8\ 000\ (\mathrm{MB})}{\left(10 \times \frac{10\ (\mathrm{Gbps})}{8\ (\mathrm{bits})}\right) - \frac{10\ (\mathrm{Gbps})}{8\ (\mathrm{bits})}} = \frac{8\ 000\ (\mathrm{MB})}{11.25\ (\mathrm{GB/s})} = 0.7\ \mathrm{s}$$

These deep buffers permit aggregation at significantly higher link utilisation and/or higher aggregation ratios. The buffers can hold multiple seconds of captured packets at 10 GbE coming in across multiple links. In short, deep buffers can smooth out much longer bursts and weather far longer periods where egress bandwidth is oversubscribed without dropping packets.

Scale Capture/Analytics for average rather than peak aggregated rates

MetaWatch also implements flow control to capture devices via IEEE 802.3x PAUSE on its aggregated egress ports. When receiving them the port will back pressure the buffer to reduce the egress bandwidth. Any device consuming the aggregated stream supporting Ethernet Flow Control can take advantage of this useful feature.

For example, the ixgbe Linux driver for Intel 10 GbE adapters has Ethernet Flow Control enabled by default and the majority of Ethernet Switches support it. Capture and Analytics devices that can generate PAUSE Frames can therefore be scaled down in performance and hence, cost, to take advantage of the deep upstream buffer.

In Summary

- Buffers are required in any realistic aggregation scenario
- The size of the aggregation tap's buffer dictates how many ports can be aggregated
- The ASICs in most aggregation taps only have around 10 MB of buffer allowing line rate bursts at 10 GbE on two ports simultaneously of only single digit milliseconds
- MetaWatch has buffers of 8GB or 32GB allowing far greater aggregation ratios to be configured with a vastly reduced likelihood of packet loss
- MetaWatch can also moderate its aggregated egress rate in response to Ethernet Flow Control configured on the consuming device allowing a less costly device to be used

Santa Clara—Corporate Headquarters
5453 Great America Parkway,
Santa Clara, CA 95054

Phone: +1-408-547-5500
Fax: +1-408-538-8920
Email: info@arista.com

Ireland—International Headquarters
3130 Atlantic Avenue
Westpark Business Campus
Shannon, Co. Clare
Ireland

Vancouver—R&D Office
9200 Glenlyon Pkwy, Unit 300
Burnaby, British Columbia
Canada V5J 5J8

San Francisco—R&D and Sales Office 1390
Market Street, Suite 800
San Francisco, CA 94102

India—R&D Office
Global Tech Park, Tower A & B, 11th Floor
Marathahalli Outer Ring Road
Devarabeesanahalli Village, Varthur Hobli
Bangalore, India 560103

Singapore—APAC Administrative Office
9 Temasek Boulevard
#29-01, Suntec Tower Two
Singapore 038989

Nashua—R&D Office
10 Tara Boulevard
Nashua, NH 03062