



**Network latency: avoid paying a tax on time** 9

In network security we are all too aware of our own latency tax – the price we pay for keeping our networks protected. Over and above the costs of equipment purchase, upkeep and related skills, this extra 'tax' is priced in direct proportion to seconds, milliseconds, even microseconds lost. So intelligently integrated, hardware-accelerated systems could be the solution to the trade-off between security and performance, says Patrice Perche of Fortinet.



## FEATURE

# Network latency: avoid paying a tax on time



Patrice Perche

Patrice Perche, Fortinet

Running late for a flight, the traveller can take numerous shortcuts. Pay extra to go in the short-stay car park, convert currency once you've landed, or even avoid baggage check-in altogether by stuffing as many essentials as possible into hand luggage. However the one stage that cannot be rushed is security. It takes as long as it takes, and there is nothing you can do about it.

Few would disagree that airport security staff should be free to take as long as they need. This is an unavoidable and entirely necessary delay, and that latency is the tax you pay for being protected against all manner of nasty threats.

In network security, we are all too aware of our own latency tax – the price we pay for keeping our networks protected. Over and above the costs of equipment purchase, upkeep and related skills, this extra 'tax' is priced in direct proportion to seconds, milliseconds, even microseconds lost.

This is because the very functions of network security – including stateful or deep inspection firewalls, Intrusion Prevention Systems (IPSs), anti-virus (AV), Virtual Private Networks (VPNs), anti-spam and web content filtering – inject the same sort of unavoidable and necessary delay that exists in every airport. So how does this tax get paid?

Make no mistake, this tax is a monetary sum that subtracts itself from every organisation in every sector. It is collected in two ways. First, by the efforts of network managers to over-compensate the bandwidth of critical network segments by investing large sums 'sticking it to the stovepipe' with upgrades to physical and switching infrastructure. Second, through the direct financial implications of running a network that isn't as fast as it should and could be.

## Security at 100Gbps and beyond

Big investments are going into network upgrades, with demand for high-end 10G, 40G and even 100G ports outstripping the rest of the network equipment market. In May 2010, Infonetics Research confirmed growth in the sales of 40G IP edge routers had reached 125% during 2009, and forecast the overall demand for 10G/40G/100G ports to grow ten-fold before 2014. All of this, the research notes, is in response to growing levels of traffic. Without these investments, traffic would get squeezed and delays reach unacceptable levels.

More bandwidth is going to mean more threats. Do the math, starting with the percentage of today's traffic that contains various types of malware, including viruses, spam, inappropriate content and other policy violations. Then consider that only 4% of your email traffic is clean of cyber-criminal intent, and the challenge becomes clear. A constant percentage equates to a larger amount as the total quantity increases.

***"Any investment in driving up security performance is going to yield substantial returns"***

More security means more latency. Network security is business-critical, but with the firewall function alone typically accounting for up to 20% of latency in

the network, any investment in driving up security performance is going to yield substantial returns.

Security has been cited as the single biggest culprit of latency inside corporate networks. It stands to reason – all that traffic needs to be 'stopped and verified' as thoroughly as possible. Returning to the airport security analogy, parallels with network security continue as both respond to increased traffic demands by building extra capacity. This creates the need for extra security too, which will mean a longer wait for everything to pass through safely.

To deal with the increase in traffic, the choice is clear: make the security zone bigger with more X-ray machines and immigration personnel, or make the security zone faster with a more responsive, smarter approach. Where network security is concerned, 'bigger' and 'faster' are not necessarily the same thing.

***"Latency stays expensive, right up there with compliance as one of the most draining challenges your IT operation has to face"***

## Time is money

Organisations are finding that migrating toward 10G/40G and 100G network infrastructures is being made easier by the fact that, like processing power and memory, bandwidth becomes cheaper over time. Latency on the other hand stays expensive, right up there with compliance as one of the most draining challenges your IT operation has to face. In fact, the only other thing more expensive is the disastrous fallout from a significant security breach. The key to understanding the



## FEATURE

tangible cost of latency often lies in the notion of 'first mover advantage'.

To switch analogies, in Formula One Grand Prix racing, competition is so intense that extremely fast cars driven by some of world's best drivers routinely finish last. Operating on millisecond margins, all the race teams know that the slightest delay will cost them everything.

Direct comparisons can be drawn between F1 and the master-mathematician hedge fund managers and algorithmic-traders working in the key financial trading centres. Throughout the financial services sector, network security is critical to effective data management, business continuity and compliance. On the other hand, network speed is critical to maintaining competitive advantage.

***"Throughout the financial services sector, network security is critical to effective data management, business continuity and compliance. Network speed is critical to maintaining competitive advantage"***

At best, in a distributed or high-volume data environment, compromised network performance can cause deterioration in customer service and increased user frustration. At worst, in an electronic trading situation, a few milliseconds of latency in the network can conceivably cost millions of pounds.

At the major stock exchanges, the ordinarily obscure and technically nuanced topic of latency is one of the most contentious issues around. Some technologically advanced traders have finely honed both their incoming market data feeds (systems like Reuters and Bloomberg that relay the latest price information) and their trading execution systems in order to gain first mover advantage commonly referred to as 'latency arbitrage'. Some traders even target the physical Layer 0/1 infrastructure for latency gains, moving operations physically as close as possible to the exchange in order to minimise cable lengths. Get to the trade first, of course, and you can detrimentally influence the price up or down for the next trader after you.

For ordinary retail investors, the practice is regarded somewhere between 'unfair' and 'illegal'. As recently as June 2010, regulators at the US Securities and Exchange Commission resolved to consider how to institute a level playing field, with some speculating this could include mandating that all fibre connections into exchanges be 100m long for everybody.

### Quest for speed

The quest for unrelenting speed has also given rise to the development of highly accurate network monitoring and packet time-stamping technology that is able to sit on large network segments to measure and report on live and historic levels of latency to within hundreds of picoseconds accuracy (a picosecond = one trillionth of a second). Much like the Hawk-Eye system used in professional tennis, this technology is used to replay events to protocol and application layer granularity. Hypothetically, it can deduce the most advantageous level of prevailing latency at multiple global stock exchanges, enabling traders to determine which exchange to make trades at, on a given day or moment. Market data providers are also getting into the act, with products such as the Reuters Latency Monitor offering financial institutions a measurable safeguard against the risk of consuming critical market information that is fractions of a second out of date.

***"In the quest to achieve minimal latency, attention has turned to re-engineering the apparently intractable trade-off between the best possible security and the best possible performance"***

While some key industry sectors are more sensitive than others about latency of microsecond magnitude, it is inevitable that the issue will intensify throughout all market sectors as organisations come to depend unreservedly upon real-time data communications. Consider a telco, who worries about the instantaneous effectiveness of its premium mobile services, a cloud computing provider (or customer for that matter) managing the integrity of SLAs, a retailer process-

ing card transactions, or a manufacturer orchestrating its supply chain with knife-edge precision. Ask them what they think about 'a few milliseconds' of latency, and stand back as their faces turn grey. Consequently, in the quest to achieve minimal latency, attention has turned to re-engineering the apparently intractable trade-off between the best possible security and the best possible performance.

CSOs and CIOs understand that the more security elements on their network, and the more traffic in need of interrogation, the longer it takes – so much so that network planners who want to achieve true 10G/40G or 100G capability need to be planning many capacity levels higher, in order to allow for the amount of network degradation.

Our technology vision is for network security to impose no more latency into the network than a switch, whether it is handling unicast or multicast feeds. There can be no shortcuts though; there is no merit in a 'firewall-lite' approach or re-spun Deep Packet Inspection (DPI) play. The challenge that the industry faces is in confronting the latency that threatens high-capacity networks with platforms that deliver advanced, integrated, multi-layered security functions.

### Network threats increase in complexity

Historically, very few organisations have geared security into their bandwidth expansion plans from the get-go as, compared to the ever-present risk of a security breach, latency was always considered the lesser of two evils. However, whether capacity planning exercises lead to a big bang approach to 10G and beyond, or a more staged approach is taken, high-speed network planning and high-performance security should go hand in hand.

Pre-conceived notions of what high-performance security looks like suggest that further education on scalable security infrastructures is needed. Just consider how individual 'point' security products multiply and silo enormous duplicated frequencies of individual 'stop and search' inspections on each and every packet that



## FEATURE

traverses your network. In the face of those pressures, can the network achieve the throughput performance levels and integrated blended response needed for the onset of 10G and beyond?

Increasingly, large enterprises are sharing the view that running an alignment of point products creates obstacles to teamwork. However, many integrated network security approaches also fall foul of this problem.

### ***“Businesses that deploy one comprehensive inspection of threats upon one accelerated platform will safeguard themselves from both known and unknown attacks”***

At the technical level, the transit of IP traffic going into one end of the network gateway security infrastructure and out the other is going to involve a degree of packet disassembly and reassembly. This is one of the core principles of security performance degradation. Going through the same practice of disassembly/reassembly, to check for a different security problem each time, results in lots of redundant processing – that is, lots of wasted throughput capacity and lots of wasted time. But doesn't an integrated network security approach also suffer from this? Well, not necessarily.

### **Interoperability question**

As well as repeatedly queuing for packet assembly/disassembly, IP traffic typically has to deal with security functions based upon multiple, disparate source codes. Herein lies the interoperability question inherent in what can only be described as ‘cobbled-together’ rather than properly integrated security architectures. A lot of very high-end network security hardware is marketed as being able to perform multiple, integrated security features – often simultaneously. Few stack up to the billing, and many provide no more advantage than a point solution architecture in terms of overall latency, risks of security ineffectiveness and complete confusion about the root of the problem should any failures arise.

The shortcomings don't end there either. Any security system is only as good as the threat intelligence that constantly updates it. A given ‘integrated security’ vendor might do its own AV research, but probably contracts out its IPS or malware filtering. Where does that leave your network?

The wrong technology approach could serve to simply scatter latency everywhere. This is because the threats themselves aren't just becoming more present, they are becoming more blended. To ensure your security infrastructure is reacting to blended attacks with a blended response at the optimum performance level, data cannot simply be interrogated IP packet by IP packet. A blended response device must be capable not only of detecting these blended threats but of doing so on the first pass; multiple handoffs in software will only help in exhausting capacity. Businesses that deploy one comprehensive inspection of threats upon one accelerated platform will safeguard themselves from both known and unknown attacks, and reap the rewards.

### **Complete content protection**

In security circles, much has been made of the breakthrough innovations underpinning so-called Next-Generation Firewall (NGFW) technology. Many seem to claim it heralds a new dawn of network protection of a kind never seen before.

One of the most touted capabilities within NGFW products is an application visibility and control capability. This elevates the inspection capabilities of a firewall from layer 4 to the application layer (layer 7), enabling threats to be identified according to application content. Since many application providers are moving to a web-based delivery model, and many prevalent threats are carried by legitimate business applications, obviously this capability is important, but not especially innovative. In fact, NGFW is little more than a logical subset of the latest developments in Unified Threat Management (UTM) technology.

The reality of security today is that deeper inspection of all content is essential, not just the application allow/deny approach offered by NGFW devices. For example, to protect against the recent Conficker virus, an enterprise would have needed a firewall, web filtering, network AV, IPS, anti-spam and host-based AV in addition to an efficient automatic updating mechanism for all of these devices. This is what integrated security (or UTM) solutions do, and they are really just a superset of NGFW products. The application policy capabilities are a feature, but the technologies are more focused on scrutinising the content of legitimate applications as well as blocking unwanted applications to ensure that threats are passed via application communications.

### ***“Truly integrated high-performance security can only be derived from a truly uniform architecture approach”***

### **Content protection**

So, complete content protection is enabled when application control and application security are combined to integrate content-based security technologies into the firewall, in order to identify threats within trusted application content. Returning once again to latency concerns, this approach surely represents the most efficient way of applying the most effective security to critical data.

Yet one person's ‘integration’ is another person's ‘amalgamation’. Only one will mitigate latency. Truly integrated high-performance security can only be derived from a truly uniform architecture approach, in which each security function has been developed on the same source code in order to optimise security performance and eliminate redundant traffic processing. How? Through the use of specialised hardware based on Application-Specific Integrated Circuit (ASIC) technology to accelerate the security inspection process.

The development and application of specialised hardware security-specific ASIC processors to accelerate



## FEATURE

integrated security has demolished institutionalised thinking around this security approach. Those harbouring lingering concerns about the applicability of integrated network security within their high-speed, real-time critical networks should only have cause to worry if they follow the wrong kind of approach.

It stands to reason that intelligently integrated, hardware-accelerated security can

make the real difference to secure business performance as organisations find themselves competing in both the security arms race and the battle of the bandwidth.

### About the author

*Patrice Perche is vice president, international sales and support at Fortinet. He has nearly two decades of experience in the IT security industry. In his current position, he is responsible for managing Fortinet's*

*international sales and support operations in EMEA and Asia. Perche has extensive experience in launching innovative security technologies into new markets and working closely with channel partners. Prior to joining Fortinet, he was co-founder and CEO of Risc Group, a France-based company focusing on IT security with pan-European operations. Perche graduated from Insa Lyon and holds a masters degree in computer engineering.*