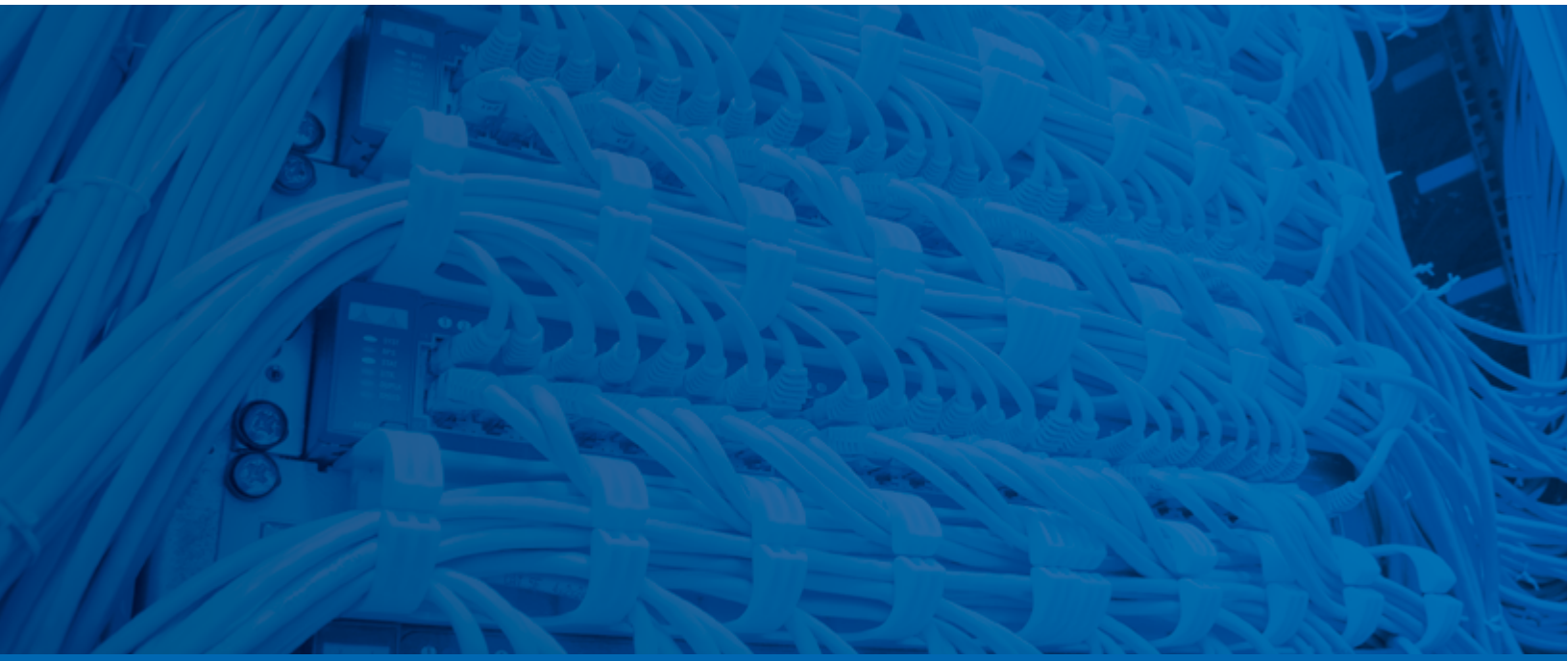


Big Data, Big Challenges

How the data deluge is great for business but an enormous challenge for network operations

Sponsored by

APCON
Solutions for Networks

Enormous Opportunity Brings Unprecedented Challenges

Few people know that the term “Big Data,” a ubiquitous buzz-phrase now, can be traced back more than 20 years to John Mashey, chief scientist at Silicon Graphics, a cutting edge company that created special effects for Hollywood productions and spy video technologies for the government.

Mashey wrote no academic papers using the term, but he made lots of presentations to potential Silicon Graphics investors and customers using slides in which he repeatedly used the phrase in a context quite similar to the context in which it is used and understood today: As a collective noun for the process of turning exponential increases in data volume into actionable information.

Instead of becoming overwhelmed by the unprecedented amount of data that many in those days worried would create a global logjam in computing, Mashey foresaw a world in which businesses and other organizations could harness all that information to gain better insight into customer behavior.

Doing that, he said, would improve productivity, cut costs, improve sales and financial performance, and make life better for billions of people. Businesses actually would be able to anticipate the future with great clarity and react in time to meet it.

Big Data's Prophet Didn't Predict Ripple Effects

Today, Big Data generally can be thought of as "the collection and analysis of data and metadata sets in amounts so enormous that they far exceed the capability of commonly used software tools to capture, curate, manage and process in real time or something very close to real time."

It turns out Mashey was a prophet. His term stuck and entered the common vocabulary. More importantly, so did his vision for turning gargantuan amounts of raw data into actionable information. Now, Big Data as we know it represents one of the big keys to the future, with previously unimaginable opportunities.

However, the early forecasts didn't include previously inconceivable challenges that Big Data has presented for network operations and security managers. Many companies are only now beginning to grapple with the implications while struggling to keep up with the competition.

Hundreds of practical business problems have sprung up:

- How to protect the information from opportunistic crooks?
- How to control network management costs without losing total visibility?
- How to analyze all that information without delaying it for even a fraction of a microsecond?
- How to reliably monitor all that information without hiring an army of engineers?
- How to filter for only the relevant information in a sea of data?

Network Monitoring Solutions Emerge as a Key Component

As the concept of Big Data comes to fruition, there is a direct correlation between enterprises' ability to take advantage and their adoption of state-of-the-art network monitoring solutions that help tame the sea of data, effectively making it more digestible for networks and network tools.

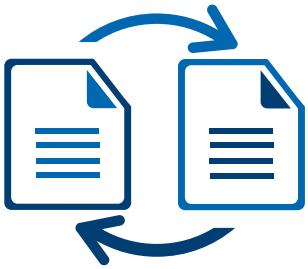
These solutions perform tasks such as packet deduplication, which culls the stream of duplicate information, which can be up to 40 percent of network traffic; or packet slicing, which increases the capacity of monitoring tools by stripping bits out of the packets that are unnecessary for monitoring. Such solutions increase security by stripping confidential details – credit card and social security numbers and other personal information – from the flow of data, and they help capture and store data from any number of sources for immediate or later review.

Mr. Mashey might not have envisioned 20-some years ago that technological solutions would play a major role in building an effective answer to each of the challenges presented by Big Data.

We'll address the three most obvious challenges that any business, government entity, institution or other enterprise faces when it decides to dive into the world of Big Data: The so-called "3V's" of velocity, volume and variety.

Velocity

Volume



Variety

These challenges all revolve around one word: "Big." Just how big is Big Data? And what does it mean for network operations and security managers charged with ensuring every bit and byte is properly captured, sorted, scanned, sampled, monitored, analyzed and stored without breaking the bank?

The Challenge of Velocity

The “big” in Big Data doesn’t even begin to describe the enormity of the subject.

An analyst named Doug Laney, now at the Gartner research firm, years ago came up with probably the best explanation for what Big Data does and how it can benefit any enterprise. He came up with the “3V’s” concept for getting one’s head around the enormity of challenges – and opportunities – surrounding Big Data today.

Laney’s first “V” is velocity, referring to the speed at which data is created, stored, analyzed and visualized.

In previous times, data typically were processed and transmitted in batches. Databases typically issued updates to users nightly or even weekly. That’s because ordinary computers and servers took significant amounts time and processing power to process and update all the computers connected to the database. During that time, access to the data being updated either was extremely limited or not available at all.

Big Data has changed all that. Data – especially if it’s used to advance business goals – must be collected, processed, analyzed, stored, retrieved and retransmitted in real-time. The rate at which all that happens in a Big Data computing environment boggles the mind.

For example: People upload about 100 hours of YouTube video every minute. In that same minute, they also send about 200 million emails, type nearly 300,000 Tweets and query Google about 2.5 million times. That doesn’t begin to tell the story of the velocity at which data are being generated by businesses, governments, militaries and markets.

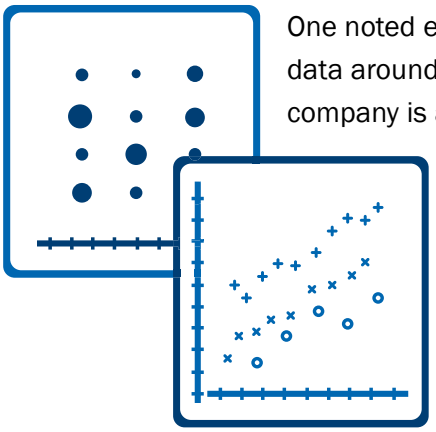
To keep up, companies that have had their toes in the Big Data waters for several years now are taking a deep dive by upgrading their networks from a 1G environment to 10G, 40G or even 100G processing environments to stay ahead of the velocity curve.

High Velocity = Data Overload

One extreme example of the need for speed can be found in the realm of science. At Europe's Large Hadron Collider, two high-energy particle beams are sent racing through a 17-mile circular tunnel so scientists can observe what happens when those particles smash into each other. The data numbers created are off the chart: Each of the nearly 150 million sensors in the tunnel record and transmit data about 40 million times per second. There are nearly 600 million collisions per second. The resulting Niagara of data is too much even for the best networks to process.

Still, the 0.001% of the sensor stream data that the physicists at the LHC choose to process amounts to 25 petabytes of data per year. After that sample data set goes through replication (so it can be processed through multiple analytical tools), the LHC collects and stores about 200 petabytes of data per year.

Today, enterprises are increasingly likely to generate, process, analyze or store extraordinarily large amounts of data at high speeds. Many are waking up to the value of monitoring hundreds upon hundreds of data points, and they are processing, storing and analyzing data literally at the speed of light.



One noted example is retail giant Wal-Mart, which constantly monitors transaction data around the world to spot and react to customer trends. For example, the company is able to monitor credit-card transactions for error codes that could indicate a problem with a card issuer's system even before that issuer notices it.

Even with the most powerful analytics tools available, real-time intelligence gathering is not possible without network throughput that can accommodate potentially hundreds of thousands of data streams in real time without losing visibility.

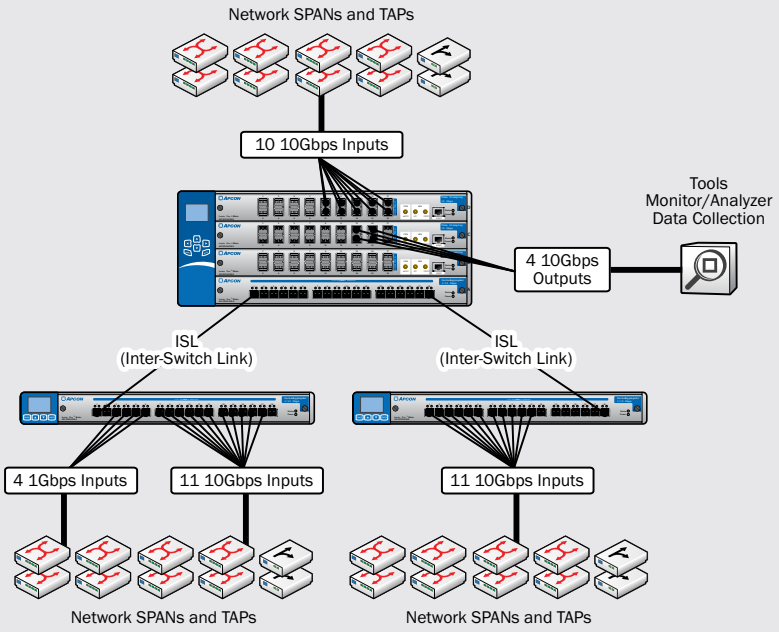
To compensate, network architects have turned to Ethernet switches that allow them to thin the data stream through deduplication or packet slicing, or that allows them to practice "trunking" to combine throughput of multiple connections.

An Example from Financial Markets

Finance is among the industries most affected by today's data speeds. At banks and brokerage houses, high-frequency trading requires – often by service-level agreement – the secure transmission of time-sensitive, proprietary information. Data moves so fast that its arrival is measured in the hundredths, thousandths or even billionths of a second.

A typical network monitoring implementation at a financial institution integrates the traditional functions of aggregation and load balancing with other important functions, from monitoring latency via time-stamping to recording data and validating trades.

Multiple TAP and SPAN inputs are aggregated and directed to monitoring and analysis tools. And, all of this happens at "line rate," the actual speed the data is flowing.



Firms deal with the velocity problem by turning to high-capacity network switching solutions that can handle speeds of 10G or more. Technology can also integrate the traditional functions of aggregation and load balancing – which keep high-dollar analysis tools from being overloaded – with the function of time stamping. Data packets are "stamped" as they pass through various points to measure latency and ensure compliance with SLAs. Through packet slicing, the switches can simultaneously eliminate sensitive bits of high-velocity information, not only enhancing the security of the transmission but also lowering the physical number of bytes being transmitted to keep from overloading analysis tools.

The various ways of attacking the first "V" also help control the "Big C" – costs. The expense of individual network monitoring tools can easily run to the six figures. Reducing the amount of high-velocity data sent to the tool – by any means – can help forestall the purchase of additional tools.

And that leads us to Laney's second "V" – volume.

The Second ‘V’: Volume

Naturally, data generation at the staggering speeds we’ve already discussed leads to not just mountains of data, but to whole Himalayas-sized mountain ranges of data being created and stored.

A decade ago, the data generated in the history of the world doubled every 40 months. Now data doubles in just over two years. Indeed, 90 percent of all data EVER created came to be in the past two years. And because the rate of data generation is accelerating, the volume growth is, too. By 2020 – there will be 50 times more data in existence than existed in 2011.

We could go on, but you get the idea. The volume already being generated, organized, analyzed, stored and transmitted today exceeds the comprehension of most non-technology professionals. So the ever-larger volumes expected to be created in the future makes it almost impossible to adequately describe using mere words.

Big Data is just that big.

How Large Is an Exabyte? Soon, We'll All Know

Today, there are estimated to be fewer than 200 petabyte-sized storage drives in existence, and only a handful of larger-still exabyte drives (1,000 petabytes, or 1 billion gigabytes) in operation. But within just few years there will be thousands of petabyte drives and hundreds of exabyte drives in use around the world.

To be sure, most businesses are unlikely to ever have such enormous data storage needs. But even small businesses that move into Big Data will need exponentially more storage than they ever imagined before.

To handle it all, even with the most modern and efficient software frameworks, companies need to seriously consider not only how much storage capacity they need, but also the network architecture, network switching solutions and tools they'll employ to get the most data management bang for their corporate bucks.

All that capability is necessary to balance the loads of new data coming in but also to deal with the duplications of that incoming data that must be made so that the data can be analyzed simultaneously, and in real-time, by a wide array of analytic tools.

How big is a Petabyte?

- If you counted all the bits in one petabyte at one bit per second, it'd take 285 million years.
- If you counted one byte per second, it would take 35.7 million years.
- It would take 223,000 DVDs (4.7Gb each) to hold 1Pb.
- It would take 746 million 3.5-inch high-density floppy discs (1.44Mb each) to hold one petabyte; 746 billion floppy discs weigh 13,422 tonnes (if each one weighs 18g). This is just under the size of two Type 45 destroyers, such as the newly built HMS Duncan, which has just left the Clyde
- Estimates of the number of cells in a human body vary, but most put the number at approaching 100 trillion, so if one bit is equivalent to a cell, then you'd get enough cells in a petabyte for 90 people – the rugby teams of the Six Nations.

That's how "data mining" works. It looks at often widely divergent strings of seemingly unrelated data and metadata sets to discover patterns and outliers. In identifying patterns, it gives managers insights on how to reduce costs, build better products, deliver services better, market products and services more effectively, and predict consumer behaviors in a complex, matrixed marketplace.

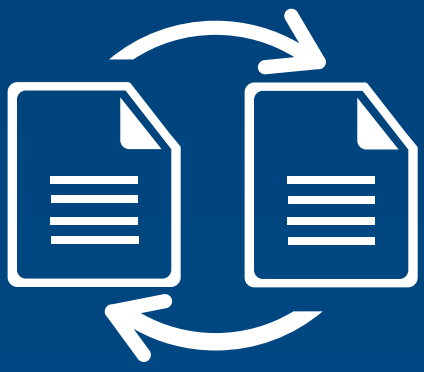
Before Big Data Can Work, Volumes of Information Must be Managed

But in order to reach actionable business decisions using Big Data approaches, the data must first be collected, organized, filtered, and otherwise manipulated for efficient analysis.

To manage that process, intelligent network monitoring switches aggregate multiple data streams copied from SPAN, Mirror, and Tap ports in many locations on a network. How many? Well, for security reasons alone, the best advice is to "listen in" on 100 percent of network traffic between all switches, routers and other points of access in your network.

The mass data stream is then filtered and passed on to network analyzers, recorders, and other data mining tools. Advanced features such as packet deduplication and packet slicing further refine the data stream for efficient analysis on information of any variety – which leads us to Laney's third "V."

The Challenge of Variety



Only 15 years ago, pretty much all data was structured. That is, it was searched for, processed, analyzed and stored by category. Today, however, data that companies are monitoring, generating, analyzing and storing is just as likely to be unstructured as structured.

Data is collected from a variety of monitoring points – as we said before, as many points as you can – and in every imaginable shape, from myriad people and things. It can be email, financial records, customer records, etc., in the form of text, video, audio or log files.

The varieties are endless, and the data enters the network without having been quantified or qualified in any way. It is truly raw data: Numbers seemingly without any meaningful context. The challenge, obviously, is how to make sense of it all.

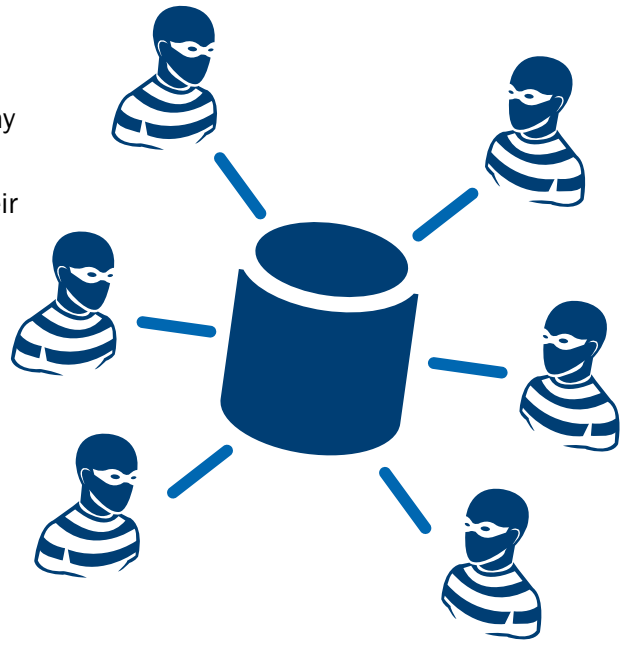
For businesses and other organizations working to discover meaningful – even profitable – patterns in all that unstructured data and getting it to the right places, the challenge is getting the right data and ONLY the right data to each tool in your inventory. That's the deal with variety – each category of data needs to get to one or more tools. Multi-stage filtering is the solution that can sort out this problem.

New techniques can simplify the process of filtering data packets as the data flows in and improve the data accuracy delivered to network analytics tools. Filtering processes, by sending only the necessary information to analytics tools, also help mitigate tool oversubscription and the subsequent problems of network congestion and packet loss.

The data aggregation capabilities built into state-of-the-art network monitoring solutions also weeds out redundant packets recorded at various monitoring points in your network so tools aren't tasked with analyzing the same information several – or dozens of – times.

By extension, those capabilities ultimately become a cost-saver by significantly extending the life of monitoring tools while retaining 100-percent visibility and efficiency.

Data Variety Makes Security a Taller Order

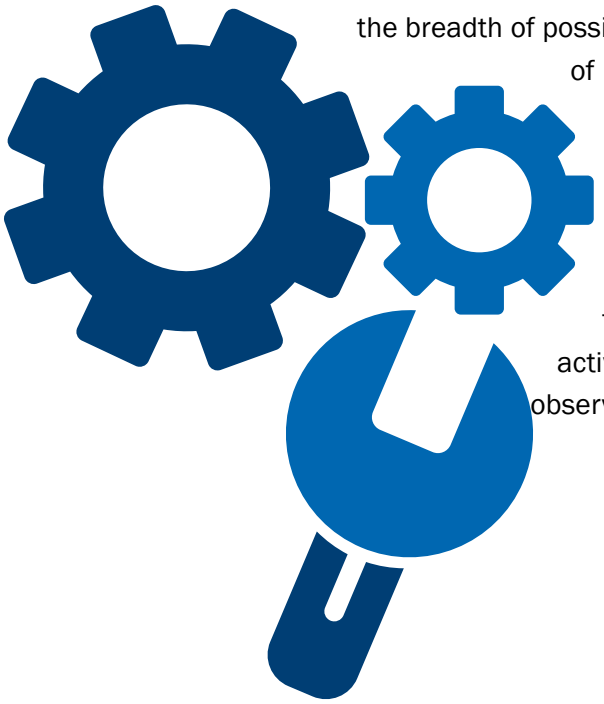


Data variety also makes network security a more difficult job. For one, the bad guys come at you from all angles, and in many forms. Sometimes, they're inside your network before you can fend them off. Network security managers will tell you that their job requires both the ability to defend against attacks and to forensically determine the scope of attacks after they are detected.

Operators need Intrusion Detection Systems (IDS) that provide the ability to analyze all activity coherently, even when data is streaming in from a variety of monitoring points. After all, if you're not scrutinizing all the data – from a variety of sources and in all forms – then you don't have visibility.

Today's network-monitoring technologies also integrate the ability to save data already filtered from one or more links for a predetermined amount of time. That allows post-facto forensics to help answer many questions: How did the attackers get in? Who are they? What activity patterns serve as markers of their presence? How do we keep them out in the future?

With so much diverse data moving so fast, advanced capabilities are the only way operators can hope to answer crucial questions about the scope of an intrusion and the breadth of possible damage. They can draw on high-quality logs of past activity from disparate forms of information, saving time and potentially fending off greater damage.



To make use of these tools, security managers need an intuitive user interface on their IDS that provides a unified, high-level view of network activity, including all components and points of observation across a wide swath of time.

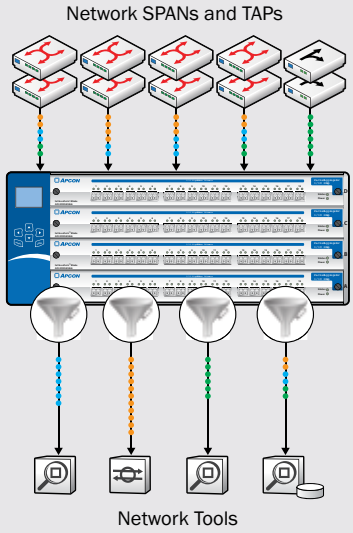
A Way Forward for Big Data

As we've seen, the challenge Big Data presents for network operations and security managers is a beast with many tentacles. Achieving the potential of unprecedented business insights requires far more than a commitment to enormous storage capabilities.

Indeed, the new volume, velocity and variety of Big Data requires a sophisticated approach to network monitoring with the integrated capability to create order from disorder. These newer solutions can not only make the costs associated with acquiring advanced Big Data resources a lot more palatable, but they actually help Big Data provide measurable and even handsome returns on such investments.

Recent advances in network monitoring switches mean that upgrading from a 1G network environment to a 10G – or even 40G – environment can be up to 75 percent less expensive than would be the case using old older approaches. The trick is to pack more network monitoring and switching capabilities power into a single switch, increasing the capacity and capabilities of these switches so you can access more monitoring points on the network. In effect, you can use the switch structure as an aggregation point for all the tools you need to monitor throughout the network.

With the growing volume and variety of data passing through packet-switched networks, effectively monitoring production network traffic – and keeping tools from being overloaded – becomes a challenge. Effective and efficient packet filtering in the network monitoring system has become critical.



Chief among the benefits is far greater visibility – up to 100 percent visibility of all monitoring points – and access for any monitoring tool to see every point in the network 100 percent of the time.

It is possible for companies to digest Big Data while significantly extending the life of their monitoring and analysis tools. The answer is in the box.

APCON
Solutions for Networks

APCON, Inc.
9255 SW Pioneer Court
Wilsonville, Oregon 97070 USA
Tel: +1 503-682-4050
Toll Free: 1-800-624-6808

Engineering Design Center
501 W President George Bush Highway,
Suite 100
Richardson, Texas 75080 USA

E-mail: sales@apcon.com