Forbes insights

The Need for Speed: How the Internet of Things Will Live on the Edge

IN ASSOCIATION WITH:

TABLE OF CONTENTS

INTRODUCTION: THE DATA FLOOD AND PROCESSING CHALLENGES

Edge computing is here. Forbes Insights and Intel surveyed 700 industry executives to determine the current state of their organizations' Internet of Things (IoT) efforts. We asked about the present and future roles that the edge will play in their businesses.

The number of data-generating devices—from sensors embedded in retail signage to smart cameras on streetlights, to massive crop inspection devices that can tell weeds from produce and kill the weeds, to millions of other devices—is expanding at incredible speed. In addition to merely creating data, the devices must also properly handle it—process it on the device, send it to another device or send it into the cloud.

Technology is also enabling the rapid increase in the number of edge devices, and the devices that will be handling the data flow between the edge and the cloud. This technology is the 5G communication standard.

The 5G communication standard currently being implemented worldwide will enable more devices to be connected, and to transmit higher-resolution images and stream much more data than previous devices because of increased bandwidth. The higher-data bandwidths provided by 5G will enable an increase in the resolution and performance of computer vision devices—enabling them to not only recognize general facial features, but also to detect mood, to detect smaller objects and to more rapidly stream relevant images onto the internet. Another benefit: the ability for edge devices in remote areas where internet connectivity was once unavailable to communicate with the internet over 5G.

In this report, we'll define edge computing, explore underlying technologies, examine key aspects of their use and plans for IoT, and highlight the current state of executives' edge computing efforts.

EDGE COMPUTING DEFINED

"AI edge computing is AI running on IoTs," says Anusua Trivedi, senior data scientist at Microsoft. Trivedi is working to develop artificial intelligence (AI) that functions on edge devices that interface with different IoTs and add intelligence to the process. For example, Trivedi's research connects medical scanners, cameras or other diagnostic equipment to devices such as iPads, iPhones and Android devices. The data that is produced by the diagnostic equipment is transferred to the iPads, IPhones and Android devices, where AI is used to evaluate the images.

Trivedi is interested in bringing advanced technologies to locations that have limited connection to the cloud. The AI processing is often done locally on edge devices, based on learned profiles for advanced diagnostics. For example, a digital X-ray machine—the "edge" device—takes an X-ray of a person's chest. Data is then transferred from the X-ray machine to a tablet or other devices that are used to interact with the X-ray machine. This edge device applies AI to the image(s) and "reads" the X-ray, and then provides some predictive analysis to the doctors for easier evaluation of images. Also, when internet access is available, the images can be transmitted into the cloud, and a cloud server may further evaluate the image and return a report to the edge device. Additionally, the scanned data may be combined with thousands of other scans in the cloud server, and an improved probabilistic model can be developed. This new "model" can eventually more accurately interpret the images.

Edge computing can be broken into what Mahadev Satyanarayan (Satya), Carnegie Group professor of computer science at Carnegie Mellon University, describes as tiers.

TIER 1 is the server or data center located somewhere inside the cloud. This device is located farthest from the edge devices but has the most processing power. In this paper, we'll just refer to the Tier 1 devices as cloud servers.

A **TIER 2** device can go by a variety of names—edge server, edge cloud or cloudlet are the names most frequently used. A Tier 2 device is located close to the edge device and operates between the edge devices and the cloud data centers. These provide significant advantages, effectively enabling edge computing. In this paper, we'll refer to these as edge servers, although an edge server can be a complex array of hardware and software that more closely resembles a small data center than it does a mere server.

Finally, **TIER 3** is the device at the edge. It's the X-ray machine that generates image data, the robotic manufacturing device that performs processes using instructions received through the internet, the quality-control video camera installed on a production line to detect defects, a consumer product like a cell phone or virtual reality glasses, or any of millions of other devices that are located at the edge, generating and transmitting data, performing processes in response to incoming data, or applying knowledge built into the device to perform analytic or other processes. Edge devices are located at what Satya calls "the extreme, extreme edge," because these devices are the first devices in the data chain—capturing data and processing it on the device, or the last devices, the final devices, in the data chain, acting on data returned from the edge servers or the cloud.

HOW THE EDGE SERVER ENHANCES EDGE DEVICES AND ASSISTS CLOUD SERVERS

Before discussing the benefits of the edge server, we will take a closer look at what it is. Satya refers to these devices as "cloudlets." He says that a cloudlet is like a small data center: "A cloudlet is like a cloud. It uses multitenancy (meaning that there may be hundreds of users at once) and many technologies that are comparable to a cloud data center. It can be a rack of servers in a closet at Starbucks, or a rack of servers on a ship, or at the back of a large aircraft," says Satya. It can also be a server mounted on a cell tower, on a streetlight or in other locations that are close to edge devices. As noted earlier, a Tier 2 entity can be called a cloudlet, an edge cloud or an edge server.

Placing edge servers near the edge devices provides many benefits:

REDUCED LATENCY: By locating the edge server near devices at the edge, the amount of time it takes to transmit data between edge devices and the edge server is significantly reduced versus the time to transmit data further up and back through the cloud. Reducing latencies to a few milliseconds enables such applications as virtual reality, medical imaging and other applications that require almost instantaneous responses. Reduced latency, plus the ability of the edge server to rapidly evaluate the data and return a response in milliseconds, makes applications requiring millisecond responses possible. For example, a person who is using a virtual reality headset relies on visual and auditory cues, relative to his location and head position. When he turns his head, he expects the field of vision to change so quickly that changes in the visual field appear to be smooth and without jitter. This requires the transmission of head position and other data to an edge server—and new video sent from the edge server—enabling an update of the visual images in milliseconds. If the data coming from and returning to the VR headset had to go to the cloud, the goal of millisecond updates wouldn't be achievable. The VR application would fail.

BETTER COORDINATION WITH OTHER DEVICES: Placing edge servers close to edge devices and other edge servers enables high-speed interaction with other devices. Placing each edge server close to other edge servers creates a "web" or "fog" that can share data and processing tasks. For example, in the case of a traffic accident or road blockage, edge servers can rapidly send information to other edge servers or other edge devices, and coordinate response or provide continuous information. In the case of a vehicle of interest that is being tracked, smart cameras can relay images of the vehicle of interest to an edge server located closest to it. As the vehicle moves from the

field of vision of one camera to another, the process of tracking can also be switched from camera to camera and from edge server to edge server, assuring the most accurate tracking possible.

REDUCING CERTAIN CLOUD REQUIREMENTS: Edge servers will be configured to be high-performance "mini data centers." These will probably be populated with high-speed CPUs, field programmable gate arrays (FPGAs) and adequate storage. Edge servers can cache data that is downloaded from the cloud data centers and deliver it to edge devices when it is needed. For example, the tablet that is used to interpret X-ray images can be used to receive updated learning models from the cloud data center. The edge device doesn't have the capacity to handle a high-speed data transfer from the cloud data center. Here's where the edge data center comes in—it downloads the entire update from the cloud data center, storing it locally. When the edge device (the tablet) is ready to download the update data, the edge server transmits the data at speeds that the edge device can handle. In effect, the edge server acts as a buffer between the edge device and the cloud data center.

PERFORMING MUCH OF THE PROCESSING THAT WOULD OTHERWISE BE PERFORMED BY EDGE DEVICES: For example, a point of sale (POS) terminal is used to capture and authorize all sales transactions. The device has a limited amount of memory storage. The data may be sent to the edge server for storage. At the end of the day, the edge server performs the tasks of generating a variety of daily reports that are then sent through the internet to the cloud server. Effectively, using an edge server can provide the edge devices with "edge server" performance, by giving the edge device the computing capabilities of the edge server.

ACTING AS "CUSHIONS" IN THE EVENT OF A CLOUD FAILURE: Although these rarely happen, hacking of cloud servers, power failures or cyberattacks may temporarily bring down centralized servers. In the event of such an incident, edge servers would be able to perform many of the functions that would otherwise be sent to the data centers. When the data centers return online, the edge server can update the data centers with changes that occurred while the data center was down.

ENABLING HIGHER LEVELS OF SECURITY: Edge servers will provide the processing power that will enable high levels of security. They will be able to run complex algorithms that can shield edge devices from data breaches and can also provide high levels of encryption and decryption. Edge servers will provide an additional, potentially impenetrable layer of security to the edge.

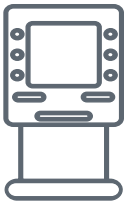
BENEFITS OF EDGE COMPUTING: EDGE DEVICES

IoT at the edge enables intelligent processing in places where it wasn't previously possible, such as: a smart video camera at a corporate facility that performs facial recognition to determine access rights for people approaching the area where the smart camera is deployed; a credit card authorization system that matches a face to the user of a credit card, adding an extra layer of security; or a scanner that passes over fields of crops, distinguishing weeds from healthy produce and instructing a robot to remove the weeds.

The intelligent processing capabilities at the very edge provide benefits that were never possible before now, including:

GREATER NETWORK AVAILABILITY: By processing data on the edge device, instead of sending it to the cloud, the edge device puts less load on the internet infrastructure than it would if it sent all data to the cloud for processing. (For example, a camera may count cars over a period of time, sending aggregated totals at intervals, rather than sending all video to the cloud for constant processing). Processing data on the edge device will reduce the amount of network traffic used, in comparison to "dumb" sensors that continually send data into the cloud. Because images are pre-processed in the edge device, and only significant images are sent to the cloud, there is no longer any need for storing unnecessary images that are sent streaming from a dumb camera into the cloud. The smart camera manages the interpretation and detection of image data.

POTENTIAL COST CONTROL: Local processing on the edge device reduces the amount of processing that will be done on a cloud-based data center, which can help to control costs.



DATA COLLECTION OR CUSTOMER INTERACTIONS AT THE EDGE: Applications installed on the edge devices can be used by the customer or the edge device to input data or interact with the organization. For example, a wireless kiosk can be used by the customer for numerous tasks—catalog review, ordering, participation in surveys. Portable devices can extend the reach of the organization to wherever the customer is located. At the end of the day, or at preset intervals, data can be transferred from the edge device to an edge server or directly to a cloud data server.

INTELLIGENT MANAGEMENT OF DATA, REDUCING DATA LOAD ON THE NETWORK AND THE CLOUD: For example, an intelligent thermostat will regularly monitor the temperature at the thermostat and at attached sensors. It controls the functions of the heating and cooling systems. The data may stay in the thermostat, rather than being sent into the cloud. If temperatures go above or below certain levels (say, for example, 125 degrees or 20 degrees Fahrenheit), the thermostat may sense an issue with the heating or cooling systems and transmit a warning message of some sort into the cloud. Although the edge devices are able to communicate with the cloud, some devices may very rarely do so.

AUTONOMOUS OPERATIONS. For example, a drone can fly over a field, counting plants and also detecting plants that aren't maturing adequately, and detecting weeds. It can send a message to a robotic device on the field, instructing it to remove the weeds and immature plants. In this case, the drone and the robot are edge devices—the drone uses AI to interpret the computer vision, and the robot uses GPS signals from the drone as well as computer vision to find the targeted object, and device controls to remove the weed or immature plant. The autonomous processes described here are performed using no interactions with the cloud.

WORKLOAD CONSOLIDATION/VIRTUALIZATION. In organizations where edge devices are dedicated to that organization's processes (as opposed to devices "in the cloud" and not dedicated to a particular organization), such as robots, diagnostic equipment or other devices, it's possible to switch the operation of the devices to that of a virtualized asset to the organization. Edge devices that are equipped with FPGA chips can be reprogrammed in milliseconds from devices that perform their usual edge functions into virtualized machines that assist the organization. Edge devices with FPGAs can provide extra computing capabilities that are not available in other edge devices.

STATE OF PLAY IN ORGANIZATIONS TODAY

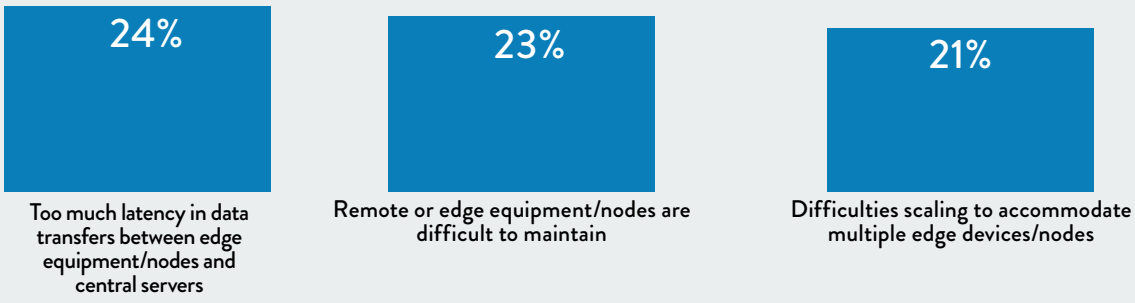
If you apply Microsoft's Trivedi's definition of an edge device as "anything IoT," organizations are already making good use of the edge. Nearly half (47%) of organizations that we surveyed have begun implementing visual analytics on a limited scale. Another 33% reported that they've implemented visual analytics across selected functions or business areas.

Reporting organizations indicated that they use a variety of edge devices or nodes that are linked to their IoT networks. Some of the top devices are sensors, cameras/visual sensors, audio/microphones and mobile phones.

Of the above edge device functions, visual analytics seem to be extremely important. Over the next three years, 85% of survey respondents expect increases in the use of visual analytics in networks of cameras or visual sensors to increase by up to 49%. Currently, 61% of organizations using visual analytics are able to recognize and analyze people. The recognition numbers are similar for objects (60%) and situations (60%).

Of all the input methodologies used for data collection and input, it's been said that visual input will be the most prominent. In the coming years, there will be a need for more images, and more high-resolution images. The increase in the amount of data that is generated by the visual systems will potentially stress the infrastructure that will handle the data. Because of this, much of the image processing will be performed directly on the edge device producing the images, or on a Tier 2 device.

Organizations indicated that there are many issues they've faced in building out IoT (and edge computing) to its fullest capabilities. These include:



With the emergence of edge servers, these challenges should go away. Edge servers are just emerging as key elements in the overall infrastructure. Developers are working on standards for edge servers. It's not clear how they'll be monetized—will they be available for all businesses and users and treated like a utility? Will there be different implementations (public safety, private subscription)? It's likely that they may be run as a subscription service, much like cloud providers, with much of the maintenance handled by the edge server provider.

The fact that edge servers will be located near the edge means that issues of latency and data transfers should be minimized. Transfers to central servers or other edge nodes will be handled by the edge server. Finally, scaling

may also become a relative non-issue, due to the high capacity and flexibility of edge servers.

Artificial intelligence is a key feature in many edge devices. It is used in visual processing units to evaluate incoming data and by "speech recognition" appliances. AI will be an integral part of the edge infrastructure. Over the past few years, AI has seen a sharp rise in use and capability. It is being applied to IoT data, either in the edge devices, in edge servers or at cloud data centers. AI is not "intelligence" per se—it's the rapid processing of large quantities of numbers.

Until recently, adequate performance to run the processes that enable AI has not been available. Today, higher-performance CPUs, FPGAs and video processing units (VPUs) can provide the performance that's been missing. Devices are now able to perform the "learning" that's necessary for AI. Other devices are able to match the "learned" data to input, and perform actions based on the data. The increased performance of CPUs, FPGAs and VPUs will enable organizations to upgrade their systems and perform more processing at the edge. The move of processing to the edge will reduce the workload on cloud data centers and reduce the traffic over the cloud. The benefits of higher-performing CPUs and FPGAs should be clear—more processing on the edge means better data and less reliance on cloud resources.

The continued growth in processor performance, coupled with the speed and flexibility of FPGAs, visual processing chips and neural computing processors, will accelerate advances in AI. Specialty neural network processors being developed by Intel will further accelerate the development of AI applications.

WORKLOAD CONSOLIDATION

Because edge devices can collect and process data, they can also be valuable as a source of pre-analyzed or summary data. For example, a smart camera mounted on a streetlight can count and categorize vehicles. At preset intervals, it may send the data—number of cars, trucks, bicycles, pedestrians that pass through an intersection in certain time periods. This summary data may be sent through the cloud and used for traffic prediction and analysis.

A POS terminal may record all transactions for a 24-hour period. Rather than sending a complete list of transactions, the edge device can summarize data in a way that can be more easily digested at a cloud data center—along with summarized data from thousands of other POS terminals.

A camera that is used for quality control, scanning thousands of items passing beneath it each second, may summarize the data, reporting on types of defects, times that defects are detected (perhaps suggesting that a particular machine or worker is producing more defective products) and other data. This summary of data will be processed at a central location—but processing the summary data will be much more efficient than analyzing the gigabytes of raw data that are generated by the visual inspections.

Pre-processing data, performing sophisticated analysis and automating other tasks on the edge—rather than at more centralized data centers—is a tremendous advantage that edge servers and edge devices provide.

A CHECKLIST FOR DEVELOPING AN EDGE STRATEGY TODAY

Although many organizations already have edge strategies today, it's useful to consider a number of factors that will impact adoption of edge and AI strategies:

- What types of devices or applications will be relevant to your business?
- Can you develop applications to exploit devices at the edge?
- What specific areas of edge computing will make your organization valuable to users at the edge? Or, what kind of edge devices can you develop that will set your organization apart from the competition?
- How can your organization best use workload consolidation to automatically orchestrate your infrastructure all the way to the edge, optimizing use of resources and speeding deployment of new services?
- Are there services that you can link into edge technologies? (Think, for example, of the companies that record video from cameras and perform analytics on the video.)
- Are you identifying edge server providers and carefully monitoring how edge server technologies and licensing will be implemented?
- How will you plan for the future by ensuring that your edge deployments have the extra processing power that is needed to enable flexibility, growth and upgrades?

Think all this through carefully: Working out an edge strategy today will give your organization a competitive edge tomorrow.

METHODOLOGY

Forbes Insights, in partnership with Intel, conducted a survey of 700 executives. The survey was global, focused on the United States (57%), followed by Canada (9%), the United Kingdom (5%), France (4%) and India (4%). Seventy percent of respondents were in the C-suite of their organization, including chief information officers (17%) and chief technology officers (16%). Respondents were divided equally across seven key industries: communications, energy, financial services, healthcare, manufacturing/industrial, retail and transportation.

ACKNOWLEDGMENTS

Forbes Insights and Intel would like to thank the following individuals for their time and expertise:

- **Mahadev Satyanarayan (Satya),** Carnegie Group Professor of Computer Science, Carnegie Mellon University
- **Anusua Trivedi,** Senior Data Scientist, Microsoft

Forbes insights

ABOUT FORBES INSIGHTS

Forbes Insights is the strategic research and thought leadership practice of Forbes Media, a global media, branding and technology company whose combined platforms reach nearly 94 million business decision makers worldwide on a monthly basis. By leveraging proprietary databases of senior-level executives in the *Forbes* community, Forbes Insights conducts research on a wide range of topics to position brands as thought leaders and drive stakeholder engagement. Research findings are delivered through a variety of digital, print and live executions, and amplified across *Forbes*' social and media platforms.

EDITORIAL & RESEARCH

Erika Maguire
EXECUTIVE EDITORIAL DIRECTOR

Kasia Wandycz Moreno EDITORIAL DIRECTOR

Hugo S. Moreno EDITORIAL DIRECTOR

Ross Gagnon RESEARCH DIRECTOR

Scott McGrath RESEARCH ANALYST

Mark Brownstein REPORT AUTHOR

Zehava Pasternak DESIGNER

PROJECT MANAGEMENT

Casey Zonfrilli
DIRECTOR, ACCOUNT MANAGEMENT

Tori Kreher
PROJECT MANAGER

Brian Lee
PROJECT MANAGER

Todd Della Rocca
PROJECT MANAGER

SALES

North America
Brian McLeod VICE PRESIDENT
bmcleod@forbes.com

Matthew Muszala EXECUTIVE DIRECTOR
mmuszala@forbes.com

William Thompson DIRECTOR
wthompson@forbes.com

Kimberly Kurata MANAGER
kkurata@forbes.com

Europe
Charles Yardley SVP MANAGING DIRECTOR
cyardley@forbes.com

Asia
Will Adamopoulos
PRESIDENT & PUBLISHER, FORBES ASIA
wadam@forbesasia.com.sg

499 Washington Blvd. Jersey City, NJ 07310 | 212.367.2662 | **https://forbes.com/forbes-insights/**