Association for Computing Machinery
Advancing Computing as a Science & Profession

*Ubiquity,* an ACM publication
November 2015

Ubiquity Symposium

# The Internet of Things

## Fog Computing: Distributing data and intelligence for resiliency and scale necessary for IoT

***By Charles C. Byers and Patrick Wetterwald***

### Editor's Introduction

*The Internet of Everything (IoE) is more than a $19 trillion opportunity over 10 years. Fifty billions of devices will be connected to various networks in 2020. This is bringing new technical challenges in all domains and specifically in the data processing. Distributed intelligence is one of the key technological answers. We call it "fog computing." Fog can provide intelligent connection of people, processes, data, and things in hierarchical Internet of Things networks. By supplementing the cloud and providing intermediate layers of computation, networking, and storage, fog nodes can optimize IoE deployments—greatly enhancing latency, bandwidth, reliability, security, and overall IoE network performance. The article will analyze the architecture and main design choices of this technology.*

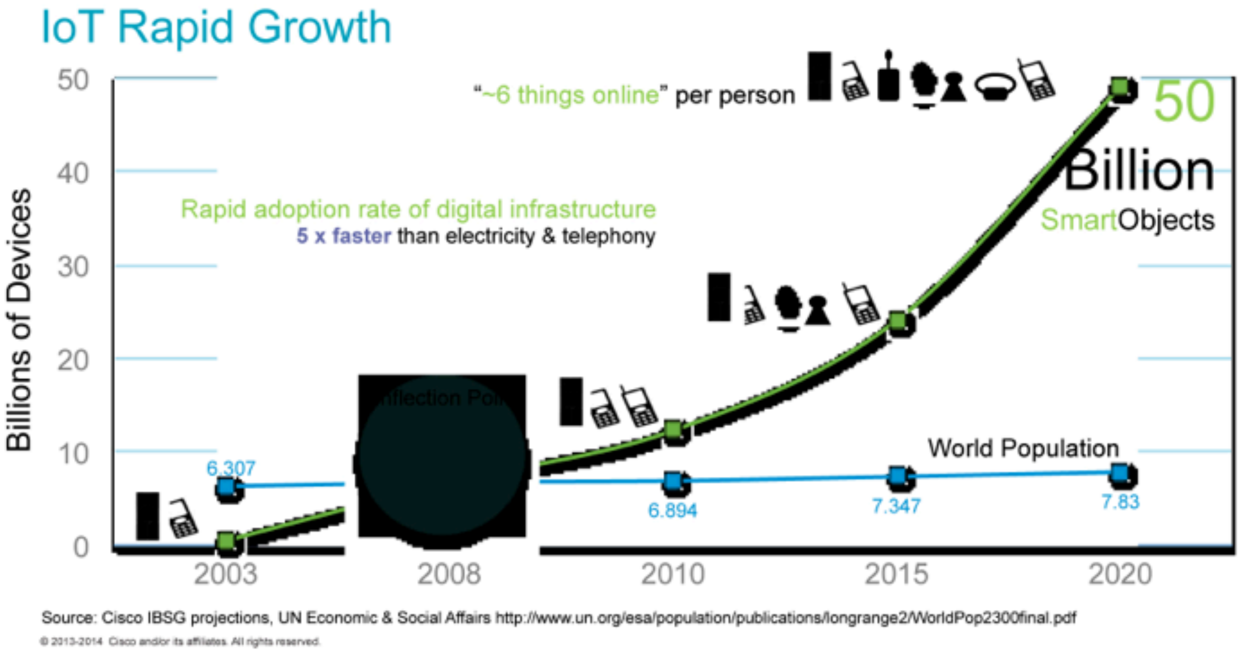
## Ubiquity Symposium

# The Internet of Things

## Fog Computing: Distributing data and intelligence for resiliency and scale necessary for IoT

***By Charles C. Byers and Patrick Wetterwald***

The number of Internet connected devices will cross the incredible total of 50 billion by 2020. Beyond the consumer market, a large part of this growth is occurring in the industrial space including smart grid, smart cities, industrial automation, and transportation. The diagram below illustrates this rapid growth and the inflection point, which already happened in 2008, when the number of connected devices surpassed the number of people on Earth.



Source: Cisco IBSG projections, UN Economic & Social Affairs http://www.un.org/esa/population/publications/longrange2/WorldPop2300final.pdf

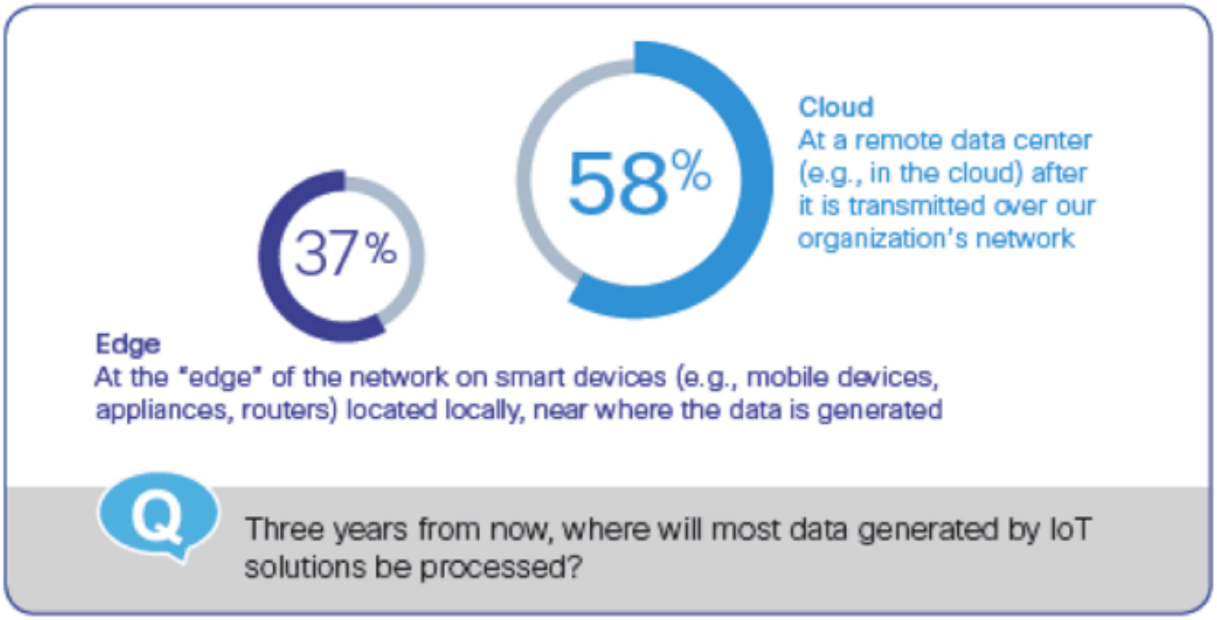
© 2013-2014 Cisco and/or its affiliates. All rights reserved.

Usual cloud-based architecture, where application intelligence and storage are centralized in server wire centers, satisfies the need of most of the Internet of Things (IoT) applications, but begins to break down when real-time requirement (control loops), high volume of data, or limited network bandwidth play an important role in the deployment model. The need for decentralized processing is emerging. Some references will call it the "edge computing."

This is not just about aggregation or concatenation of sensed physical data (like a typical gateway will do), but really about distributed intelligence, where effective real time and deterministic processing is needed to implement a functionality.

The move from cloud computing, or centralized computing, to edge computing has already begun as the following diagram shows.  According to a 2014 Cisco survey, by 2017 37 percent of IoT computing will be located at the edge of the network.

Among survey respondents, the move toward edge computing is already underway.



Source: Cisco Consulting Services, 2014

Some endpoint devices—such as intelligent sensors and actuators—and smart handheld devices include significant computing, networking, and storage capabilities. These will be valuable in complex Internet of Everything (IoE) systems. However, endpoints can suffer from space, power, bandwidth and security constraints.  The intermediate layers of networking equipment and IoT gateways are often the perfect places to host IOT processing for systems, which require gathering data coming from different sources (sensors, distributed data bases, etc.). They are located on the data path and their processing capabilities now allow running complex IOT applications.

## Why Distributed Intelligence?

Several reasons include:

- **Scalability.** Large deployment of smart metering systems involves millions of end points, which makes the use of intelligent concentrators mandatory. A digital factory represents a few tens of thousand of sensors and actuators. A smart city with parking lot management, road traffic control, and environmental monitoring over a very large territory is bringing its own deployment complexity. The centralized approach is not sufficient to handle this increasing volume of end devices and its geographical specificities. Data is most relevant, or safest, if it is processed close to the edge of the network.

- **Network resource preservation.** The volume of data generated by all type of sensors has a direct impact on the network bandwidth necessary to carry this new created information (we may call it "little data" by opposition to the term "Big Data"). Some remote locations are only connected using wired or wireless connections with limited bandwidth (2G/3G/4G, ADSL, or satellite link). Distributed processing helps relieving the constraints on the network by sending to the cloud or operation center only the necessary information and by doing most of the data processing, like video analytic for instance, at the remote site much closer to the data's source.

- **Close loop control.** Low latency is required to create stable behavior in real-time systems. Large delays found in many multi-hop networks and overloaded cloud server farms prove to be unacceptable, and the local, high performance nature of distributed intelligence can minimize latency and timing jitter. Many critical applications such as industrial automation, inflight control system, electrical tele-protection system, medical applications, or internal vehicle networking have very tight requirements in term of latency and jitter. Only local processing could satisfy the most stringent requirements. Very often, this is combined with advanced networking technologies like deterministic networking, where the guarantee of delivery of packets in a bounded time is granted.

- **Resilience.** It is of most importance that mission critical processes run even if communication with the operation center is not effective. An architecture based on distributing processing is not only recommended but is often the only valid solution.

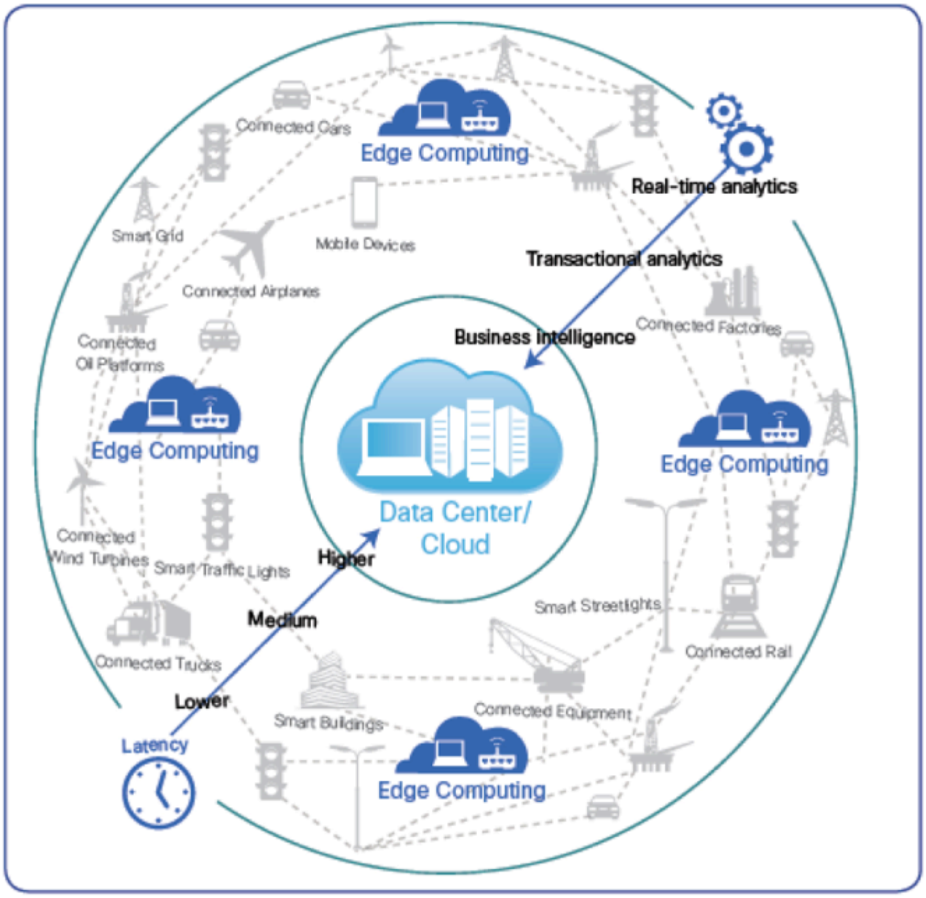
- **Clustering.** Moving from individual devices to clusters. For example, a connected vehicle has many sensors and actuators, but they are seen from the outside as a single unit (vehicle) that communicates with other vehicles or the terrestrial infrastructure.

## The Role of Real-Time Data Analytics

Due to the increasing volume of data and the real-time requirements, we see an evolution in the processing of this information. Processing must happen at the right place at the right time, so real-time analytics is hosted close to the location where the data is generated.

The following figure illustrates this concept from real time analytics to transactional analytics to business intelligence.

Edge computing helps ensure that the right processing takes place at the right time and place.

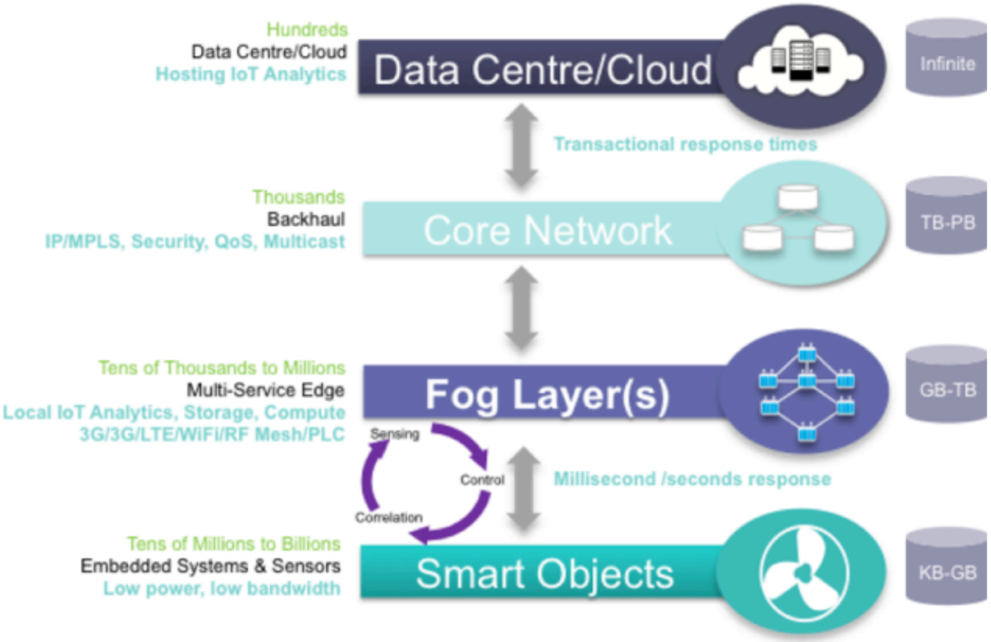


Source: Cisco, 2014

Not all raw application data needs to be entirely processed in the cloud. It is often more adequate to do a pre-processing at the edge and send only the relevant information to a centralized data center. The complete application being composed by edge and cloud computing.

## Fog Architecture Overview

The type of architecture to address the above challenges is called “fog computing” (closely related to cloud computing, but “closer to the ground”). Fog computing is a system-level architecture optimized for the distribution of computational, networking, and storage capabilities in a hierarchy of levels within an IoE network. It seeks to provide the correct balance of capacity among the three basic capabilities at exactly the levels of the network where they are the most optimally located. The figure below depicts the fog computing layers, located at intermediate levels of the network hierarchy between the cloud and the smart-object layer. The fog can form a connected graph, with fog nodes existing at many layers of a hierarchy, with rich interconnect topologies between them, and to smart objects, core, and cloud layers. Different layers have different properties, with respect to storage capacity, network bandwidth, response time, etc.

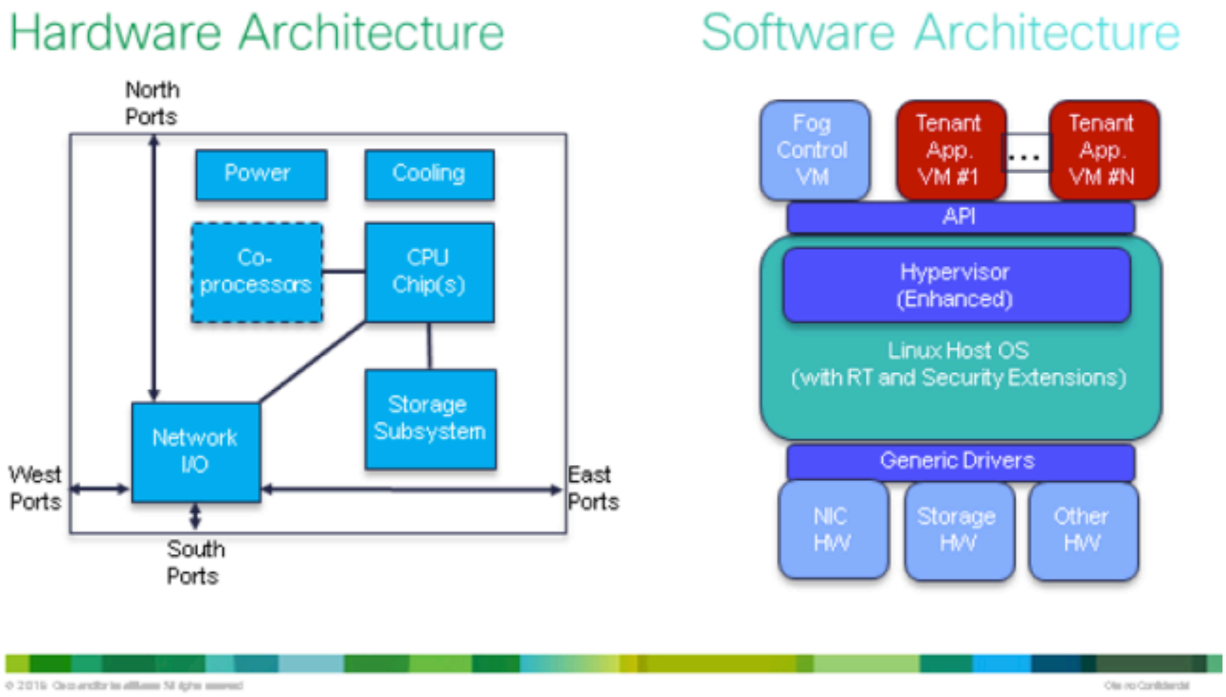

Hundreds
Data Centre/Cloud
Hosting IoT Analytics
Data Centre/Cloud
Infinite
Transactional response times
Thousands
Backhaul
IP/MPLS, Security, QoS, Multicast
Core Network
TB-PB
Tens of Thousands to Millions
Multi-Service Edge
Local IoT Analytics, Storage, Compute
3G/3G/LTE/WiFi/RF Mesh/PLC
Fog Layer(s)
GB-TB
Sensing
Control
Correlation
Millisecond /seconds response
Tens of Millions to Billions
Embedded Systems & Sensors
Low power, low bandwidth
Smart Objects
KB-GB

Fog computing builds upon the basic capabilities of cloud computing, but extends them toward the edge of the network, and often right up to the intelligent sensors, actuators, and user devices that constitute the IoE. Many familiar cloud techniques, such as virtualization, orchestration, hypervisors, multi-tenancy, and excellent security are seamlessly extended from the cloud through the fog layers. With the fog layers augmenting the cloud, network capabilities that are difficult or impossible exclusively in the cloud can be provided. Fog offers performance, scalability, efficiency, security, and reliability advantages compared to cloud only solutions for critical IoE applications. Fog usually doesn't replace the Cloud (which has many advantages due to centralization and scalability), fog supplements the cloud for the most critical aspects of network operations. Fog also supplements and greatly expands the capabilities of intelligent endpoint devices and other smart objects.

## Fog Nodes

The fundamental element of the fog computing architecture is called a "fog node." It is a collection of modular hardware and software elements that can be configured to perform the specific functions required as various levels of the network hierarchy by the mix of IoE applications the network is expected to support. Fog nodes lower in the network hierarchy (closer to the endpoints) often have relatively simple hardware and software configurations and modest capacity and performance specifications, while those higher in the hierarchy (closer to the cloud) are often quite sophisticated, with performance approaching the high-end servers and high bandwidth networking equipment found in major data centers.

Fog nodes are interconnected with a wide variety of wired, wireless, and optical networking facilities. A common theme to these networks is reliance on the IPV6 protocol, because of the large address space it supports, necessary for extended systems with tens of billions of nodes. Ethernet is also common in fog networks, because of its scalability, performance, and ubiquity. Each fog node has a number of interfaces. Southbound interfaces connect to lower fog layers, and ultimately to the IoT's sensors, actuators, and other intelligent endpoints. Northbound interfaces connect to higher order fog nodes, and ultimately to the Internet core network and the cloud. East-west interfaces interconnect fog nodes at approximately the same layer of the network hierarchy, carrying inter-node traffic in support of fog features like distributed processing or storage, load balancing, network resiliency, and fault tolerance. The following figure describes the preferred hardware and software architecture of a fog node at a high level.



**Fog node hardware architecture.** From a hardware perspective, the preferred implementations of fog nodes fall into two broad clusters: single board configurable platforms and high capacity modular platforms. Fog nodes can be implemented as ancillary functions on traditional network elements (like routers, servers, storage engines, appliances, gateways, edge devices, or access points), or as stand-alone fog boxes. Fog nodes are configurable with different options for main processors (including X86, ARM, and other CPU and coprocessor choices in several performance grades), networking (including programmable packet processing, security accelerators, and many types of wired, wireless, and optical transport interfaces), and storage (including large DRAM banks for in-memory databases, solid state drives, and rotating disks).

Fog nodes are scalable and adaptable to many different deployment environments. This scalability can include dimensions of processor performance, network capacity, memory, and persistent storage size. Fog hardware is typically versatile in its ability to support deployment with various power sources, cooling strategies, physical enclosure types, and environmental hardness levels, serving a broad range of IoE use cases and deployment scenarios.

**Fog node software architecture.** From a software perspective, fog nodes are highly virtualized machines with multiple VMs running under a highly capable hypervisor. That hypervisor includes real-time enhancements and security extensions needed to serve the needs of critical fog applications. The figure above shows the high-level software architecture of a fog node.

The tenant application processes shown on top are all virtualized. They are carefully isolated from each other by the fog software infrastructure, and communicate between functions via APIs.  This isolation is critical, so no process uses more than its allocated share of the fog node's resources, and no process can interfere with the operation of other processes or the fog node's shared software infrastructure (the application processes may be written by or owned by many different entities who are tenants of a given fog node, and not all of these are necessarily trustworthy).

The fog software infrastructure provides services to the tenant applications. There is a basic Linux host operating system, with extensions for real-time operations and enhanced security. A hypervisor runs on top of this, providing a virtualized environment. Many resident software modules are parts of the fog node's software infrastructure, including network management, packet processing, network-function virtualization (NFV), file systems, database management systems, and cryptography processors.  Fault tolerance is also (at least partly) implemented as a shared fog software infrastructure service, providing resiliency in the presence of failed fog nodes or network links.

## Cloud and Fog in IoE Networks

As mentioned above, the fog can supplement the cloud for a specific subset of applications that can benefit from fog's unique capabilities.  We don't dispute the cloud deployment model is ideal for a large subset of IoE applications, but when certain critical application requirements are taken into account, supplementing the cloud with fog leads to more optimal system properties. The following discussion will highlight how the Fog architecture satisfies some of the situations where Fog Computing is needed in IoT networks:

**Latency.** Being located close to the source of information (smart objects), latency and jitter will be optimized to meet close loop functioning requirements.  Fog nodes can routinely achieve millisecond-level latency for local control loops, while realistic cloud based deployments may have worst-case latencies two orders of magnitude larger.

**Geographic diversity.** The geographic location of fog nodes often holds advantages compared to pure cloud deployments. Data can be located at the optimal depth in the network, intelligence can be localized as appropriate, and caching structures can be optimized.  Often, data is most relevant or safest if it is processed close to the edge of the network.  If data needs to be aggregated or anonymized before transmission to the cloud, fog nodes are the perfect location for that function.

**Network bandwidth.** Network bandwidth scarcity is an ongoing concern. Collections of IoE elements will quickly overload interconnect links if all of their raw traffic is carried back to the cloud. By interposing fog layers, the large volumes of data associated with sensors and actuators can be managed close to the source, and the bandwidth on backbone links toward the cloud can be preserved. Fog nodes can filter data, perform analytics, and provide local storage to help minimize the impact of IoE on network backbones. Analytics, in particular, benefit greatly if critical parts of the algorithms are located in fog nodes only a network hop or two from the data sources. The southbound interfaces collect the potentially large data streams from the sensors; the analytics algorithms digest the data using processor resources in the fog node; and only high level post-analysis results (which require orders of magnitude smaller network bandwidth) are sent to the northbound interface.

**Reliability.** Many IoE applications will be mission critical or even life critical. These applications must continue to operate as expected even if cloud resources or the network links needed to reach them are down, or seriously overloaded. Local fog nodes can provide back-up service logic even if the cloud isn't responding. They may not have the full capabilities of the cloud, but often have enough basic local functionality to maintain critical services until cloud processing can be restored. Collections of several fog nodes can act as fault tolerant sets, supporting the application on remaining good fog nodes even if other nodes in the set are malfunctioning.

**Intelligent endpoints and the fog.** Some IoE network architectures may attempt to push as much of the computation, networking, and storage as they can down to highly intelligent endpoint devices, making the data network between the devices and the cloud "dumb pipes." While this may work for a subset of use cases, in general we believe highly capable fog nodes provide a more optimal network deployment. Intelligent endpoints may be under physical constraints, lacking the energy, physical space, or environmental control to provide adequate processing power, network throughput, or storage capacity. They typically lack the modularity of fog nodes, making customization or upgrades difficult. Fog nodes can have better network and physical security than endpoints (which may lack strong crypto support, or be more vulnerable to physical attack).

## Security

Finally, let's address security. IoE networks may transport highly sensitive data, and could have control of very dangerous actuators and high power systems. Data breaches and hacking must be avoided at all costs. Fortunately, Fog Nodes have the correct mix of span of control, computational power, and network connectivity to serve many security purposes. Fog nodes can act as proxies to run very strong cryptography, while sensors and actuators below them may lack the computational throughput or energy reserves to do so. Fog nodes can have full hardware roots of trust, where their security status is guaranteed as a function inherent in their hardware regardless of the application software running upon them. This trust can be selectively extended to any applications running on the fog nodes; applications on other fog nodes, and less intelligent sensors and actuators that are connected.

## Conclusions

In conclusion, the IoE is more than a $19 trillion opportunity over 10 years, and advanced techniques like fog computing are needed to facilitate its deployment. Fog can provide intelligent connection of people, processes, data and things in hierarchical IoT networks. By supplementing the cloud and providing intermediate layers of computation, networking and storage, fog nodes can optimize IoE deployments, greatly enhancing latency, bandwidth, reliability, security, and overall IoE network performance.

## About the Authors

Charles C. Byers is a platform architect with Cisco's Chief Technology and Architecture Office. He works on the architecture and implementation of media processing systems, Fog Computing platforms, and the Internet of Everything. Before joining Cisco Systems, he was a Bell Labs Fellow at Alcatel-Lucent. During his 28 years in the telecommunications networking industry, he has made significant contributions in areas including voice switching, broadband access, converged networks, VoIP, multimedia, video, and modular platforms. Byers received his B.S. in electrical and computer engineering and M.S. in electrical engineering from the University of Wisconsin, Madison. He holds 51 U.S. patents.

Patrick Wetterwald is a thought leader in architecture and standardization for the smart grid, industrial sensor networks, and Internet of Things (IoT). He held various engineering manager positions within Cisco where he successfully led advanced technology projects in the domain of wireless sensor networks, wireless communication, layer 3 mesh, and IPv6 network mobility. Wetterwald is an initial founder of the IPSO Alliance, where he served for five years as the president of the Board of Directors. He has written several articles on behalf of the Alliance and delivered many keynotes at international events. He is leading and participating to several standardization efforts for ETSI, IETF, IEC, ITU, European Mandates and EC expert groups. He is now focusing on Industrial Automation technologies including Deterministic Networking. Before joining Cisco, Wetterwald spent the last 25 years in the telecommunication industry working for Lucent Technologies, IBM and Airbus Industry. He graduated from the Ecole Nationale Supérieure des Télécommunications de retagne. He filed 86 patents and got (PMP) certification in 2001.

**DOI:** 10.1145/2822875

## References

Andy Noronha, Robert Moriarty, Kathy O'Connell, and Nicola Villa. Attaining IoT Value: How to move from Connecting Things to Capturing Insight. White paper. Cisco. 2014.