TECHNICAL REPORT

ISO/IEC TR 23188

First edition
2020-02

Information technology — Cloud computing — Edge computing landscape

Reference number
ISO/IEC TR 23188:2020(E)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Fax: +41 22 749 09 47
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of document should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents) or the IEC list of patent declarations received (see http://patents.iec.ch).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT) see www.iso.org/iso/foreword.html.

This document was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 38, *Cloud computing and distributed platforms*.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

Edge computing is increasingly used in systems that deal with aspects of the physical world. Edge computing involves the placement of processing and storage near or at the places where those systems interact with the physical world, which is where the "edge" exists. One of the trends in this space is the development of increasingly capable Internet of Things (IoT) devices (sensors and actuators), which generate more data or new types of data. There is significant benefit from moving the processing and storing of this data close to the place where the data is generated.

Cloud computing is commonly used in systems that are based on edge computing approaches. This can include the connection of both devices and edge computing nodes to centralized cloud services. However, it is the case that the locations in which cloud computing is performed are increasingly distributed in nature. The cloud services are being implemented in locations that are nearer to the edge in order to support use cases that demand reduced latency or avoiding the need to transmit large volumes of data over networks with limited bandwidth.

This document aims to describe edge computing and the significant elements which contribute to the successful implementation of edge computing systems, with an emphasis on the use of cloud computing and cloud computing technologies in the context of edge computing, including the virtualization of compute, storage and networking resources.

It is useful to read this document in conjunction with ISO/IEC TR 30164[1)] [27], which takes a view of edge computing from the point of view of IoT systems and the IoT devices which interact with the physical world.

1) Under development. Current stage 10.99.

Information technology — Cloud computing — Edge computing landscape

1 Scope

This document examines the concept of edge computing, its relationship to cloud computing and IoT, and the technologies that are key to the implementation of edge computing. This document explores the following topics with respect to edge computing:

— concept of edge computing systems;

— architectural foundation of edge computing;

— edge computing terminology;

— software classifications in edge computing, e.g. firmware, services, applications;

— supporting technologies, e.g. containers, serverless computing, microservices;

— networking for edge systems, including virtual networks;

— data, e.g. data flow, data storage, data processing;

— management, of software, of data and of networks, resources, quality of service;

— virtual placement of software and data, and metadata;

— security and privacy;

— real time;

— mobile edge computing, mobile devices.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 22123-1:—[2], *Information technology — Cloud computing — Part 1: Terminology*

ISO/IEC TS 23167, *Information technology — Cloud computing — Common technologies and techniques*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO/IEC 22123-1, ISO/IEC TS 23167 and the following apply.

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at http://www.iso.org/obp

— IEC Electropedia: available at http://www.electropedia.org/

2) To be published.

3.1 Edge computing

3.1.1
distributed computing
model of computing in which a set of *nodes* (3.1.5) coordinates its activities by means of digital messages passed between the *nodes* (3.1.5)

3.1.2
edge
boundary between pertinent digital and *physical entities* (3.2.8), delineated by networked *sensors* (3.2.9) and *actuators* (3.2.1)

Note 1 to entry: Pertinent digital entities means that the digital entities which need to be considered can vary depending on the system under consideration and the context in which those entities are used. See 5.2 for more detail.

3.1.3
edge computing
distributed computing (3.1.1) in which processing and storage takes place at or near the *edge* (3.1.2), where the nearness is defined by the system's requirements

3.1.4
lightweight node
node (3.1.5) with limited processing, storage and networking capacities

3.1.5
node
networked machine with processing and storage capabilities

3.1.6
edge computing system
system providing functionalities of *edge computing* (3.1.3)

3.1.7
endpoint
combination of a binding and a network address

[SOURCE: ISO/TR 24097-3:2019, 3.4]

3.2 IoT terms

3.2.1
actuator
IoT device (3.2.4) that changes one or more properties of a *physical entity* (3.2.8) in response to a valid input

[SOURCE: ISO/IEC 20924:2018, 3.2.2]

3.2.2
Internet of Things
IoT
infrastructure of interconnected entities, people, systems and information resources together with services which processes and reacts to information from the physical world and virtual world

[SOURCE: ISO/IEC 20924:2018, 3.2.1]

3.2.3
Internet Protocol
IP
protocol specified in RFC 791 (IP version 4) or in RFC 2460 (IP version 6)

[SOURCE: ISO/IEC TR 21890:2001, 3.4]

3.2.4
IoT device
entity of an *IoT system* (3.2.6) that interacts and communicates with the physical world through sensing or actuating

Note 1 to entry: An *IoT device* (3.2.4) can be a *sensor* (3.2.9) or an *actuator* (3.2.1).

[SOURCE: ISO/IEC 20924:2018, 3.2.4]

3.2.5
IoT gateway
entity of an *IoT system* (3.2.6) that connects one or more proximity networks and the *IoT devices* (3.2.4) on those networks to each other and to one or more access networks

[SOURCE: ISO/IEC 20924:2018, 3.2.6]

3.2.6
IoT system
system providing functionalities of *Internet of Things* (3.2.2)

Note 1 to entry: IoT system is inclusive of *IoT devices* (3.2.4), *IoT gateways* (3.2.5), *sensors* (3.2.9), and *actuators* (3.2.1).

[SOURCE: ISO/IEC 20924:2018, 3.2.7]

3.2.7
operational technology
OT
hardware and software that detects or causes a change through the direct monitoring and/or control of physical devices and systems, processes and events in the organization

3.2.8
physical entity
entity that has material existence in the physical world

[SOURCE: ISO/IEC 20924:2018, 3.1.26, modified — Note 1 to entry has been removed.]

3.2.9
sensor
IoT device (3.2.4) that measures one or more properties of one or more *physical entities* (3.2.8) and outputs digital data that can be transmitted over a network

[SOURCE: ISO/IEC 20924:2018, 3.2.9]

3.3 Real time

3.3.1
real time
processing of data by a computer in connection with another process outside the computer according to time requirements imposed by the outside process

[SOURCE: ISO/IEC 2382:2015, 2122900, modified: words 'pertaining to' removed to improve substitutability of definition; Notes 1 to 3 to entry have been removed.]

3.3.2
real time system
system in which processing meets *real time* (3.3.1) requirements

3.3.3
hard real time system
real time system (3.3.2) whose operation is incorrect if results are not produced according to specified timing requirements

3.3.4
soft real time system
real time system (3.3.2) whose operation is degraded if results are not produced according to specified timing requirements

4 Symbols and abbreviated terms

AC	Alternating current
ASIC	Application-Specific Integrated Circuit
BYOD	Bring Your Own Device
CDN	Content Distribution Network
CSC	Cloud service customer
CSP	Cloud service provider
DDoS	Distributed Denial of Service
EPG	Electronic Programme Guide
EPROM	Erasable Programmable Read Only Memory
FPGA	Field Programmable Gate Array
Gb	Gigabyte
GPS	Global Positioning System
GPU	Graphics Processing Unit
IETF	Internet Engineering Task Force
IoT	Internet of Things
IP	Internet Protocol
IPTV	Internet Protocol television
LAN	Local Area Network
MDM	Mobile Device Management
OS	Operating system
PC	Personal Computer
PII	Personally Identifiable Information

RAM Random-access Memory

RFC Request for Comments

ROM Read Only Memory

TPM Trusted Platform Module

VM Virtual Machine

VoIP Voice over Internet Protocol

VPN Virtual Private Network

5 Overview of edge computing

5.1 General

Over time, the forms of computing have varied between centralized and distributed, depending on the nature and capabilities of the computing devices and of the networks used to connect them.

Mainframe computers represent a form of centralised computing, where the main computer systems are placed in a data centre, containing processing and storage units. Originally, almost the whole of the computing system was situated within the data centre. Gradually, time-sharing terminals were located in remote locations to provide user access to the mainframe systems. Terminals were typically little more than a display with a keyboard for input and the associated network connection had limited bandwidth, perhaps involving a dial-up modem.

The personal computer (PC) represents a distributed form of computing. The PC has significant processing and storage capabilities and can be used very effectively in a standalone mode. However, PCs are more typically used in a networked mode. Initially, the networks were used for simple communications such as (text based) email, but as the network bandwidth increased over time, increasingly sophisticated activities took place, with file transfer and eventually peer-to-peer capabilities being used.

The availability of higher bandwidth networking encouraged the development of the client-server application architecture, with the PC used for the client, connected to a centralized server which performs the main processing and storage of the application. Client applications can include quite substantial software elements performing significant processing activities. Data might also be stored locally for faster access, although the main database(s) are held centrally.

The advent of the internet and the World Wide Web (WWW) represents the appearance of another form of computing. In this form of computing, web servers serve up web pages and related material which are accessed through client web browsers. Devices running web browsers can be relatively low in compute power, while the web servers for some of the more popular and high demand web sites can involve massive compute power spread over many machines in a large data centre.

Cloud computing is a computing paradigm that makes available all types of computing resources in an on-demand, highly scalable fashion via cloud services. Cloud computing in practice is made possible through a highly centralised architecture, with computing resources concentrated in large data centres. However, cloud computing in practice also has some distributed computing features. It is typical for cloud service providers (CSPs) to have multiple physically separated data centres and cloud service customers (CSCs) commonly distribute their applications and data across multiple data centres – for resilience, to reduce latency and for disaster recovery purposes. In addition, the favoured design paradigm for cloud native applications is to distribute multiple instances of each application component across different machines within the cloud computing system. This design paradigm and the technologies that support it are of significance to edge computing.

5.2 Concepts of edge computing

Edge computing is distributed computing in which data processing and storage takes place on nodes which are near to the edge. The edge is marked by the boundary between pertinent digital and physical entities, i.e. between the digital system and the physical world, delineated by networked sensors and actuators.

Pertinent digital entities means that the digital entities which need to be considered can vary depending on the system under consideration and the context in which those entities are used.

An example of varying pertinence are the servers within a cloud computing data centre. From the perspective of CSCs building systems using cloud services running on these servers, these IoT devices are anything but "at the edge". However, from the perspective of the CSPs having to manage the cloud computing data centre, it is highly likely that the servers are instrumented with a variety of sensors capable of reporting various physical properties of the servers, for example their temperature. Those sensors are at the edge.

The concept of nearness to the edge also needs explanation. Nearness for edge computing is usually based on minimising the latency for communication between the IoT devices that are at the edge and the place(s) where data processing and storage occurs. Nearness can mean placing the edge computing nodes physically close to the IoT devices. In the most extreme cases, nearness means combining the sensors and actuators and edge computing into a single node, as might happen with a smart phone. In other cases, the edge computing nodes are separated from the IoT devices but are placed physically close to the IoT devices and have a proximity network connecting them designed to minimise latency. Nearness can also be influenced by the nature of the networks and the volume of data flowing to and from the IoT devices – where large volumes of data and high data rates are involved, edge nodes are placed so as to reduce the latency of handling this data to the minimum necessary to meet the requirements of the use case.

Digital systems can observe and affect the physical world. Sensors and actuators are at the edge between the digital systems and the physical world. Edge computing systems generally combine these IoT devices with distributed computing resources to provide the capabilities of the system. In edge computing systems, actions often need to occur within specific timeframes, i.e. edge computing systems can also be real time systems, and latency considerations affect system design and the choice of the placement of data processing and storage to achieve timing requirements. Edge computing helps to meet those timing requirements.

Edge computing is characterized by networked systems in which significant data processing and storage takes place on nodes near the edge, rather than in some centralized location. Edge computing can be contrasted with centralized computing where the centralized nodes are remote from the edge. However, it is important to note that edge computing is complementary to centralized forms of computing and that in any given system, edge computing is often used in conjunction with centralized computing.

There are multiple reasons for the rise in the use of edge computing. One reason is the arrival of new devices combining significant processing power and storage with low power usage. Smart phones have been one of the driving factors in this area, with billions of such devices in daily use. The Internet of Things (IoT) is another reason, with small, low power, low cost IoT devices enabling the creation of IT systems which can sense and act on real world entities.

5.3 Architectural foundations of edge computing

Edge computing involves nodes that are highly heterogeneous and which are commonly arranged in tiers of compute and storage capabilities. A simplified view of the organization of edge computing nodes and the networks connecting them in edge computing is shown in Figure 1.

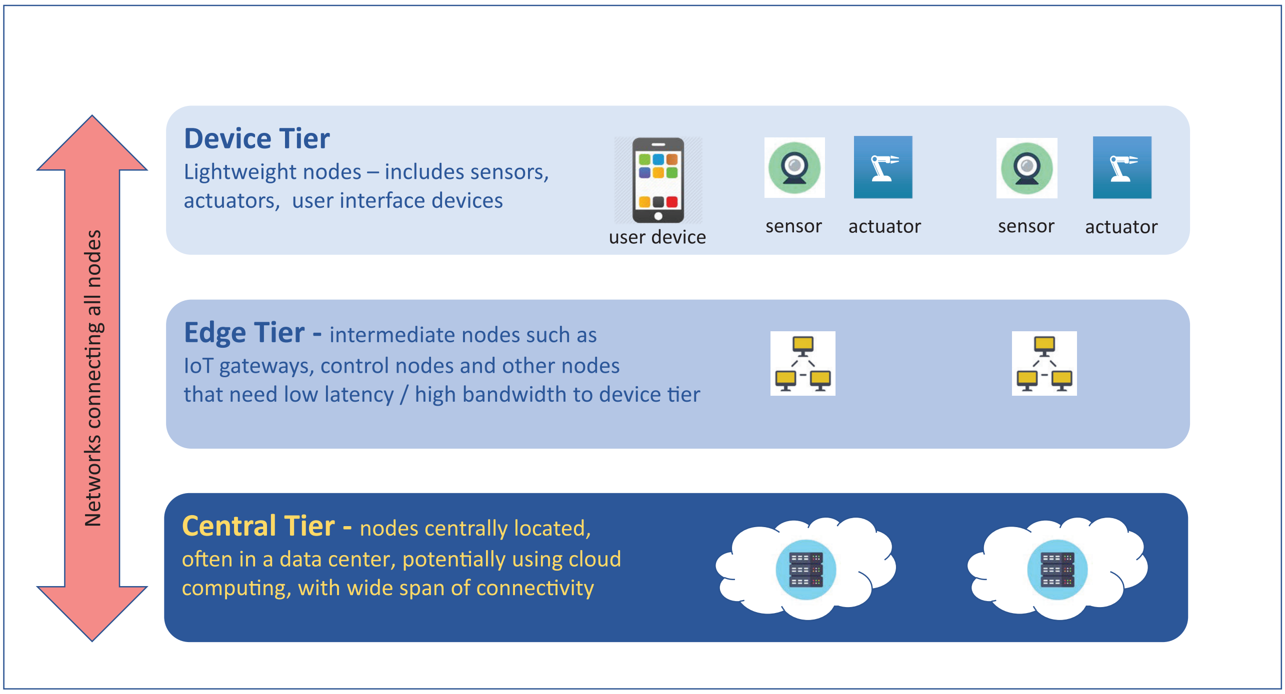

Figure 1 — Organization of nodes in edge computing

The tiers shown in Figure 1 are essentially a conceptual model (containing physical elements) and are illustrative rather than definitive – in reality the number of tiers and the type of node in each tier and the networks connecting them are variable, depending on the nature of the system involved. What is important to understand is that there are multiple tiers, containing varying types of nodes, all connected by networks which can also vary in nature depending on the tiers involved.

The device tier is at the edge. It typically contains lightweight nodes which commonly contain sensors or actuators or user interface devices. Such devices often have limited compute and storage capabilities. The networks used by this tier are often proximity networks, with limited bandwidth and limited range (see 6.1.1 in this document and ISO/IEC 30141:2018[2], 10.2.3.2 and 10.4.1.2 for more detail about proximity networks).

The edge tier typically sits near to the device tier (where "near" is a relative term and depends on the particular system and use case) and its role is to provide direct support to the nodes in the device tier. One type of node in the edge tier is the gateway node, for which an IoT gateway is one example. The role of the gateway node is to interconnect proximity networks to IP-based wide area networks. This may involve message and protocol syntax and semantic conversions. This role could include message encryption, deduplication and backup functions.

Another type of node in the edge tier is the control node. The control node receives data from nodes in the device tier – typically data from sensors or input from user interface devices – and responds by issuing instructions to other nodes in the device tier. Other types of node may be placed in the edge tier to meet other edge computing functional requirements. This may include management nodes, security nodes and software support nodes.

Control nodes are usually placed in the edge tier due to issues of latency and timing. The response of a control node is often time constrained (sometimes called real time – see Clause 14), such that the response must be given before some deadline following the receipt of some data or an event. One factor in this time constraint is the transmission time of messages to and from nodes in the device tier – this leads to the need for the nodes in the edge tier to be placed physically close to the device tier nodes and to the need to reduce the number of hops that the messages must take. These constraints can also influence the type of proximity network used and the protocol used over those networks. Similarly, the control nodes must have appropriate processing capacity and storage for the processing that is necessary to produce appropriate and timely responses.

As an example, if the input data is a video stream from a camera device, which is a type of sensor, and the processing required is an analysis of the video to detect the movement of some object with the intent of influencing the movement via some actuators (which are different devices from the camera), this could take a substantial amount of processing power and also require the handling of a substantial amount of data – the control node must have appropriate processing power and storage to successfully carry out its task.

The central tier represents a tier of nodes provided by centralized facilities. The nodes in the central tier offer the ability to provide very substantial compute power and storage. The central tier is an excellent place to conduct analytics or other processing that requires both a lot of compute power and access to a lot of information. The central tier can hold large stores of information which can come from many sources – this can be from across the other tiers or from outside locations, potentially sourced from other organizations. The central tier can provide services to the other tiers, including services for processing data in various ways or for holding information or providing information as required. The central tier usually has a wide span of connectivity, meaning that it is commonly connected to many other systems, including many of the distributed nodes in both the edge tier and in the device tier.

It is often the case that the central tier is implemented using cloud computing. A fuller description of the relationship of edge computing to cloud computing is given in 5.4.

It is typical of the nodes in the central tier to communicate to the other tiers using high bandwidth networks, typically the internet but possibly dedicated networks. It is also possible for the nodes in the central tier to be arranged in a highly available resilient configuration, with multiple instances of applications and services allied to replicated or redundant copies of information.

The tiers described in Figure 1 can become blurred when considering the many different types of device that are available. A significant example is the smartphone, which combines a number of elements into a single device, as follows:

- a number of sensors of various types, including GPS (location sensor), accelerometer, barometer, health monitors (such as heart rate monitoring);
- camera – both for static images and video;
- microphone & loudspeaker for audio;
- display screen and user interface;
- significant compute power (e.g. quad or octo core systems, 2 to 4 Gb RAM);
- significant local storage (e.g. up to 256 Gb);
- networking and connectivity.

These capabilities in a single device span the device tier and the edge tier and provide for dynamic addition and update of software on the device, enabling a very wide range of capabilities. Combined with excellent networking capabilities, smartphones enable some forms of edge computing in their own right – with the added advantage of their being mobile.

5.4 The relationship of edge computing to cloud computing

Edge computing can exist on its own, without any relationship to cloud computing. In terms of the tiers described in 5.3, systems can exist in which cloud computing is not used in any of the tiers. This implies that the system has no need of the capabilities offered by cloud computing. Older industrial systems are of this nature – designed to be self-contained and with fixed functionality.

However, for many systems cloud computing or cloud computing technologies are used in one or more of the tiers. This is especially so for the central tier – it is very common for the nodes in the central tier to be part of a cloud computing environment, either a public cloud or a private cloud. Cloud computing can also be used in the edge tier – as more powerful and lower-cost hardware becomes available, it is increasingly possible and desirable to use cloud computing in the edge tier. Although today use of cloud computing in the device tier is unusual and rare, there is nothing in principle to preclude its use once the device nodes are sufficiently powerful and well-connected.

It is worth noting that edge computing rarely exists on its own, but is connected to both processing and information which is held in the central tier. This means systems have processing and storage spread right across the various tiers and types of nodes. The principle is the right placement of processing and storage elements. Right placement in that processing and storage take place on nodes that are best suited to the task involved.

Figure 2 illustrates how both IoT and edge computing can relate to different parts of the cloud computing ecosystem. This can include cloud services built on a private cloud (both on-premises and more remotely) and a public cloud (including public clouds designed to serve a specific jurisdiction, or multinational or even global cloud services). Hybrid clouds of various kinds can also be employed according to business needs, as in the example shown in Figure 2 where a public cloud in the central tier is combined with a private cloud in the edge tier.

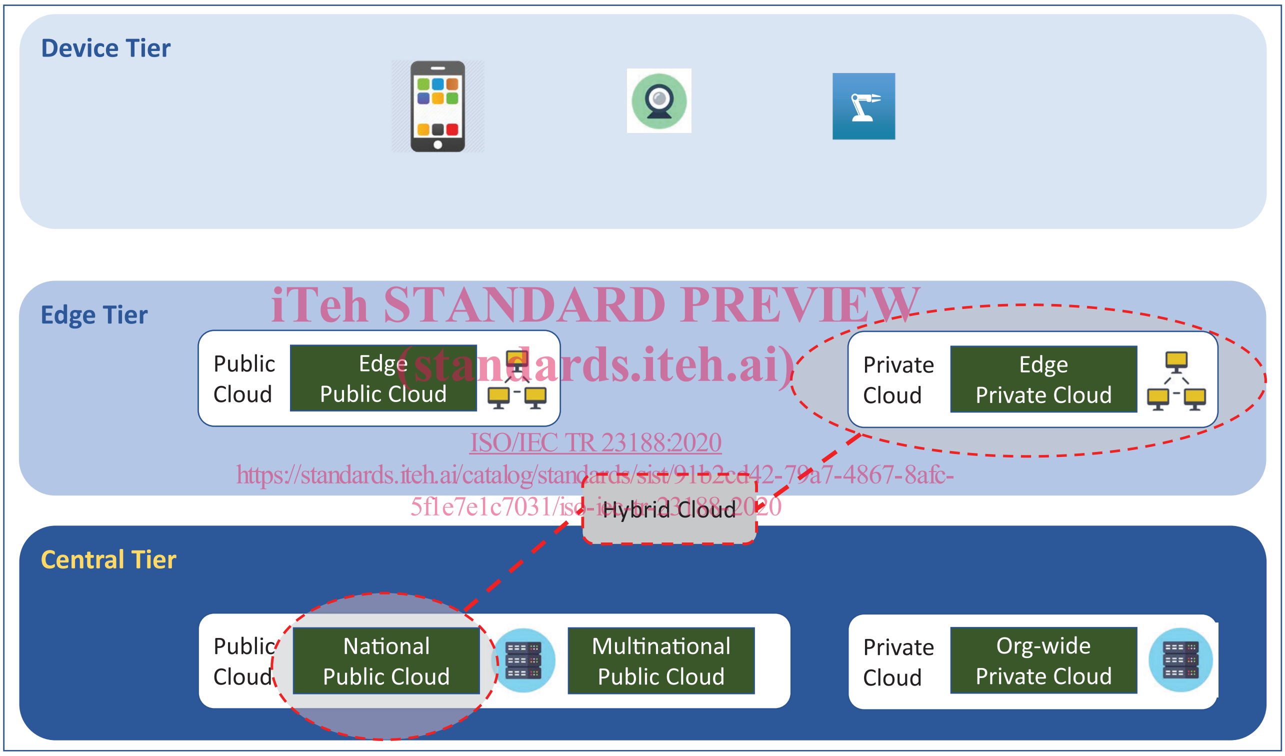

Figure 2 — Relationship of edge computing to cloud computing

The central tier can be implemented using public cloud services, and these can be implemented using either a multinational public cloud (i.e. multiple data centres in a number of jurisdictions) or a national public cloud (for example, where there are restrictions on where the data can be stored and processed). The central tier can also be implemented using an enterprise-wide private cloud.

The edge tier can be implemented using an on-premises private cloud, physically located to suit the needs of the edge system. However, it is the case that some public clouds make available cloud services on a more physically localised basis, potentially suitable for edge computing scenarios. Compute capabilities in cellular telephone towers are an example. Such distributed public cloud offerings could be used for the edge tier. As an example in this latter case, the device tier nodes could be directly connected using mobile phone networks to nodes running in cellular telephone towers.

As an example, to reduce latency and achieve timing goals, it could be necessary to place control processing on edge tier nodes, where those nodes are physically close to the device tier such as a co-located on-premises private cloud. However, for processing that involves analysis of large volumes of data that arise from many different sources, it is likely that centralized storage is the best approach,

