CHAPTER

DATA CENTER FACILITIES

4

This chapter (1) defines the data center, (2) provides brief information on how to plan data center capacity, (3) discusses the site selection process if the business decision is to build new data center facility, (4) provides performance metrics for a data center, and (5) describes the details of data center space and the cost estimation that underpins one of three pillars (space, power, and cooling) of the data center facility cost foundation (see Figure 4.1).

4.1 BASIC UNDERSTANDING OF A DATA CENTER

Before discussing the technical details of a data center facility, the first few questions that we may encounter are: What is a data center anyway? How does it matter to the cloud computing solution or cost modeling? What are differences between a conventional data center and a cloud data center?

To answer these questions, let us look at the data center from a historical perspective. Actually, there have been many names for a data center. We can summarize various names into four different perspectives (see Table 4.1).

4.1.1 DEFINITION OF DATA CENTER

Surely, some names describe certain data center functions. Others just mean different sizes of data center in terms of capacity. Still others are just spelled differently. All these names are right if one is just focusing on one aspect of data center function but they are not very comprehensive. Eric Bauer et al. [85] tried to define this term comprehensively:

> A data centre is a physical space that is environmentally controlled with clean electrical power and network connectivity that is optimized for hosting servers. The temperature and humidity of data centre environment are controlled to enable proper operation of the equipment and the facility is physically secured to prevent deliberate or accidental damage to the physical equipment. This facility will have one or more connections to the public Internet, often via redundant and physically separated cables into redundant routers. Behind the routers will be security applications, like firewalls or deep packet inspection elements, to enforce a security perimeter protecting servers in the data centre. Behind the security appliances are often load balancers which distribute traffic across front end servers like web servers. Often there is one or two tiers of server behind

Cloud Data Centers and Cost Modeling. DOI: http://dx.doi.org/10.1016/B978-0-12-801413-4.00004-0

the application front end like second tier servers implementing application or business logic and a third tier of database servers. Establishing and operating a traditional data centre facility – including IP routers and infrastructure, security applications, load balancers, servers' storage and supporting systems – requires a large capital outlay and substantial operation expenses, all to support application software that often has widely varying load so that much of the resource capacity is often underutilised.

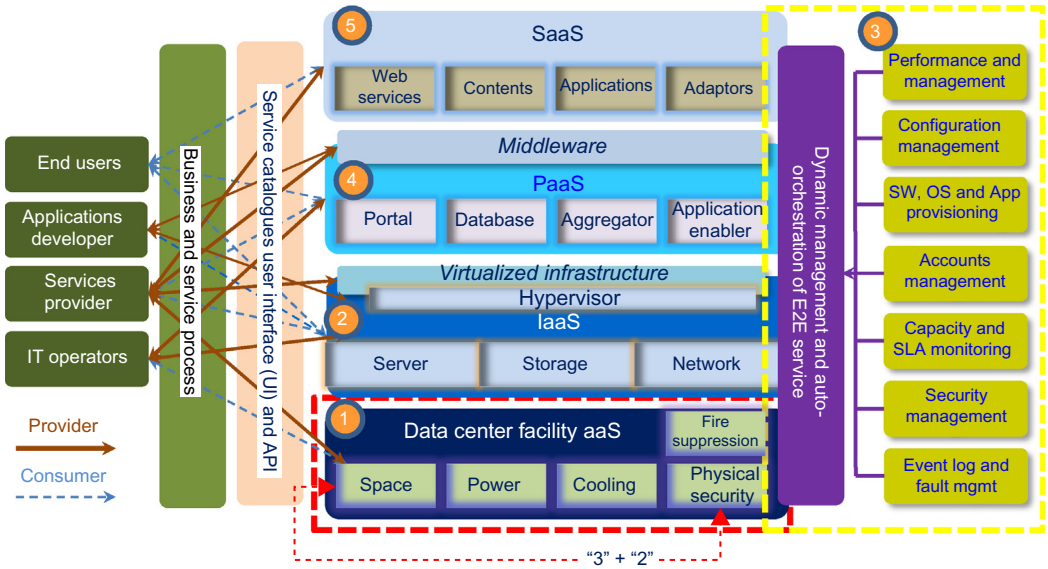

FIGURE 4.1

The big picture of the cloud cost framework.

Table 4.1 Different Data Center Names

Engineering Perspective	Communication Perspective	Storage Perspective	Computer Perspective
Mechanical rooms	Wiring closet	Network storage room	Server room
Electrical rooms	LAN rooms	Storage space	Computer room
	Network closet or room	Backup room	Data center
	Telecommunication room	Disaster recovery site	Datacentre
	Console room or Network operation center		

Eric Bauer and colleagues' definition of data center is quite comprehensive. It includes data center space, power, cooling, physical and virtual security, IT infrastructure (server, network connection, storage, and load balance), IT applications, redundancy, data center operations, and management. The term does not only include hosting servers but also a mainframe. However, it seems to be too long.

Maurizio Portolani et al. [86] gave the definition of the data center from a network perspective as follows:

> Data centers house critical computing resources in controlled environments and under centralized management, which enable enterprises to operate around the clock or according to their business needs. These computing resources include mainframes; web and application servers; file and print servers; messaging servers; application software and the operating systems that run them; storage subsystems; and the network infrastructure, whether IP or storage-area network (SAN). Applications range from internal financial and human resources to external e-commerce and business-to-business applications. Additionally, a number of servers support network operations and network-based applications. Network operation applications include Network Time Protocol (NTP); TN3270; FTP; Domain Name System (DNS); Dynamic Host Configuration Protocol (DHCP); Simple Network Management Protocol (SNMP); TFTP; Network File System (NFS); and network-based applications, including IP telephony, video streaming over IP, IP video conferencing, and so on.

Maurizio et al.'s term is a relatively narrow definition. It only emphasizes the functionality of hosting and networking. Perhaps this is because the majority of existing data centers are very small, what we often call a micro data center or computer room. Normally, they only accommodate a few racks. It may just utilize existing comfort cooling or the office air conditioning environment rather than a precision cooling facility. They are only to host a few applications and websites.

In order to serve the purposes of this book, we define a data center facility as follows:

> The data center is a place where can accommodate many computing resources that collect, store, share, manage, and distribute a large volume of data. It consists of all necessary data center facility elements (space, power, and cooling) and IT infrastructure elements (server, storage, and network) based on business requirements.

A data center can be varied from a micro data center with a few servers to a warehouse scale that can accommodate thousands and even millions of racks that are configured with thousands or millions of servers. This is dependent on the particular business requirements or needs. Table 4.2 illustrates that different sizes or types of business may need different size data center facilities.

Table 4.2 Business Size and Data Center Scale Correlation

Business Size	No. of Employees	Data Center Size	No. of Racks
Micro Business	1–4	Office base	None
Small Business	5–19	Computer room	None or 1–5
Medium Business	20–200	Small or medium DC	5–20 or 20–100
Large Business	More than 200	Medium or large DC	20–100 or >1,000
Huge Enterprise Business	More than 10,000	Mega DC	>10,000

4.1.2 DATA CENTER ARCHITECTURE

Once the size of data center has been decided, the next step is to plot out the data center architecture. A typical contemporary data center should have nine basic function rooms or areas to support data center facility services (see Figure 4.2).

These nine basic function rooms are:

- Entrance Room
- Main Distribution Area (Computer Room)
- Telecom Room
- Mechanical Room
- Electrical Room
- Network Operation and Support Room (or NOC)
- Staging Area, Storage Room, Loading Dock
- Common Areas
- General Office

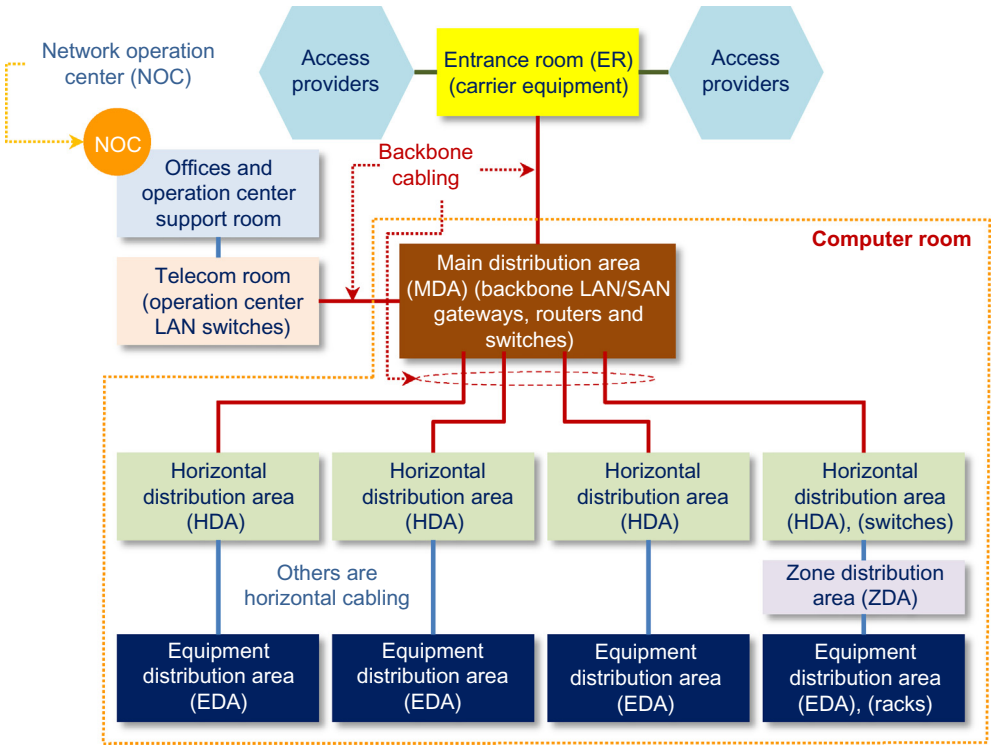

FIGURE 4.2

TIA-942 data Center topology (2005).

Each function room should provide unique features for the data center infrastructure. The architecture of these functional rooms is basically based on the TIA-942 standard data center topology. Before we further discuss these nine functional rooms in the section on data center space, we have to decide on data center capacity first. This exercise is a part of data center planning. In order to have the right data center architecture, it is important to have the right data center plan in terms of capacity. This plan will not only satisfy the existing business requirements but also fulfill future business needs.

4.2 DATA CENTER CAPACITY PLANNING

Business requirements will decide the data center workload, capacity, and tiering. It does not matter what size the business is. Whether it is a small, medium, large, or mega size business, the goal of data center capacity planning is to build or find the right data center facility for the business activity. It should not only meet the normal or existing business activities or operations but also accommodate the growth volume of future business demands. It is a long-term investment.

If we are talking about future growth, forecasting would be an inevitable step for our planning process. By human nature, we are not very good at predicting the future, particularly when we encounter the distant future, because there are so many uncertainties and variables. As a common idiom says, "I do not know what the future is, anything is possible." The longer the period of time the more inaccurate the prediction may be. It is quite obvious. We may be able to estimate the outlook of a business in 10 months but it would be very challenging to forecast the future business demands in 10 years.

However, as John Allspaw indicated, "You wouldn't begin mixing concrete before you know what you are building. Similarly, you shouldn't start to build a data center facility before you determine your data center's capacity. Capacity planning involves a lot of assumptions related to why you need the capacity. Some of those assumptions are quite obvious, others are not." [87] Actually, only very few assumptions may be quite obvious or direct and the majority of them are uncertain or fuzzy. Many assumptions involve subjective judgments or guesses. Others may have to be defended by certain probabilities, especially when we add the time domain for a longer period.

Why is this so? One of reasons is our subjectiveness in how we understand the world, as Mortimer J. Adler described: "Men are creatures of passion and prejudice. The language they must use to communicate is an imperfect medium clouded by emotion and coloured by interest, as well as inadequately transparent for thought" [89]. This means that the decision or judgment that we make is often highly subjective and based on personal perception or opinion. Very often, we see the world through an incomplete or even incorrect information medium.

Another reason may be that just as the German scholar, Reinhard Selten indicated, we only make the decision or judgment to the certain point of own satisfaction rather than to be rational [90]. In other words, when we make assumptions, we might not have enough time or patience to find a rational or perfect answer. An assumption could be the truth, but it is ultimately just what people think. Until it is proved by the time, the assumption may also be false or incorrect.

The last reason is probability. An assumption may be true at this moment but it might not be true at another moment. We are facing a dynamic world. Nothing holds still. When a prediction or assumption is made, we often defend it with the certain probability, which measures or estimates

the likelihood of the assumption being true. Unless the probability is 0% or 100%, for a particular event, anything may occur. It actually makes the world much more interesting. If everything were predictable, the world would become very boring.

In essence, we should always be aware of the nature of assumptions when we make a series of assumptions for both the near term and future status of data center capacity. Assumptions are assumptions. They are not facts. Good future assumptions are those that you can probably defend with a definable certainty or probability. If the assumption cannot be defended by probability, it should be explicitly spelled out. Once we understand the nature of assumptions for capacity planning, we should be able to make the assumptions of capacity planning much more objective or scientific for our cost modeling exercise.

If we look from a major cost components' perspective at the data center facility, there are three key cost categories. These cost components should consist of more than 90% of the total data center facility costs:

- Space
- Power
- Cooling

Moreover, there are also four additional mandatory cost components that we should take into consideration for a TIA-942 standard tier data center facility:

- On-site security
- Fire suppression system
- Cable and cabling
- Racks or cabinets

How can we predict these key cost components when we are planning the capacity of the data center? We can probably adopt some common analytic methodologies, such as the cost benefits analysis (CBA), benchmark analysis, SWOT, agile, Delphi, and analytic hierarchy process (AHP) methods. (In addition to the common approaches, there are many other analytic approaches to define the cost. Readers can refer to Chapter 14 and see Appendix E for the details). We can also use these methodologies or processes in combination. It is really dependent on the phase of data center planning. In addition, we can also leverage different mathematic tools, especially statistical analysis or simulation techniques, to predicate future status. The ultimate goal is to eliminate subjective opinions and to have more scientific evidence so that we can use this evidence to support our assumptions and future forecasting results.

When we are talking about capacity planning for a data center facility, we may face one of the following seven scenarios due to different business circumstances:

- Building a brand new data center
- Purchasing or selecting colocation space
- Expanding the existing data center capacity
- Consolidating the existing data centers
- Moving or migrating current business applications into the cloud
- Building a business continuity facility
- Data center relocation

There may be other possible scenarios, but the above seven scenarios should be sufficient for us to discuss the cost models or framework. Once we understand these scenarios, we should be able to translate the business requirements into two series of key parameters for different scenarios:

- Performance of data center facility
  - Service Level Agreement (SLA)
  - Customer experiences
  - User or stakeholder expectations
- Resource capacity of data center facility
  - Resource ceilings
  - System capability
  - Reliability and availability

The purpose of data center planning is to make sure that all these parameters will be aligned with business requirements. One of the common approaches to specify these parameters is the so-called SMART approach recommended by Armando Fox and David Patterson [91]. SMART stands for:

- **S**pecific
- **M**easurable
- **A**chievable
- **R**elevant
- **T**ime boxed

"Specific" means each business requirement should be explicit, which implies that the expected result of the business requirement is measureable in dollar value. "Measureable" means the metric can be measured by a specified value, time, or dollar term. Here is an example of a specific and nonspecific metric:

- "The business requires that the hosting data center should be very reliable." This requirement is too simple, nonspecific, and unmeasurable.
- In contrast, the business requirement should state: "The hosting data center must exceed 99.95% availability, the number of unplanned outages in one month should be less than 0.01%, impacted customers should be less than 50; if an unplanned outage occurs, the impacted customers should be not more than 100, the response time should be less than 5 minutes, and the resolution time should be less than 2 hours." This requirement is specific and measurable.

"Achievable" means the result would be realistic based on the current technology level. For example, the business asks for a guarantee of no unplanned outages for a large-scale data center facility within the next 12 months. This may be an unachievable target for a tier 1 data center. The data center facility service (DCaaS) provider has to also review its database infrastructure and understand its data center capability in terms of meeting business demand.

"Relevant" means the business requirements must have business value to one or more stakeholders (data center facility's investor or sponsor). Suppose the business asks you to build a data center facility that requires a backup generator; one of the techniques to drill down to the real

business value (also suggested by Armando Fox and David Peterson) is to keep asking five "why" questions. For example:

1. Why does the business require a generator? Because the business needs a very reliable hosting data center facility.
2. Why does high reliability matter to the business? If the data center is down, the business will lose significant sales revenue.
3. Why will the downtime impact the sales revenue? Because all CRM clients connected to the hosting server will be down as well.
4. Why will the hosting server be down? If we lose power in the data center.
5. Why will the data center lose power? If the power supply is disrupted and we do not have a generator.

Figure 4.3 is a simple flowchart to ask several "why" questions to determine whether the data center needs a generator or not. This process will drive the business value for both customers and stakeholders.

"Time boxing" means that you know where to stop if the business requirements have exceeded time and technology capabilities and/or budget limit. Of course, it is possible to divide a large amount of business requirements into a number of manageable and deliverable tasks or functions and then go back to stakeholders and fully communicate with them and get their approval for further resource (time, capex, and third-party support) commitment.

Once we understand how to effectively identify both performance parameters and resource metrics for data center facility, it is easier to follow the process of determining the data center capacity during the planning phase (see Figure 4.4).

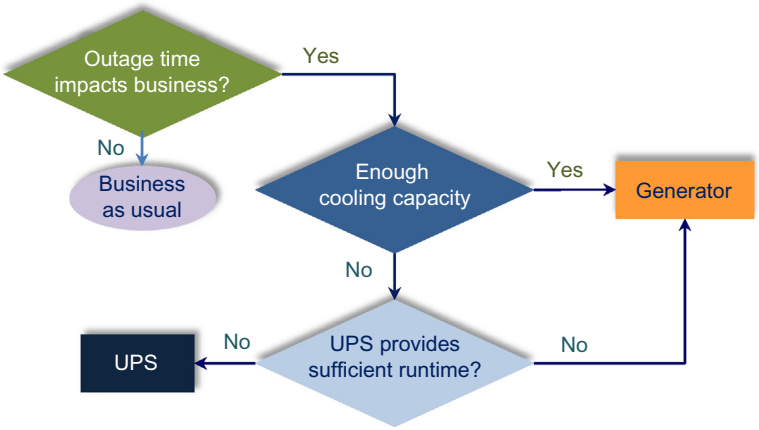

FIGURE 4.3

Determine if a generator is required.

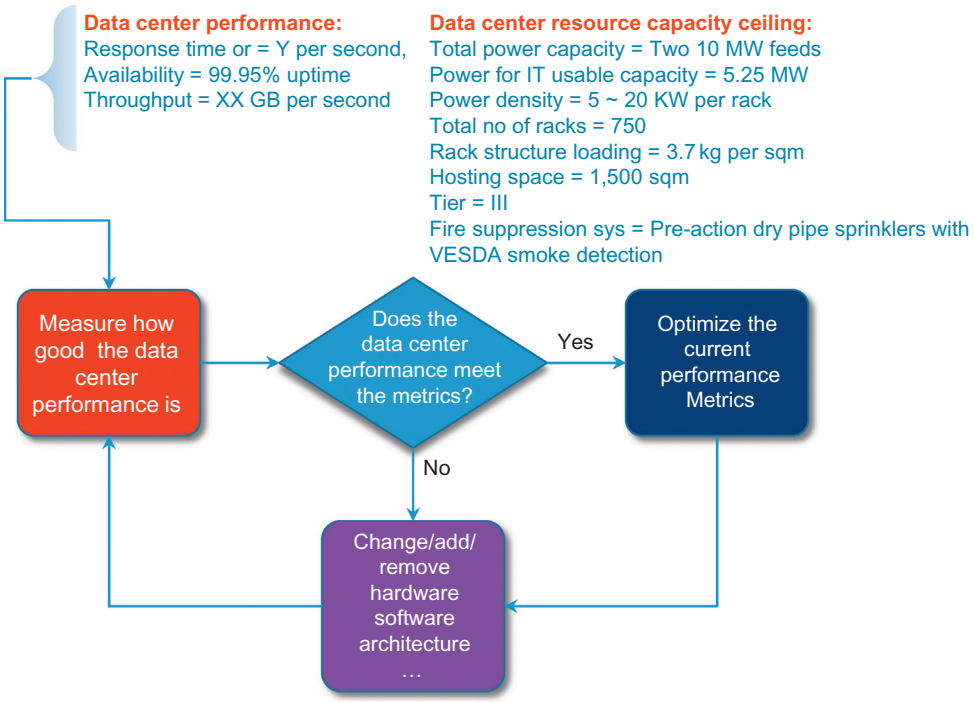

FIGURE 4.4

A simple process for data center planning.

When we implement the above process, there are a few important questions that should be asked. The art of capacity planning is dependent on asking the right questions in the right order. There are four main types of questions that should be asked:

1. What is strategic goal of the business? You must figure out the strategic goal of the business and how this strategic goal will be implemented in orderly way by subdividing it into a number of tactical objects.
2. What is being described by stakeholders in detail? As a data center facility planner you must discover the main facts and assumptions along with defendable probabilities to support the strategic goal and tactical objects within a specified time frame.
3. Is the defined strategic goal achievable in whole or in part? Surely, you cannot answer this question until you have answered the first two. You have to know what is being defined before you can decide whether the goal is achievable or not. If it is not achievable, you have to work it out in a different way.
4. How significant is the strategic goal? You have to understand why the goal is so important to the stakeholders. What is the business value in both explicit and implicit dollar terms?

Here is a typical example of asking the right questions in the right order for data center capacity planning:

1. What is your sales revenue target for the hosting service business in three years? How can you achieve it in each year or each quarter?
2. What are the main assumptions for your target? For each quarter's sale target, what is the probability, most likely, likely, or unlikely?
3. Is the defined sales revenue realistic? What are the risks in term of the current market conditions?
4. How significant is it for the company to achieve this strategic goal? What is the dollar value?

If the business is not new and just asking for business expansion, the second level of questions should be as follows: Does the existing data center facility meet the current business demand? If so, how well it is working? If not, why not? Do we need to expand the existing data center capacity? If the workload is growing at x% per annum for existing infrastructure, will it maintain the current performance? If not, what is the right strategy to expand the data center capacity?

- Build a brand new data center?
- Purchase colocation?
- Expand the existing capacity?
- Move a part of workload to a cloud?
- Consolidate the existing infrastructure?

By investigating like this, we will be able to make the cost of the data center facility align with the business goal and objects and the cost of data center will become much more transparent to both customers and stakeholders. Subsequently, it will also drive the unit cost of the data center facility to be more competitive.

Finally we have to point out the above assumptions we have made here are based on a so-called linear process model or single point of revenue prediction. It is the baseline estimation plus some standard deviation. We will have further discussion on this topic regarding a nonlinear model of revenue predication in Part V of this book.

Once we have fully understood how to effectively translate or interpret what the business wants, namely the strategic goals and the process of data center capacity planning, the next step is to examine the contents of data center facility planning in detail, which will include three critical elements:

- Data center site selection
- Data center performance
- Data center resource ceiling

4.2.1 DATA CENTER SITE SELECTION

A data center facility will be strategic assets or critical infrastructure for any enterprise that owns the infrastructure, especially for a business that provides Data Center as a Service (DCaaS) or Infrastructure as a Service (IaaS). The selection of a data center at the right site or location will be absolutely critical if the business is to be competitive in the marketplace. It will not only impact on return of the capex investment but also operational efficiency. It will ultimately impact on the unit cost of DCaaS and IaaS.

Table 4.3 Australian Data Center as a Service Catalog [92]

Service Categories	No. of Cloud Services	Percentage
IaaS	858	54%
PaaS	105	7%
SaaS	225	14%
Managed services	350	22%
Others and blank	51	3%
Total	1,589	100%

The cloud market is very competitive. Based on the latest Australian government's DCaaS service catalog (Department of Finance and Deregulation, version 5.1, releases date: June 26, 2014) there are 104 cloud service providers that supply more than 1,589 cloud services; nearly 54% of them are IaaS, 22% of them are managed services, and only about 21% of them are either PaaS or SaaS (see Table 4.3). Within this DCaaS catalog, some big global players, such as Amazon AWS and Rackspace, are not present. However, they have formed different alliances with different companies and provide IaaS for the Australian government. In short, they have already gained their foothold in Australia, where they have built a few data centers, especially in Sydney and Melbourne.

In comparison with the United States, United Kingdom, or even HongKong (see Appendix C) [93], the Australian cloud market is relatively small. For such a small market, over 104 providers and 1,589 cloud products is quite a crowd. In addition, the number of cloud service providers is not fixed. It is growing. This is good news for cloud consumers but for many cloud services providers, it means increased data center competitiveness and a need for efficiency in terms of building and operating DCaaS and IaaS. We believe the efficiency of data center infrastructure will be at the heart of the competition for many cloud service providers. However, when you make a decision either to build a new data center or to select one of the DCaaS providers, two primary issues must be considered. They are data security and data center efficiency. It doesn't matter what the size of the business is, whether it is small, medium, or large, even if the business is searching for colocation capacity. The bottom line of site security and efficiency will ultimately determine how to select the data center site. A decision maker has to implement a due diligence process for data center site selection.

When we implement a due diligence process for data center site selection, John Rath [94] suggests two primary factors to be examined:

- Natural disasters (mainly weather) and unnatural disasters
- Workforce resources and business climate

In addition to these two primary factors, John Rath also believes that a decision maker should also examine many different categories of issues. Some of them may be very special. However, the first thing would be the natural disaster and weather issue. When we talk about natural disasters, we can probably categorize these events into three types:

- Climatic
- Geological
- Hydrological

A climatic disaster event would include a blizzard, cyclone, hurricane, tornado, severe hail, excessive rain, heavy snow, ice and/or high wind, an electrical storm, or a severe weather pattern sustained over a period of time including very low or very high temperature. A geological disaster event will be events such as earthquakes and volcanoes. Hydrological disaster events will be tsunamis and floods. Despite these disaster events, other disaster events may be triggered by climatic, geological, and hydrological disaster event or events that are a combination of these natural disasters. The typical types of these disasters are:

- Fire/wildfire
- Avalanche
- Landslide/mudslide
- Drought

In contrast to natural disaster, some unnatural disasters events should also be taken into consideration for data center site selection, such as terrorist attacks and pandemics.

There is another way of making assessment of data center location, which was proposed by Karl F. Rauscher et al. [95] from Lucent Technologies. They suggested an eight-ingredient model. It is presented from a business continuity perspective (see Figure 4.5). The eight ingredients are environment, power, payload, policy, human, network, hardware, and software. Let's walk through these ingredients in detail.

4.2.1.1 The environment

Within the generic framework of the ingredients required for a data center facility to operate, the environment is the foundation. This means the geographic location of the data center. It should include climatic, geological, and hydrologic factors, which we have mentioned above. In addition, the site environment should also include the following five elements:

- Average temperature pattern
- Aa central location to all major customers, relatively
- Close to a power plant that provides the power

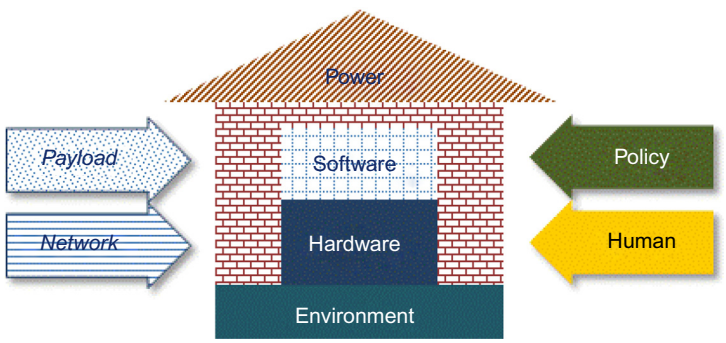

FIGURE 4.5

Eight-ingredient model.

- Many dark fibers
- Site has potential space for data center expansion

The main reason of including the average temperature pattern is to maximize the benefit of free cooling by leveraging economizers (for details, refer to Chapter 9) so that we can reduce the operating costs of the data center facility.

The meaning of having a central location to all major customers is that a data center should be close to the sources of major traffic or data streams so that the majority of customers can backup and retrieve data easily. For example, if all major customers are located in Sydney, it would be a good idea to allocate a data center in Sydney rather than Melbourne or Tasmania.

In order to reduce power transmission loss, a data center should be allocated as close as possible to the power plant. This will increase a data center's Power Usage Effectiveness (PUE) factor.

The dark fiber is another very important issue to consider. When you select the potential data center location, you would prefer the location already has many dark fibers that are ready via different carriers so that you can obtain a competitive price for the traffic of the data center network (see Figure 4.15 for the network ingredient).

During the last one year or so, unstructured data is growing dramatically at an exponential rate. According to IDC and EMC's forecasting in 2012, the digital universe will grow 50 fold (from 800 EB to 40ZB) between 2010 and 2020 (see Figure 4.6) and the latest IDC study in 2014 indicated that by 2020, digital data will reach 44 ZB (IDC updates this study every year http://idcdocserv.com/1678). This means that data growth is accelerating. To experience the real data growth, we can visit a live statistics site on the Internet [97]. Therefore, it is important to consider future growth for data center expansion.

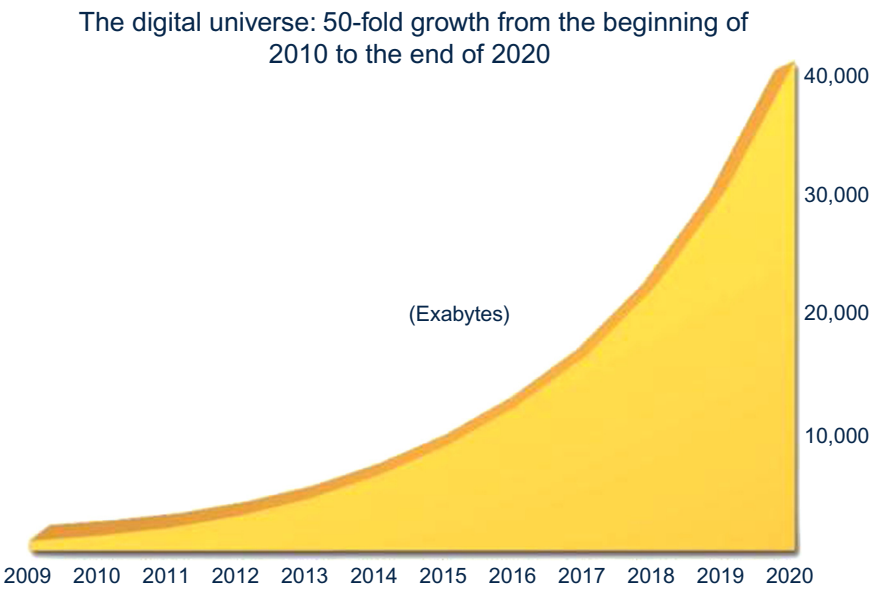

FIGURE 4.6

Data growth between 2010 and 2020 [96].

4.2.1.2 The power

Power means the external power supply from a power plant. For internal data center power distribution, we will give more details in the next chapter. When considering the data center location, the price of the power should be one of the key factors. It should be very competitive, reliable, and sufficient for future growth capacity. For example, when Priest Rapids Dam started supplying abundant cheap and renewable energy in Quincy, Washington (WA), it bought many high profile data centers to Quincy (see Figures 4.7 and 4.8).

Based on Colo & Cloud's information [100], the dam provides hydroelectric power at a price even lower than Rock Mountain Power, which was claimed to be the lowest price in the United States. The commercial power price for a data center is around US$0.025 per kWh. Microsoft reports its price is only about US$0.019 per kWh.

In contrast, the average price of electricity in Australia was around A$0.25/kWh in 2011–2012, which is one of the highest prices in the world according to the report by the Energy User Association of Australia (EUAA) (see Figure 4.9).

FIGURE 4.7

Priest Rapids Dam and hydroelectric project [98].

FIGURE 4.8

High-profile data centers in Quincy, WA, USA [100].

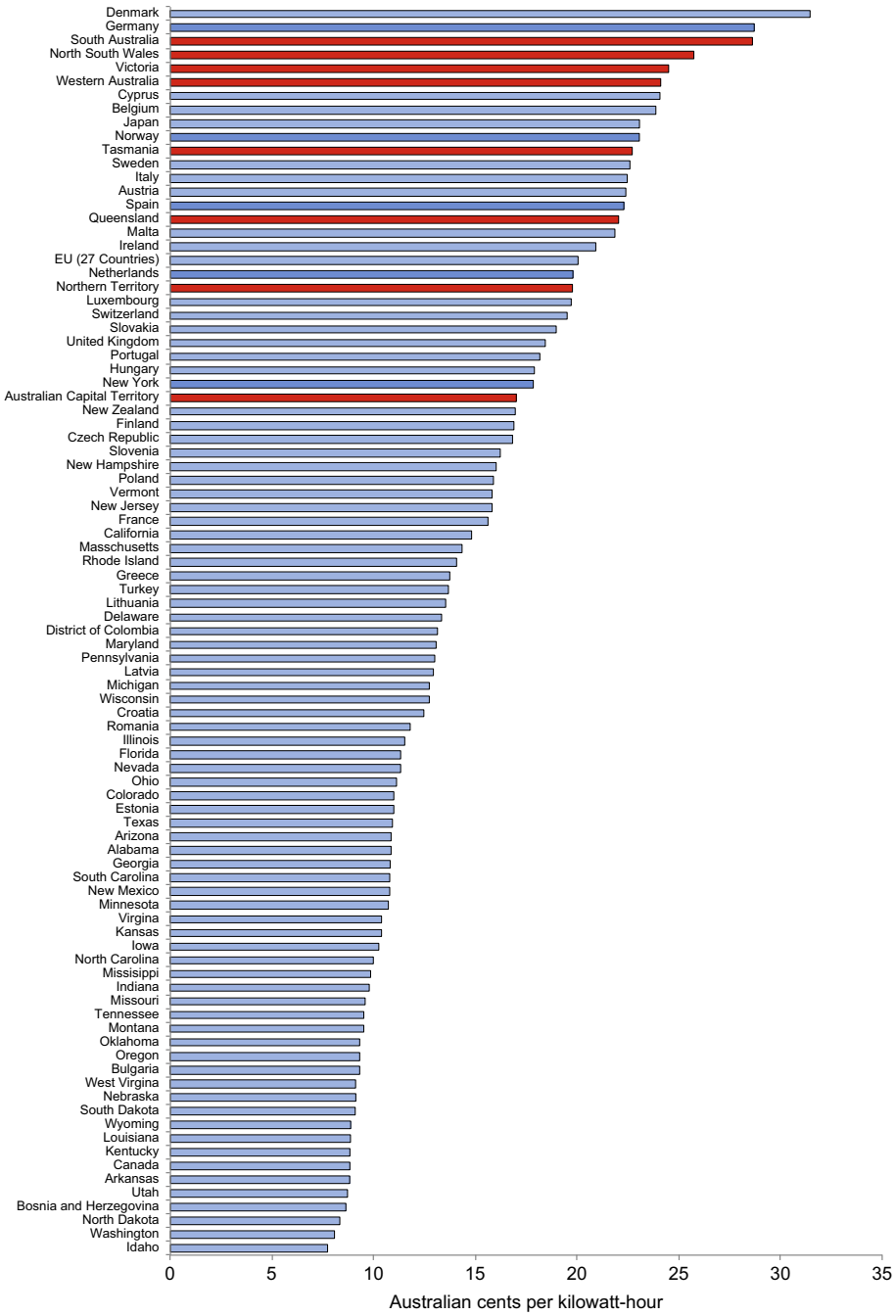

FIGURE 4.9

Electricity price by country, state, and province [103].

This is about ten times higher than Quincy, WA, USA. Of course, this comparison is not very accurate because the retail price is different from the commercial or the wholesale price. However, the power price is Washington is still the second cheapest in the EAAC price list. Normally, the commercial price would be lower than the retail one.

Based on the above EAAC's figure, if you would like to select a new data center location in Australia and just consider the power price factor alone, the favorable state would be ACT (around A$0.17/kWh) and the worst state is South Australia (around A$0.28/kWh). However, we have to be aware of the population density of the data center location. If the location of the data center is far away from a higher population density area, for example Sydney and Melbourne in Australia, the power consumption cost for network distribution may exceed the cost savings for power.

4.2.1.3 The payload and IT workload

If the price of the external power supply is uncontrollable, then the alternative to reducing the power supply cost is increasing the data center efficiency, which is the workload. The power usage efficiency is dependent on accurate prediction of data center workload because the higher the utilization of the data center the higher the efficiency is. Normally, the average operation cost for a data center will be five times more than its capex because the average life cycle of a data center facility is over 20 years. In other words, the opex is more than 25% of capex per annum. (This opex should include a 12.5% maintenance and support cost for many hardware components.) The optimal power supply load level is between 40% and 60% capacity (see Figure 4.10). It is dependent on what type of PDU is used.

Ideally, the data center should be operated at 100% (see Figure 4.11) for IT workloads. The higher the IT load, the better the data center efficiency is.

4.2.1.4 The policy

This refers to a local government policy that hopefully encourages data center facility investment. It is any kind of understanding and agreement between governments and data center facility companies. It can also be the government agreements, standard, policies, and regulations (ASPR). It may consist of a tax break for a high-tech data center facility that is built locally or incentives from by federal and local governments. The government policy could be considered a part of the business climate.

4.2.1.5 The human factor

This is the human capital. For example, when Google built one of its mega data centers in Council Bluffs, Iowa (see Figures 4.12–4.14) in 2007, Google was not only looking for clean energy, namely hydroelectric power, very close to the power source, but also looking to establish a partnership with nearby education institutions. Since 2009, Google has awarded more than $665 k to local schools in Council Bluffs, Iowa. Additionally, Google also partnered with City of Council Bluffs to launch a free WiFi network.

4.2.1.6 The network

This means the availability of dark fiber for a potential data center location. When the workload of a data center is growing, the demand for dark fiber capacity will be absolutely critical for data center scalability. This is because the demand for security enhancement and more bandwidth or traffic between the data center and head office or operations center will be increased. Moreover, if the data

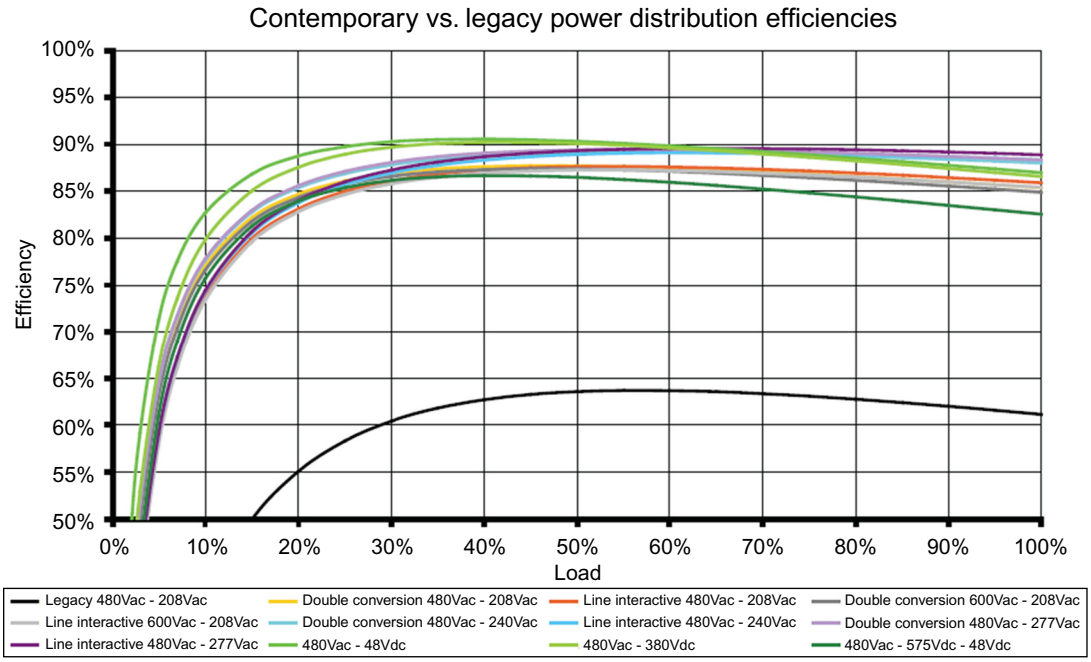

FIGURE 4.10

Power distribution or power supply efficiencies vs. workload [104].

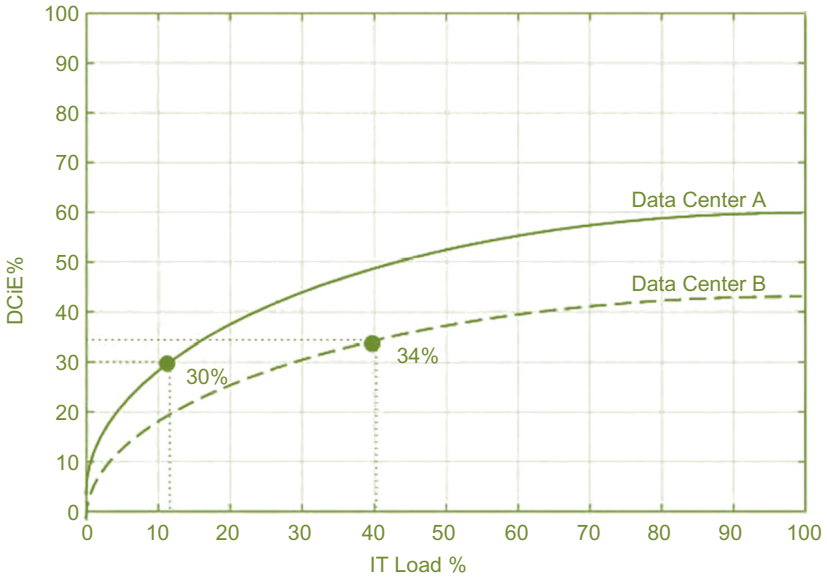

FIGURE 4.11

Data Center Efficiency (DCiE) metrics as a function of IT load [105].

FIGURE 4.12

Cooling towers of Google data center in Council Bluffs, Iowa [106].

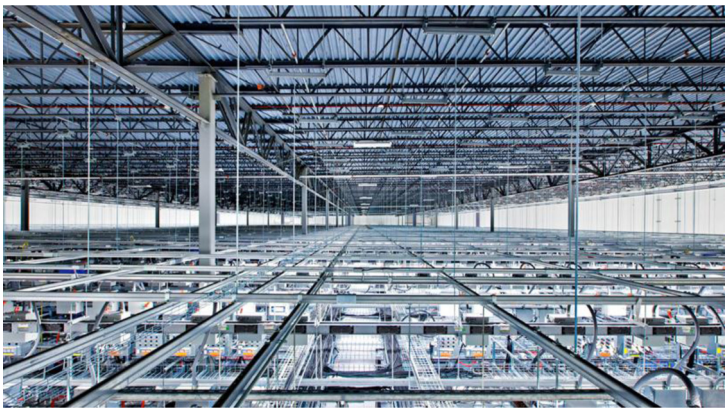

FIGURE 4.13

Google's data center hovering above the floor in Council Bluffs, Iowa [107].

center workloads are designed for mission critical applications, the data center should not only be connected to disaster recovery sites with geographic diversity but also link to multiple backup data centers.

Taking the example of the Quincy, WA, area, the connectivity is provided by many wholesale dark fiber providers such as NoaNet and Grant County PUD built the fiber network. Some carriers, such as Zayo, Level3, Verizon, and Frontier have purchased capacity from them in order to offer their services to the data center facilities (see Figure 4.15).

In the US, state or local governments will point you in the right direction for dark fiber wholesale carriers. Of course, some companies such as Fibrelocator.com (http://www.fiberlocator.com/)

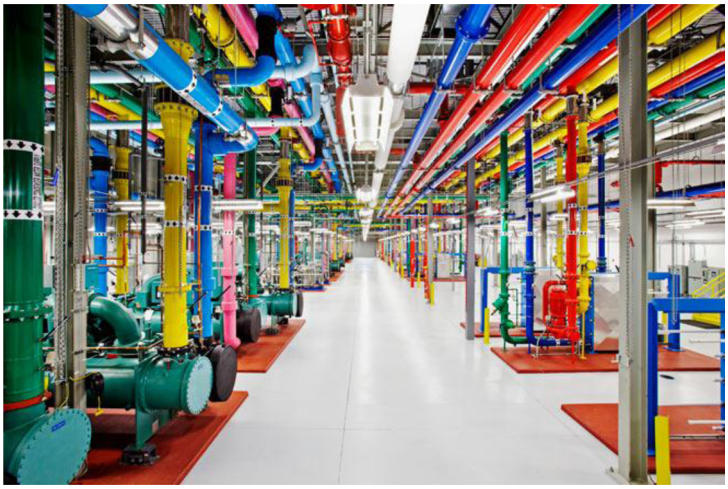

FIGURE 4.14

Google's data centers use color-coded pipes to indicate what they're used for. Referring to Google's URL, you can see all the details of color-coded pipes (Council Bluffs, Iowa) [108].

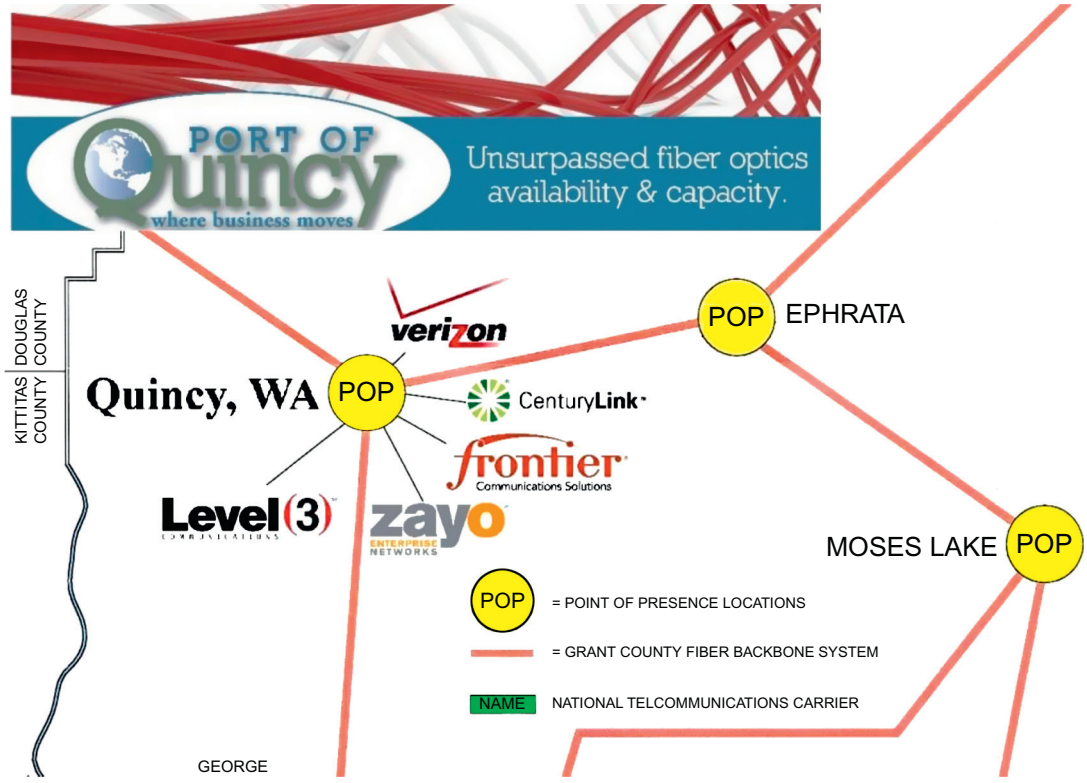

FIGURE 4.15

Dark fiber connectivity in Quincy, WA [100].

and CFN Services (http://www.cfnservices.com/) will provide research services concerning allocation of dark fibers.

In Australia, a new company established in 2008, Vocus Communications, may be able to provide information about dark fiber and connectivity for data centers both in Australia and New Zealand. The traditional carriers, such as Telstra, Optus, and AAPT, will provide their wholesale services for optical fiber network connectivity across Asia or the world.

In terms of data center hardware and software ingredients, we will give details in the following chapters.

The selection of a data center site is not only one of the essential processes for building a new data center but also a critical step in the selection of colocation and disaster recovery sites, migration of an existing data center to a cloud, consolidation of a number of data centers, and even moving legacy data centers to a new data center. It holds the future for many companies.

As a rule of thumb, the life cycle of a data center is about 20 years. The ongoing opex cost will be five times more than the initial capex investment. Data center site selection cannot only be based on the traditional criteria that are either just being close to head office or based on decision makers' intuitions. It should be vigorously analyzed via a proper process, such as the eight ingredients methodology. Once the site is selected, then we can work out the data center performance, which we are going to have a close look at in the following.

4.2.2 DATA CENTER PERFORMANCE

What does performance mean? Actually, "performance" has many different meanings, such as "an act of presentation" or to "play" a role in a movie, or "executing a task" or "fulfillment of a promise," etc. Here, the meaning of performance means "quality of function" against an agreed and measurable standard or metrics.

In Section 4.2, we briefly touched on the topic of data center performance. Here we will give more details about data center performance based on the business requirements or needs.

Data center performance means a set of metrics. It should measure all aspects of the data center facility, which may include capability, high availability, reliability, scalability, manageability, latency, response or resolution time, throughput, security, disaster recovery or business continuity, resource utilization rate, energy efficiency or carbon footprint, fault tolerance, return on investment (ROI), and total cost of ownership (TCO).

At a glance, there are too many performance parameters to be measured. Which one is the most important? Which one is the least important? It is really dependent on who you are and what kind of business you are running or what kind of business applications you are hosting. From a cloud customer's perspective, data center performance may mean:

- Data center availability or reliability (cost of downtime)
- Security or data protection
- Throughput
- Latency
- Scalability
- Problem response and resolution time
- Business continuity or disaster recovery

For a cloud service (IaaS) provider, data center performance should include all aspects of measurements from the data center facility to the hosted applications and from cloud computing resource utilization to ROI/TCO. Let's have a look at the following critical performance metrics that will impact significantly on both data center service consumers and providers in terms of cost:

- Site availability
- Problem response and resolution time
- Scalability
- Utilization rate
- Latency and throughput

4.2.2.1 Site availability

From an end-to-end perspective, the most important performance measurement that every business will pay attention to is the system or data center availability. According to the NSI/TIA-942 standard (see Appendix H), one of the key metrics is site availability. It differentiates four tiers of data centers. This can be translated into a resource allocation, in terms of construction cost per square meter.

For traditional dedicated hosting services, we know that the application and hardware infrastructure is not shared. It is a one-to-one relationship. Subsequently, if one hardware component fails, it will not impact on other applications or customers. In contrast, the cloud environment is built on a virtual infrastructure or shared hardware. An issue with one piece of hardware may impact on many applications. In other words, one problem with a hardware device could lead to not only one but many applications that have performance degradation issues.

Therefore, if you are a cloud service provider and host many applications for your end users, availability does not only mean to one end user but also to all of your end users, so the availability measurement should include the number of end users. For example, a cloud service provider may claim that it can guarantee 99.95% availability (see Table 4.4). Does this mean to all end users or just 99% of end users per month or per annum. When a cloud service business customer signs a Service Level Agreement (SLA) with a cloud service provider, the service contract should be specified very clearly because it may impact on response and resolution time.

Table 4.4 Summary of Availability Information of ANSI/TIA-942 for Data Center Tiering

Attributes	Tier I	Tier II	Tier III	Tier IV
Power & Cooling Delivery Paths	1 active	1 active	1 active + 1 passive	2 active
Redundant Components	N	N + 1	N + 1	2(N + 1)
Support Space to Raised Floor Ratio	20%	30%	80 ~ 90%	100%
Annual IT Downtime due to Site	28.8 hours	22.0 hours	1.6 hours	0.4 hours
Site Availability	99.671%	99.749%	99.982%	99.995%

4.2.2.2 Problem response and resolution time

From a cloud service operator's perspective, a cloud problem or a fault that impacts on the quantity of end users will decide the severity level of the fault. Consequently, it will trigger different levels of response and resolution time. For example if a cloud fault impacts more than 100 end users, the response time should be less than 5 ~ 10 minutes and the resolution time should be within 2 hours. However, if the fault just impacts one end user, the response and resolution time will be much longer, such as a 2 hour response time with an 8 hour resolution time.

Very often, the service operator may use an operational metric to measure the availability with different levels of support. These metrics are mean time to repair or recover (MTTR) and mean time to failure (MTTF). There is a relationship between availability, MTTF, and MTTR:

$$Availability = \frac{MTTF}{MTTF + MTTR}$$

Here the MTTR is equal to the resolution time. Based on the reliability engineering theory, we often use another metric to measure system reliability, which is mean time between failure (MTBF). The relationship between MTBF, MTTF, and MTTR is equal to:

$$MTBF = MTTF + MTTR$$

To support the above availability, a cloud service may use different service levels, such as Platinum, Gold, Silver, Bronze, on call, normal business hours, etc. The price for different SLAs will vary from very expensive to very economic. Many managed service providers adopt this metric to differentiate their products and services in the cloud market. Table 4.5 shows an example of different cloud service support levels.

4.2.2.3 Scalability

From a cloud customer's perspective, cloud scalability means a cloud's elastic characteristics or flexibility. In other words, the end user only pays for what he/she has consumed of the cloud resources. The cloud resources cannot only be quickly scaled up but also be rapidly shrunken down.

Table 4.5 An Example of Cloud Operation Support Levels

Support Type	Support Time	Response Time	Resolution Time	No. of Impacted End Users
Platinum	24 × 7	<5 min	<2 hours	>100
Gold	24 × 7	<10 min	<4 hours	>50
Silver	7 × 21 × 7	<30 min	<8 hours	>5
Bronze	7 × 21 × 5	<45 min	<24 hours	>2
On Call	Call based charge	<60 min	To be decided	To be decide
Normal	Normal business hours excluding public holidays	<4 hours	Next business day	<1

From a cloud service provider's perspective, scalability means to accommodate various users' demands. This means that the cloud service provider does not only have the capability to scale its resource pool but also to scale it up.

In order to increase the cloud infrastructure utilization rate during resource idle times, the cloud service provider has to have a different price structure. For example, Amazon's EC2 and S3 services include a so-called "Spot Instance" price model. Others, such as Salesforce.com adopt the subscription price model, from which the cloud service provider can forecast or predicate the potential resource demand.

The question may be raised, "Despite having different price models to regulate cloud consumer's behavior during idle times, how is it possible for a cloud service provider to accommodate peak demand?" The simple answer is hardware virtualization and software multitenancy. Based on Artur Andrzejak et al.'s [101] research across six different corporate data centers, the typical utilization rate for a traditional hosting server is around 10% ~ 35%. Will Forrest et al.'s [52] discussion paper from Mckinsey also indicated that a typical data center's utilization rate is around 10%. Gartner's report [39] suggested that over 50% or 60% of the storage capacity would be idle in a typical company's data center. Our data also shows that the average utilization rate for some of hosting servers (mobile contents) was below 1% and the peak utilization rate was only around 7%.

Consequently, there will be enough infrastructure headroom for a cloud service provider to work with and still make enough profit while accommodating cloud customers' demands or workloads. The only question is how to optimize the resource pool and to synchronize each infrastructure component, including power, cooling, space, server, storage, and network. This is the topic that we are going to discuss in the next few chapters.

4.2.2.4 Utilization

The utilization rate is the key financial performance index when measuring the cloud. Theoretically speaking, the higher the utilization rate is, the lower the cloud infrastructure cost is. The fundamental reason that many enterprises move their IT workloads to a cloud environment is not to share IT resources or utilize advanced cloud technologies. It is because the existing standalone or traditional IT infrastructure is heavily underutilized. A cloud service provider can bring cloud services to its customers and increase the IT utilization rate significantly. Therefore, the cloud service provider can return cost benefits to its customers while still maintaining healthy profits.

As we will see in the later chapters, the cloud infrastructure utilization rate will be a critical factor in cloud cost modeling. It doesn't only require that the cloud service provider utilize proper capacity planning based on the business and IT workload profile but also requires a cloud service provider to build a data center in a modular fashion.

Before we move on the next topic of data center capacity, let's have a look two more key concepts of data center metrics, which are latency and throughput.

4.2.2.5 Latency and throughput

Latency and throughput are two very important performance metrics for any end user. These two performance indicators will measure all other data center performance metrics when a cloud service is online. Based on Adrian Cockcroft et al.'s [102] definition for Internet services, latency stands for "the time that it takes to complete a well-defined action" and throughput means "the

number of defined actions performed in a given period of time." In very simple terms, the latency is the execution time and the throughput is the required bandwidth to execute a task.

To illustrate these two concepts of performance clearly, we will use both airplane and bullet train transportation as examples. Suppose an airplane that can carries 100 passengers per flight has to transfer 20,000 people from one airport to another and each flight would take 60 minutes per one way flight[1] (just ignore the time for loading and unloading passages). The question is, "How long would it take to complete this mission?" (See Figure 4.16.)

- Airplane capacity = 100 per flight
- Latency = 20,000/100 × 60 × 2 = 24,000 minutes or 400 hours to complete the mission
- Throughput = 20,000/400 = 50 people per hour (PPH)

If I use a bullet train to transfer the same amount of people (or 20,000) and each bullet train's capacity = 10,000 per trip and the time from one place to another destination would take 240 minutes or 4 hours,[2] what is the latency for the bullet train? (See Figure 4.17.)

- Bullet train capacity = 10,000 per trip
- Latency = (20,000/10,000) × 240 × 2 = 960 minutes = 16 hours to execute this mission
- Throughput = 20,000/16 = 1,250 people per hour (PPH)

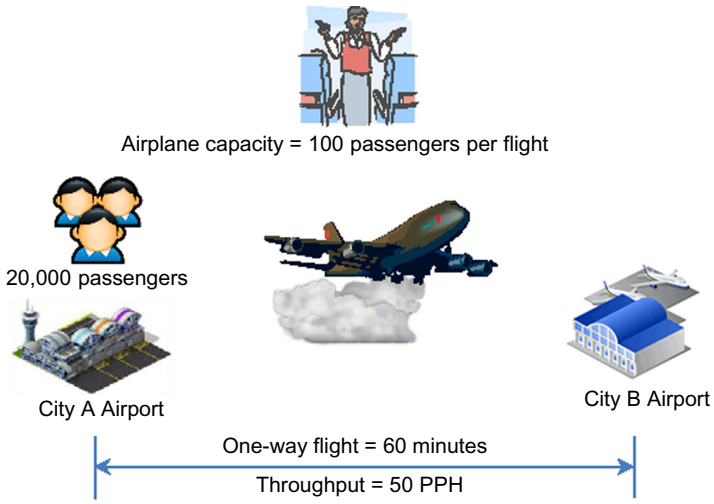

FIGURE 4.16

An example of airplane latency, throughput, and capacity.

[1]Assume that the average passenger airplane's speed = 800 km per hour.
[2]Assume that the average bullet train's speed = 200 km per hour.

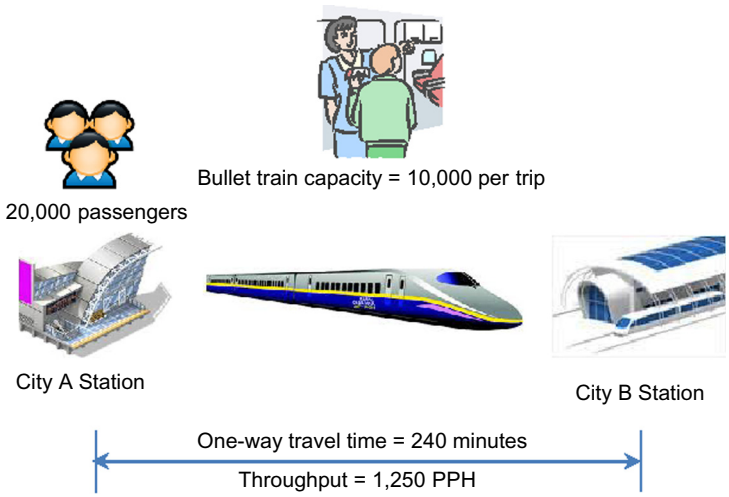

FIGURE 4.17

An example of bullet train latency, throughput, and capacity.

In comparison, the airplane is 240/60 = 4 times faster than the bullet train. Although the bullet train is slower in terms of latency, its throughput is 1250/50 = 25 times larger than the airplane.

Once we understand these two concepts, it would be easier for us to apply these concepts to the application of a high-speed link between two data centers.

Suppose we have a file of TCP window size = 1 MB (megabyte) to be transferred from data center A (in city S) to data center B (in city M). The core network connection between the two data centers is a 10GbE link with an average round-trip latency (RTT) of 20 milliseconds. Now, the question is, "What is the maximum throughput I can get? (See Figure 4.18.)

- 1 megabyte window size = 1024 × 1,024 × 8 = 8,388,608 bits
- 8,388,608 bits/0.020 second = 419,430,400 bits per second throughput, which means the maximum possible throughput = 419 Mbps

Here, the example shows that although the pipe size is a 10GbE link, the maximum possible throughput is only 419 Mbps. So, how can we increase the throughput for the given 10GbE? The answer would be to either increase the TCP window size or reduce the RTT latency. This is similar to the story of the airplane or bullet train, where you can either increase the number of passenger per flight (from 100 passengers to 200 passengers per flight) or reduce round-trip time, by increasing the airplane speed (from 120 min to 100 min).

If the pipe size and latency are fixed, what is my maximum throughput?

- The pipe size is a 10 GbE link and round-trip time (RTT) = 20 ms.
- 10,000,000,000*0.02 seconds = 200 MBits/8 = 25 megabytes per second.

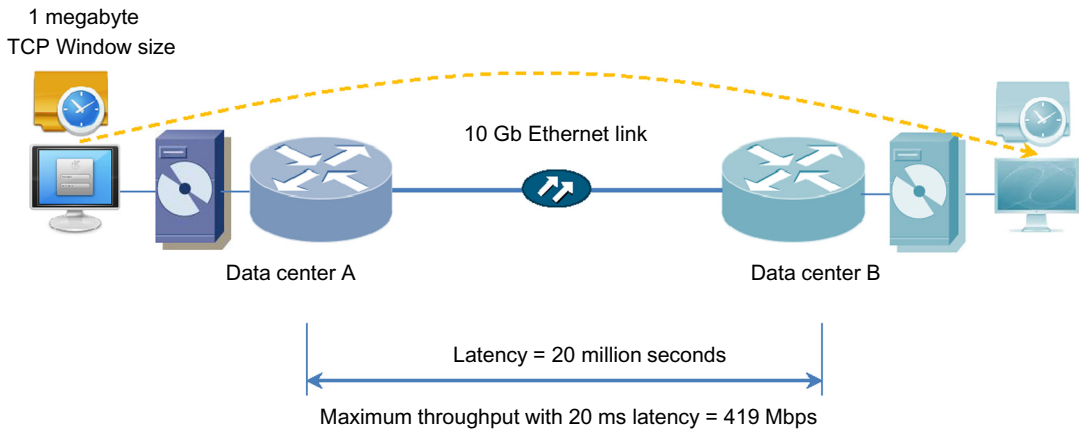

FIGURE 4.18

An example of the relationship between throughput, latency, and capacity between two data center links.

If the TCP window size is fixed, and is 1 megabyte, what is the maximum latency for transferring the file in the GbE link?

- Maximum latency = 8,388,608 bits (file size)/10 Gbits per second (desired throughput) = 838.8 microseconds = 0.838 millionths of a second.

4.2.3 DATA CENTER RESOURCE CELLING

We have touched on the topics of performance and capacity in the above sections. However, data center performance and capacity are very often misunderstood. Although they have a dependent relationship, these two measurements have different purposes.

As we have spelled out above, the purpose of performance is to meet business requirements within the specified quality or to exceed the customer's expectations. Capacity determines what the customers want and when they want it based on the current performance baseline.

From a financial perspective, most capacity planning work will be part of capex activities. In contrast, performance or performance management should be a part of opex activities. The International Standard Organisation (ISO) defines a framework for support and operation activities. We have five subactivities to underpin IT infrastructure operations including performance management (see Figure 4.19).

Similarly, we can establish a framework for data center capacity planning from a cost modeling perspective (see Figure 4.20).

There is a dependent relationship between data center capacity planning and performance management, which is the data center resource capacity (see Figure 4.21).

Why do we need data center capacity planning? It is because the resources are neither unlimited nor free and there is a resource celling in any data center. In order to meet unlimited customer demands with a limited resource pool, performance management is essential.

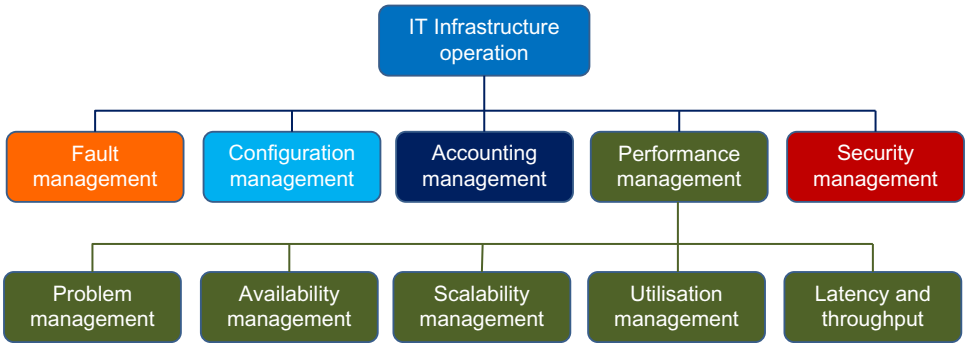

FIGURE 4.19

ISO for high-level IT infrastructure operation framework (FCAPS).

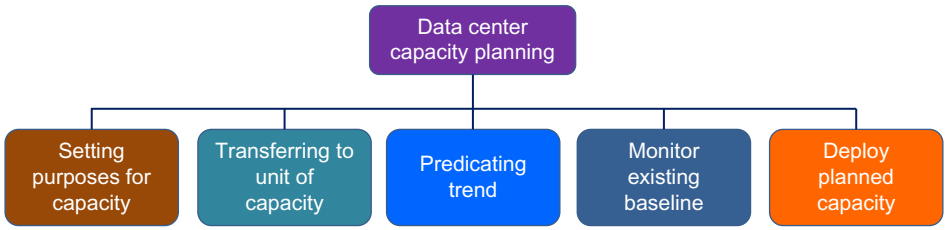

FIGURE 4.20

A framework of data center capacity planning.

Just in the above example, the 10GbE link between two data centers is the limited fixed resource capacity. In order to maximize the throughput, we have to either increase the TCP window size or reduce the RTT. If the RTT is fixed, then the only choice to increase the throughput is to increase the size of the TCP window. The formal description of this relationship is as follows:

$$Throughput\ \ (B) = \frac{TCP\ Window\ Size\ (W)}{RTT\ (D)}$$

For a cloud service provider, the data center resource celling will always be an issue. The right strategy to resolve this issue is to combine both data center capacity planning and performance management or performance tuning. Any cloud service provider should be aware of the risks of long-term workload trending down within a certain life cycle. The famous Amazon on-demand computing diagram (refer to Figure 1.15) only shows one scenario (growth). However, in reality, the IT workload may also trend down.

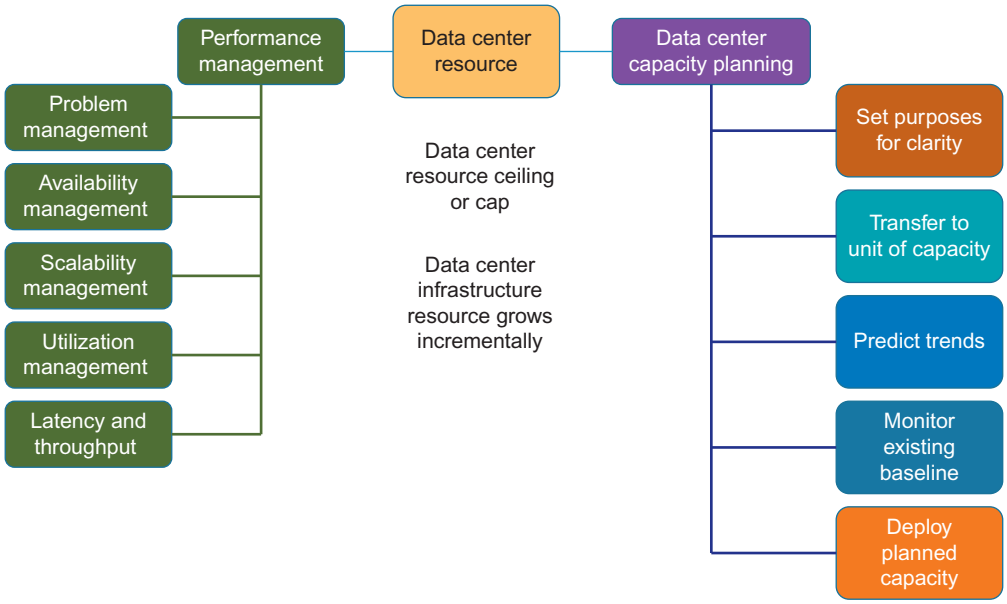

FIGURE 4.21

Data center resources, performance management, and capacity planning.

Therefore, it is absolutely critical for a cloud service provider to have the right marketing strategy and make sure that long-term workload is trending up or at least is kept stable. We will discuss more details in the last chapter concerning investment strategies.

4.3 DATA CENTER SPACE

We mentioned in the section on data center architecture the layout of a data center or its topology, which can be based on the TIA-942 (April 2004) standard (see Appendix H, TIA-942 Telecommunication Infrastructure Standard for Data Center Tier). However, this standard is basically using a telecommunication exchange perspective and is driven by the purposes of telecommunication.

Although it has covered some of the important functions of data center space that are also essential for a cloud data center, such as the entrance room, telecommunication room, office support room, and computer room, it hasn't touched on the topics that are very important from a computing perspective. Therefore, in addition to the data center functions disussed above, we should also pay close attention to the following five types of space:

- Total space (building shell)
- Total adjacent lot size (raw lot size)
- Whitespace (raised floor)
- Effective usable space (rack space)
- General Space

4.3.1 FIVE TYPES OF SPACE

4.3.1.1 Total space (building shell)

The total space is easy to understand; it is the building shell of the data center. It represents the building roof that covers all data center equipment. If you are building a new massive data center, then site selection would be the first step (refer to Section 4.2.1) and then you will build a data center in a new shell. However, many data centers are built in an old building shell, utilizing the existing building infrastructure, such as the Barcelona Supercomputing Center [109] that built a data center inside of Barcelona's Torre Girona chapel in 2005. It is a building from the 1920s (see Figures 4.22–4.24).

The main reasons to build a new data center in an old building shell were:

- Limited time (only four months) for installation
- Close to the research group at the Technical University of Catalonia

The pro is that it is very quick, which you can save a lot of time in laying the infrastructure foundation for the building shell. However, the con of utilizing an existing building is that it is not scalable or there is not enough space for expansion. In addition, the site may only have limited power and bandwidth capacity. The time to renovate an old building is quite difficult to estimate.

Selecting an old building shell may be only applied for small and medium data centers that have limited IT workload growth. It is more like a data center in a box (container) type of solution, where you would like to build a point of presence (POP) close to where your customers are, for example, a multimedia data center application.

FIGURE 4.22

Inside of Barcelona's Torre Girona Chapel [109].

FIGURE 4.23

Inside of Barcelona's Torre Girona Chapel, above the floor (in 2007).

FIGURE 4.24

Inside of Barcelona's Torre Girona Chapel, below the floor (in 2007).

4.3.1.2 Total adjacent lot size (raw lot size)

"Lot" is a real estate term. It refers to a "land lot." The term "total adjacent lot size" means that the data center will have enough potential land nearby for expansion purposes. There is another real estate term that known as "plottage" [110]. It is the process of combining adjacent parcels of land (or lots) to form one large parcel. It refers to merging or consolidating a number of adjacent lots into one large lot (see Figure 4.25).

Figure 4.25 shows an example of total adjacent lot size, which is three times larger than the original land dize of the data center space. It is a part of an investment strategy of building the data center in a modular fashion. However, when you are in the planning stage, it is important to understand the expandable land capacity, the total adjacent lot size. In other words, during the first phase of investment, the cloud workload may be difficult to predict so that you only build what you need, but when your business takes off, you should have enough land capacity to expand your business. Of course, the price of future land and specifically electrically active land is different. Based on the Uptime Institute Inc.'s research data in 2006 and 2007 [113], electrically active land was US$1,120 per square meter per month. If the inflation rate assumption is 3−4%, then today's price would be just over US$1,500 per square meter per month. We will discuss the cost model in later chapters. Empty future space may cost approximately 2/3 that of electrically active land.

4.3.1.3 Whitespace (raised floor)

The term "whitespace" means the usable raised floor data center environment that is measured in square meters or feet (it could be up to few thousand square meters). Many data centers, such as Facebook's data centers, do not use a raised floor, but the term "whitespace" may still be used to refer to the usable square footage [111] (see Figure 4.26).

Regarding the different methsod of air distribution, each air distribution approach has its pros and cons. We will discuss the strategy of cooling air distribution in Chapter 8.

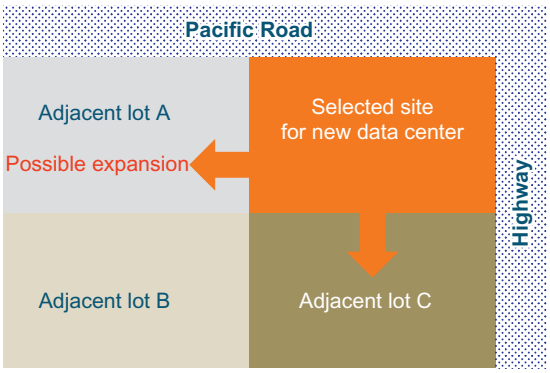

FIGURE 4.25

Total adjacent lot size.

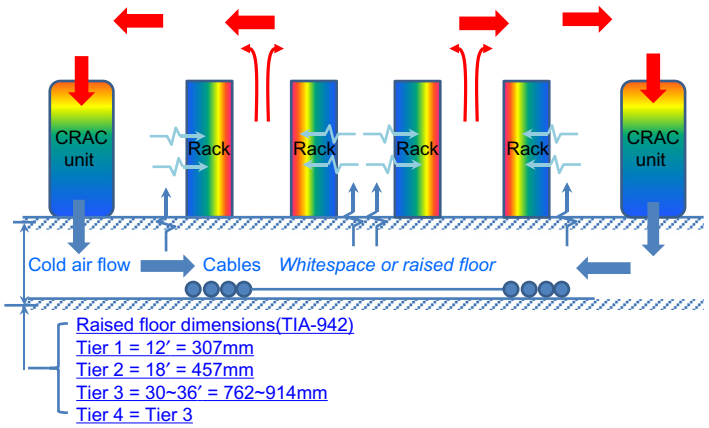

FIGURE 4.26

Whitespace (raised floor).

4.3.1.4 Effective usable space (rack space)

Effective usable space is the rack space for installing IT equipment, such as servers, storage, network equipment (routers, switches, and load balancers), and firewall equipment. A typical size of a rack is 42RU high, which is about 6 feet or 1,866.9 mm. Each rack unit (RU) is 1.75 inches or 44.45 mm high. The width of the standard 42RU (EIA-310D or E) rack will be either 19 inches (482.6 mm) or 23 inches (584.2 mm). The depth of 42RU varies from 600 mm (24 inches) to 1,000 mm (42 inches). Normally, the rack depth size is roughly that of a standard floor tile size, which is 2 feet (600 mm) × 2 feet (600 mm), if there is cold airflow via a raised floor (see Figures 4.27 and 4.28).

Based on research from Kailash Jayaswal [164], only about 50% of data center space is occupied by racks or standalone hardware and the other 50% is for data center facilities, such as power, cooling, aisles, and ramp. Douglas Alger's [109] data showed that useable rack space varies from about 14% to 80% for 14 data centers around the world. However, the average percentage is below 50%, or 43.64%.

4.3.1.5 General space

Furthermore, there are many areas for general purposes only, such as common areas (for example, parking spaces), the network operation center, and general office space. These spaces should also be considered during the phase of building the data center shell. However, the network operation center can be built outside of the data center building shell and controlled by a head office remotely.

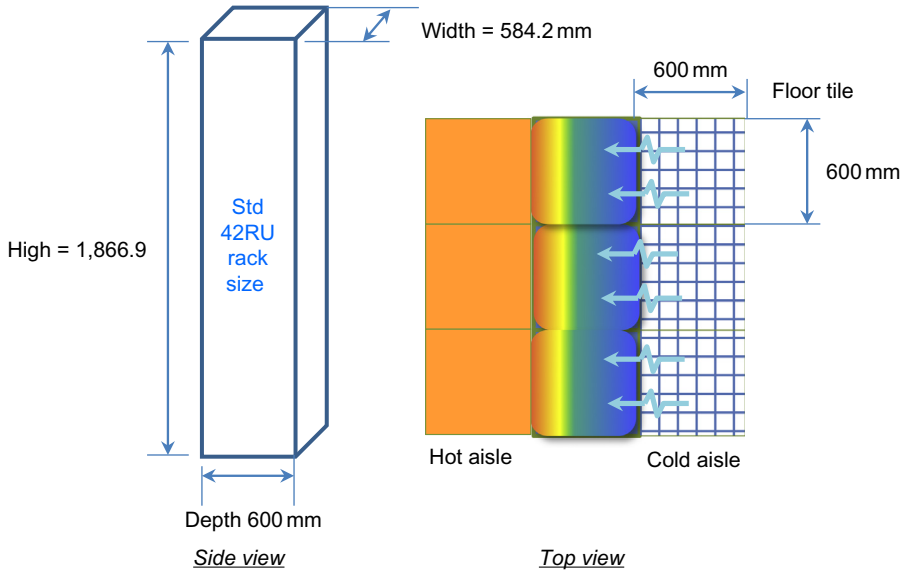

FIGURE 4.27

EIA-310 standard 42RU rack size.

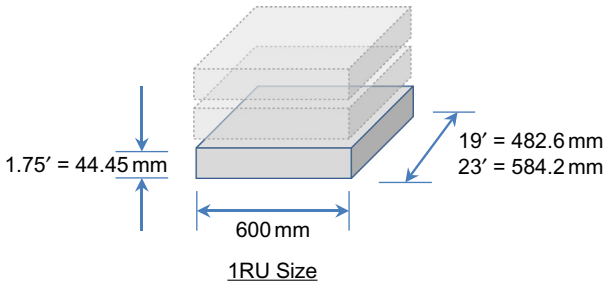

FIGURE 4.28

Standard 1RU size.

4.3.2 DATA CENTER FUNCTIONAL ROOMS

Referring to Figure 4.2 and the TIA-942 standard, a typical data center should have nine functional rooms. If we group them together, we can probably divide them into three types of special function categories:

- Utility support functions
- Computing functions
- Operational functions

Descriptions of these functions are shown in Figure 4.29. Let's explore the details of these functional rooms.

4.3.2.1 Utility support functions

The utility support function consists of three functional rooms: mechanical rooms, electrical rooms, and the staging Area including the storage and loading dock.

4.3.2.1.1 Mechanical rooms

A mechanical room is for installing cooling, compressors, water pumps, condenser units, and ventilation equipment or heating, ventilating, and air conditioning (HVAC) equipment. The room supports all HVAC or cooling functions for the data center facility.

4.3.2.1.2 Electrical rooms

An electrical room accommodates all uninterruptible power supply (UPS) units, batteries, generators, transient voltage surge suppression (TVSS), primary power panels, and service entrance transformers. The room supports a data center's power functions.

4.3.2.1.3 Staging area

This room is designed to store spare parts, faulty equipment to be sent off for repair, and all test equipment, tools, and other work gear (such as occupational and safety gear). This area also includes the loading dock for equipment transportation.

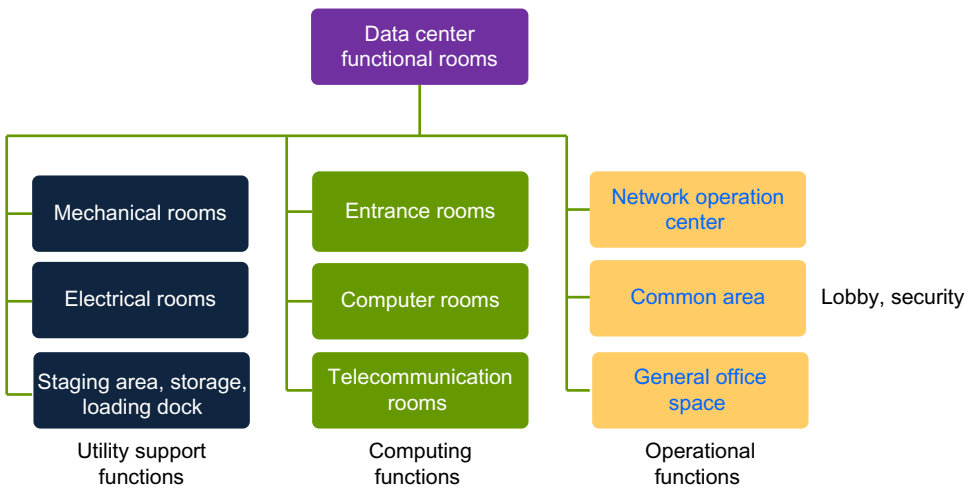

FIGURE 4.29

Data center functional rooms.

4.3.2.2 Computing functions

There are also three functional rooms in this group: entrance rooms, computer rooms, and telecommunication rooms. These functional rooms are mainly to serve the purposes of the servers, storage, and networks.

4.3.2.2.1 Entrance rooms

This is where the outside carriers' equipment meets the internal network. It is the demarcation. Normally, the equipment connects to the backbone network. Sometimes, there are many outside carriers connecting to one data center. In some instances, a carrier broker may optimize different carriers' capacity and become a reseller for a data center's customers, such as colocation customers.

The entrance room is normally located outside of the computer room so that other carriers' technicians can have access without compromising the data center security policy. In order to improve entry point redundancy and reduce excessive cable length, a large data center may have more than one entrance room to provide additional redundancy or to avoid exceeding maximum cable lengths for access provider-provisioned circuits. The entrance room interfaces with the computer room through the main distribution area. The entrance room may be adjacent to or combined with the main distribution area.

4.3.2.2.2 Computer rooms

Computer rooms are the main functional rooms for any data center. Without them, it cannot be called a data center. As we have described in the above section, normally the computer rooms will take up more than 50% of the total data center space. The key difference between a computer room and usable rack space is that the computer room has both hot and cold aisle functions but the rack space may not have both a hot and cold aisle. A rack space is within the computer room. Moreover, the computer room is divided into four different areas:

- Main distribution area (MDA)
- Horizontal distribution area (HDA)
- Equipment distribution area (EDA)
- Zone distribution area (ZDA)

Main distribution area (MDA): The MDA is the central point of distribution for all data center cabling systems. As we can see in Figure 4.2, this area is within the computer room. It may be locked into a dedicated room or area for security reasons. Normally, Private Branch Exchange (PBX), core network switches, and core routers are located in the MDA room. Some provisioning equipment for access providers, such as carrier class high-density M13 multiplexers may also be installed in the MDA room rather than the entrance room. One MDA may support multiple HDAs and multiple EDAs plus operational function rooms.

Horizontal distribution area (HDA): The purpose of the HDA is to support a dedicated area that is within a large computer room or hall for special functions or additional security. It is a centralized distribution point or horizontal cross connection (HC) for cabling to EDAs. A typical HDA may include SAN and LAN switches for the equipment within an EDA.

If a data center has multiple computer rooms that cross different floors, each floor will have its own HC to serve equipment on this floor. For a micro or small data center, an HDA would be unnecessary because one MDA should have enough capacity to handle all IT equipment within the entire computer room. However, for a large or mega size data center, there will be many HDAs.

Equipment distribution area (EDA): The name here explains itself, as this area is for connecting the end equipment or racks consisting of servers, storage disk arrays, access network switches, and other telco equipment. It should not be used for other cabling purposes, such as for an entrance room, MDA, or, HDA.

Zone distribution area (ZDA): This is another subcentralized distribution point between HDAs and EDAs. It provides the flexibility to accommodate frequent reconfigurations, such as "moves," "adds," and "changes." It is particularly useful for DCaaS providers to support colocation services.

4.3.2.2.3 Telecommunication rooms

These are dedicated to equipment related to telecommunication services. They are the common access point for backbone and horizontal pathways. The LAN network switch is also located in these rooms for the purpose of network operations. From Figure 4.2, the TIA-942 typical data center architecture or topology, we can see the telecommunication room is connected to the offices and operation support room, which is the operation center.

4.3.2.3 Operational functions

The next functional group is the operational function rooms. These are the network operation room, common area, and general office space. These three functional rooms support data center operational processes.

4.3.2.3.1 Network operation rooms

The network operation room is also called the network operation center. If it is operating a number of data centers around the world, it may be called a global operation center (see Figure 4.30). It can be located in the data center building shell. It can also be located outside the data center building and communicate via a remote control LAN or WAN. The main purpose of the operation center is FCAPS (see Figure 4.19).

4.3.2.3.2 Common area

The common area would include the lobby, site security, car parking, and cafeteria space. These areas are regularly accessed by data center staff and customers, and are designed for human comfort.

4.3.2.3.3 General office space

General office space is similar to other office areas except it is part of data center security monitoring based on the data center tiering. Again, the purpose of this space is for human comfort so the air conditioning system will be differentiated from the computer or server room.

FIGURE 4.30

An example of a network operation center.

4.4 HOW TO ESTIMATE COST OF SPACE

Many vendors have published the average cost of space. In March 2008, the Uptime Institute [112] published research that indicated the average price for usable space is about US$1,249/square feet/year or US$13,440/square meter/year for electrically active floor area. This can also be translated as US$4,840 per rack per year (with the assumption of the standard 42RU rack dimension 0.6 m × 0.6 m).

However, the author didn't give further details on how to calculate the result and how to make the assumptions for the calculation. Surely, land value will fluctuate from one location to another. Even the construction cost will also fluctuate from time to time, and place to place. It is quite challenging to pinpoint a particular number.

The best approach is to work out how to calculate or estimate the land or space value. Eric Shapiro [113] provided five different approaches to estimating the land value:

- The market approach or comparative method
- The income approach or investment method
- The residual approach or development method
- The profits approach
- The cost approach or contractor's method

Overall, the space will be counted as 15% of the total cost for a typical rack located in a data center that has 2N high availability. It doesn't matter if it is a traditional or cloud data center, they all need rack space or data center space. It is an illusion if people think that cloud computing means the data can be processed in a cloud (up in the air) without any physical infrastructure. From an IT resource usage perspective, there is no difference between cloud computing and traditional computing, such as utility, dedicated, distributed, and grid computing. They all have to consume data center facility resources. The only difference is that cloud computing uses data center resources much more efficiently by leveraging virtualization technologies, such as server virtualization and multitenancy.

4.5 SUMMARY

A data center facility is a physical place to accommodate computing resources that collect, store, share, manage, and distribute large volumes of data. The primary focus of a data center facility is the IT equipment or hardware rather than human resources. In this chapter, we have discussed two main topics:

- Site selection
- Data center space

One topic mainly focuses on the external space and the other predominantly regards the internal rooms or the area under the shell.

When we are planning to build a new data center facility or even extend one of the existing data centers, the first question that we encounter is how to select the data center location. We listed six fundamental issues or factors to be considered and weighted. Among these issues, workload and power are the most important factors because they will impact data center performance.

Because the decision on site selection will have long-term consequences (20 years), all data center ingredients should be carefully balanced out. Ultimately, the decision will impact the data center's performance from a long-term perspective. Data center performance consists of five metrics: site availability, scalability, utilization, latency plus throughput, and reliability. The reliability can also be represented with the common term mean time between failure (MTBF), which includes both mean time to failure (MTTF) and mean time to repair (MTTR). Moreover, MTTR includes both response and resolution time. Any mission critical application requires a specified response and resolution time, which we often call an SLA.

From a data center space perspective, a traditional data center facility has five different types of space: total space, total adjacent lot size (or potential expendable space), whitespace, effective usable space, and general space. As a general rule of thumb, the ratio of usable data center space is around about 50%. In other words, if the rack space needs one square meter then the supporting space of the data center facility will also need one square meter. This is not including the space for disaster recovery (DR). If we consider the DR space, the ratio of usable space for IT workload will be even lower or just above 40%.

If we look from a functional perspective, the data center can be divided into three basic functional group areas: utility functions, computing functions, and operational functions. Each function has three functional rooms. The utility function has mechanical and electrical rooms plus the

Table 4.6 Unit Cost of Data Center Space

Data Center Types	Tier 1	Tier 2	Tier 3	Tier 4
Electrically Active Floor Space	$1,120 per sqm per month or $13,440 per sqm per annum			
Future Empty Space Cost	$700 per sqm per month or $8,400 per sqm per annum			
UPS Redundant Cost per Annum	$12,000 per kW	$13,000 per KW	$24,000 per KW	$25,000 per KW
Redundant Configuration	1 active	1 active	1 active + 1 passive	2 active

staging area. The computing function has entrance, computer, and telecommunication rooms. The operational function has a network operation room, common area, and general office space.

Based on the Uptime Institute's research white paper [115], we can summarize the average cost of usable space would be US $13,440 per square meter per annum or $1,120 per square meter per month in 2007. This would include both capex and opex costs (see Table 4.6).

4.6 REVIEW QUESTIONS

1. What are the different names you know for a data center?
2. How can we define the meaning of data center facility in very simple terms?
3. Why do we need the data center if the data can be retrieved and stored in the cloud?
4. Is the cloud really free?
5. Why do we need data center capacity planning?
6. How can we ask the right questions for data center capacity planning?
7. How can we select a data center site if the plan is to build new data center?
8. How many factors should we consider when selecting the data center site?
9. What does we mean by data center space?
10. Why is the space so important?
11. How many types of spaces are there?
12. How many different functions and function rooms does a common data center have?
13. What is the average ratio of usable data center space in comparison with the total space?
14. What is the average cost per square meter or per square feet per annum for usable space?