
Whitepaper

On-Premise Data Center Analysis and Outlook for 2030

V1.0	2022

Trademarks and Permissions

Notice

The purchased products, services, and features are stipulated by the contract made between Huawei and the customer. All or part of the products, services, and features described in this document may not be within the purchase or usage scope. Unless otherwise specified in the contract, all statements, information, and recommendations in this document are provided "AS IS" without warranties, guarantees, or representations of any kind, either express or implied.

The information in this document is subject to change without notice. Every effort has been made to prepare this document to ensure accuracy of the contents, but all statements, information, and recommendations in this document do not constitute a warranty of any kind, express or implied.

FUTUREWEI® TECHNOLOGIES, INC.

Boston Research Center

Address: 111 Speen Street, Suite 114
Framingham, MA 01701
United States of America

Website: http://www.futurewei.com/

Contents

1 EXECUTIVE SUMMARY

随着数字化转型的深入，当前世界已经迈入数字优先（Digital first）时代，数字化成为用来解决各种问题的优先方案。数据拥有未来的最大价值。而这就是数据中心称为数据中心而不是计算中心。本文讨论 On-prem 数据中心市场到 2025-2030 年的市场趋势。讨论私有数据中心长什么样子，以及构建传统 on-prem，大企业自建 on-prem，和边缘 on-prem 数据中心的模型。本文从 IT 设备供应商角度讨论，而不讨论数据中心基础设施（如电源，制冷，DCIM，digital twins，机架管理，网络出口接入等）。

和公有云相比，On-prem 仍然拥有最大的客户群，以及更多的存量市场。面对竞争，on-prem 数据中心要以进化来面对，利用贴近客户的优势，取长补短，发展出最符合客户需求的产品和解决方案。虽然最近云数据中心投资超过 on-prem 数据中心，但增速已经放缓，未来可能会以 5/5 开来区分市场。

对设备提供商来说，要服务客户的需求，由于大型数据中心和小型数据中心的需求侧重点差的比较远，本文很多地方将它们分开讨论。大型数据中心会有很多与云相似的痛点，而小型数据中心会有各种部署方法，以灵活性取胜。

我们回顾了当前本地数据中心的市场状况，并总结了许多趋势、挑战和预测。 在 2025-2030 年的时间范围内，本地数据中心市场的前景非常光明。 客户调查显示，7 成客户有 pay-as-you-use（payU）市场需求。 在这种模式下，ROI 和 TCO 成为供应商的责任。 供应商可能会采用软件定义基础架构和专用基础架构的组合来提供敏捷性和优化。 专门构建的基础设施可能是不同供应商在用户体验方面的差异化因素。 同时要具有开发性，多厂商和多云支持是用户希望拥有的功能，对某些用户来说至关重要。

拼图的重要一块是所提供的服务。而服务化发放能力随着 On-prem 用户的需求不同，可能扩展到不同服务。保持多样性是数据中心服务的一项差异化优势。除直接使用 payU 很多用户都在构建自己的 IT 服务化发放能力，也对供应商的服务化能力提出要求。我们相信数据服务将成为与基础存储服务一起提供的基础服务。 数据服务包括数据保护和数据处理和分析能力。可以帮助用户实时或接近实时地有效处理数据的服务将以服务化发放。

换句话说，用户希望保持现在的服务种类和等级，然而将管理的复杂性卸载给提供商，同时以用户的身份享受具有弹性和敏捷性的服务。 根据 DC 的规模，供应商希望拥有一系列产品来满足用户需求。以下 10 点概括了数据中心的能力需要模型，提供商可以对照当前的情况，在下面的每一点构建差异化竞争能力：

1. 未来数据中心一个进化方向是硬件 offload 模板化定制化提高效率降低成本，另一个方向是软件定义提高敏捷度，同时对解决方案硬件加速。但两个方向未来会互相融合。比如，一个同时包含软件化 DC 和硬件方案加速的基础设施方案能够提供敏捷度和性能。
2. 数据中心的用户看重软件服务质量和生态，硬件差异化是未来提供更好服务和 ROI 的核心竞争力。比如，量化不同平台的 ROI。在服务化发放的前提下，一个能够增加提供商 ROI 的平台就会提高利润率。而硬件的差异化能保证客户的 SLO。

3. On-prem 数据中心也越来越注重敏捷度，并以服务化的方式提供。比如，用户的诉求是能够挑选所需要的服务，并且这些服务可能是不断变化的。如果客户要求在一小时内变更不同服务间的分布情况，作为供应商如何低成本的满足？
4. 数据中心使用率和机架密度要提高，单位能耗要降低。比如，在满足用户 SLO 的前提下，节能减低采用低功耗设备，在服务化模式下成为一个管理问题。
5. 支持跨界基础设施层面的互联互通，和数据及应用的跨界流动。On-prem 和边缘以及云的数据流动和应用搬迁。比如，能力构建支持和兼容业界的头部平台，以及大量的第三方软件生态支持。
6. 提供业界通用的云原生 DevSecOps 服务以及数据服务。比如，全套的云原生环境服务，包括开源的和第三方软件。数据服务包扩数据库，大数据，以及数据共享等等。
7. 边缘应用和边缘数据中心正在快速增长，产生新的趋势。比如，边缘新应用的接入和部署问题。数据服务在边缘的形态和 ROI。
8. 继续智能与自动化的数据中心运维，降低管理成本。比如，自动运维减低 TCO。
9. 数据中心建设需要系统快速 scale out 和交付能力。比如在 on-prem 服务化场景下，如何快速弹性扩容。打通物流环节，用小时级在数据中心解决容量问题。
10. 数据中心用户需要 Pay-as-you-go 商业模式。比如，扩大融资能力，提供给客户 self-provisioning 和付款的通道等等。

当前 on-prem 数据中心建设模型有几个趋势。在技术趋同和服务化的大背景下，通过硬件提供的差异化能提供独特的特性，更低的成本和能耗，提高方案的特有竞争力。

1. 技术趋同化。随着云原生的流行，以及开源社区的公开性，数据中心建设越来越趋同。好的特性与技术被不同厂家很快的吸收，推出类似的产品。这时候如何内部做好成本管理，外部做差异化特性就很重要。
2. 产品服务化。同样随着云服务理念的流行以及用户认可。On-prem 数据中心的服务化越来越明显。用户喜欢直接比较 On-prem 以及公有云的服务。On-prem 的定制化选择的可能性可以帮助用户得到最合适的服务，也是和公有云竞争的优势。
3. 选择多样化。保持统一服务界面的同时，在技术栈的每一层，客户都有很多的细分选择，包括异构计算。比如 CPU，可以选择 x86, ARM®, Power® systems。加速芯片也有很多家供应商或者自制。在操作系统可以选择各种 Linux®或者 windows®。在虚拟化可以选择各种 VM 以及容器化。在存储可以选择各厂家方案。虽然服务化之后，许多细节是不可见的，但是反映在性能和成本上，服务的质量和费用是不一样的。
4. on-prem 多种形态存在。规模小型或大型化。两头都有优势，而计算需求和与云竞争的规模经济使自己建立和管理中型数据中心慢慢失去成本优势。中型数据中心可以从给客户提供独特价值出发，然后通过提供商的供应链整合产生的低成本，来获得竞争优势。
5. 成本敏感化。很多选择都是基于成本。比如大型化之后，一点百分比都有很大影响，产生定制化需求。提供商可以抓住这个机会，提供多元化定制服务，提供更优惠的价格给客户。
6. 能耗加大化。每机架和数据中心能耗继续扩大。由于高密以及 AI 等应用，每 U 的能耗在不断增加。这时候提供商的能源管理策略就能做出差异化。
7. 强调 sustainability，能源绿色化。数据中心绿色能源继续流行，比如太阳能和风能。提供商能做的还有采用可回收部件，妥善处理回收的旧设备，等等。

8. 混合云随着 on-prem 和公有云的发展继续进展。前面也提到供应商要打通市场上其它的主要平台和第三方应用。

挑战方面，数据中心有很多待解决的挑战，这些问题都是用户希望提供商解决的：

1. 数据中心资源利用率仍然偏低
2. 提高能源利用率，从而使同样规模的数据中心能提供更多的计算资源
3. 继续提高 ROI，降低 TCO
4. 数据中心管理，SLO 和自动负载调度(e.g., AI-based)
5. 数据安全
6. 数据价值挖掘
7. 容灾和数据保护
8. self-driving data centers

在 2025-2030 预测方面，到 2025 年，虚拟化特别是容器虚拟化的份额在加大。Serverless 是否能在 On-prem 流行取决于应用能否跟上。大型的数据中心仍然在尝试资源池化和 composable，提高密度降低能耗。存储会继续向普及化方向发展，而数据行业也是如此，贴近数据，对信息的快速挖掘能力会如同存储一样普及化。网络也继续向普及化发展，但是高速网络的 adoption rate 仍然很低。100Gb 成本还是高，200Gb 和 400Gb 在成本上有挑战。在量子计算，memristor 等技术取得突破性进展之前，这些新技术到 2025 年在 on-prem 数据中心普及的可能性很小，本文不予讨论。但是 3 年内，在边缘数据中心，一些新的商业模式比如 VR 可能会驱动新的边缘数据中心模式，所以要保持开放和灵活接入。

在数据中心级别可能会看到的结果是，

1. on-prem 数据中心会朝以定制化降低成本提高效率以及提高敏捷度两个方向进化。性能加速后的软件平台和定制硬件吸取各自优点，开始融合。低效又不敏捷的数据中心会被市场淘汰。所以供应商要赋能用户，让他们更敏捷和高效。
2. 更超大规模的数据中心由于资源(e.g., 电力)的限制不会更经济。大型数据中心的尺寸会在 2030 年左右达到瓶颈。所以地理分布的多数据中心和规模大小不等的跨界基础设施的支持会很重要。
3. 随着更多企业采取数字第一（digital-first）战略，更多的变化会发生在边缘，小型数据中心可能会看到一些新的形式。而大型数据中心的建设格局和模型会延续这些年新兴的模块化建设风格，因为大型数据中心是多年期的投入（船大难掉头）。意味着现在的头部厂商仍然在大型数据中心的建设方面有先发优势。
4. 由于计算密度和能力的提升，小型数据中心会具有现在中型数据中心的性能，而在占地供电等方面更经济，会分散部署在各地。小型数据中心的潜力很大，如何能够更好的服务用户，把用户留在分层分布的基础设施上是一个进一步探讨的方向。
5. On-prem 数据中心的服务会慢慢标准化和模板化。一个原因是公有云和 on-prem 大厂商在推动用户习惯的改变，另一个原因是客户自己也会尝到模板化带来的好处。模板化和定制化并没有冲突，模板化是用户界面上的，而定制化是为了实现降本增效。
6. 向公有云搬迁和向 on-prem 搬迁的趋势同时存在。更需要弹性的向公有云移动，更需要成本以及控制的客户向 on-prem 方向移动。最关键的还是提供合适客户，并敏捷的解决方案。On-prem 的服务化一定要比现在做的更好。

With the deepening of digital transformation, the current world has entered the era of digital-first, and digitalization has become a priority solution to solve various problems. Data has the greatest value in the future. And that's why a data center is called a data center, not a computing center. This article discusses the market trends of the On-prem data center market from 2025 to 2030. Discuss what a private data center looks like, and models for building traditional on-prem, enterprise-built on-prem, and edge on-prem data centers. This article discusses from an IT equipment vendor perspective, not data center infrastructure (such as power, cooling, DCIM, digital twins, rack management, network egress access, etc.).

Compared with the public cloud, On-prem still has the largest customer base and more market. In the face of competition, on-prem data centers must evolve to face it, take advantage of being close to customers, learn from each other's strengths, and develop products and solutions that best meet customer needs. Although the investment in cloud data centers has recently surpassed that of on-prem data centers, the growth rate has slowed down, and the market may be divided by 50%-50% in the future.

We reviewed the market conditions today for on-prem data centers and summarized many trends, challenges, and predictions. The outlook of on-prem data center market is very bright in the 2025-2030 time frame. Customer survey shows that there are many market requests for on-prem DC as a service. Under this model, the ROI and TCO become the vendor’s responsibility. A vendor may adopt a mix of software-defined infrastructure and purpose-built infrastructure to provide both agility and optimization. The purpose-built infrastructure may be the differentiator among different vendors in terms of user experiences. The solution must also be open. Multi-vendor and multi-cloud support are features that users would like to have, and critical for some users.

Another important piece of the puzzle is the services provided. We believe data services will become a basic service provided along with the basic storage services. The data services include both the data protection and data analytic capabilities. Services that can help users process data efficiently in real-time or near real time would be provided as a service.

In other words, users would like to offload the complexity of management to the provider, while enjoying the services with elasticity and agility as a user. Based on the size of the DC, a vendor would like to have a spectrum of products to satisfy user needs.

In general, the following 10 points summarize the capacity needs of the data center:

- One evolutionary direction of the future data center is to customize hardware with templates to improve efficiency and reduce costs. The other direction is software-defined to improve agility and accelerate with hardware at the same time. But the two directions will merge in the future.
- Datacenter users value software service quality and ecology, and hardware differentiation is the core competitiveness for better service and ROI in the future.
- On-prem data centers are also increasingly focusing on agility and providing it in a service-oriented manner.
- Datacenter utilization and rack density should be increased, and energy consumption should be reduced.
- Support multi-cloud infrastructure. On-prem and edge and cloud data flow and application relocation.

- Provide industry-standard cloud-native DevSecOps services and data services.
- Edge applications and edge data centers are growing rapidly, creating new trends.
- Continue intelligent and automated data center operation and maintenance to reduce management costs.
- Data centers need the ability to scale out and deliver systems quickly.
- Datacenter users need a Pay-as-you-go business model.

There are several trends in the current on-prem data center construction model. In the context of technology convergence and the as-a-service model, differentiation provided by hardware can provide unique features, lower costs and energy consumption, and improve the unique competitiveness of the solution.

- Technology convergence is happening and will continue. With the popularity of cloud-native and the openness of the open-source community, data center construction is becoming more and more convergent. Good features and technologies are quickly absorbed by different manufacturers to launch similar products.
- As-a-service model will be popular. Also with the popularity of cloud service concepts and user acceptance. The as-a-service model of on-prem data centers is becoming more and more obvious. Users like to compare on-prem and public cloud services directly.
- Users have diversified choices. While maintaining a unified service interface, customers have many segmentation options at each layer of the technology stack, including heterogeneous computing. Such as CPU, you can choose x86, ARM®, or Power® systems. There are also many suppliers of self-made accelerator chips. In the operating system, you can choose various Linux® or Windows®. In virtualization, you can choose various VMs as well as containerization. In storage, you can choose various manufacturers' solutions.
- Small or large DCs may thrive. There are advantages on both ends, but for mid-sized data centers, the computing demands and economies of scale are getting harder when competing with the cloud.
- Cost optimization is important. Many options are based on cost. For large-scale DCs, gaining a little percentage has a great impact, thereby resulting in customization requirements.
- Energy consumption is increased. Energy consumption per rack and data center continues to expand. Due to applications such as high density and AI, the energy consumption per U is increasing.
- Green energy is popular. Green energy sources for data centers continue to be popular, such as solar and wind.
- Multi-cloud and hybrid clouds are necessary. The hybrid cloud continues to evolve with on-prem and public cloud development.

Data centers have many challenges:

- The utilization rate of data center resources is still low.
- Energy utilization needs to be improved so that a data center of the same size can provide more computing resources.
- ROI needs to increase and TCO should continue to decrease.
- Intelligent data center manager and SLO/distributed workload management (e.g., AI-based).

- Data Security.
- Find data value via data mining.
- Disaster Recovery and Data Protection.
- Self-driving data centers.

In terms of 2025-2030 forecasts, the share of virtualization, especially container virtualization, is increasing by 2025. Whether serverless can be popular in on-prem DCs depends on whether applications can keep up. Large data centers are still experimenting with resource pooling and composable infrastructure, hoping to increase density and reduce energy consumption. Storage will continue to develop in the direction of being more popular, and the same is true for the data industry. The ability to quickly mine information will be as popular as storage. The network also continues to develop towards popularization, but the adoption rate of the high-speed network is still low. The cost of 100Gb is still high, and 200Gb and 400Gb have bigger challenges in cost. Until breakthroughs in quantum computing, memristor, and other technologies are made, the possibility of these new technologies becoming widespread in on-prem data centers by 2025 is very small and will not be discussed in this article. But within 3 years, in the edge data center, some new business models such as VR may drive the new edge data center model.

What you might see at the data center level is:

- On-prem data centers will evolve in two directions, reducing costs and improving efficiency through customization and improving agility (general software-defined SDDC). The performance-accelerated SDDC and custom hardware DC learn from the other side and begin to merge. Data centers that are inefficient and not agile will be eliminated from the market.
- Larger hyper-scale data centers will not be more economical due to resource (e.g., power) constraints. The size of large data centers will reach a bottleneck around 2030.
- As more enterprises adopt digital-first strategies, more changes will occur at the edge, and small data centers may see some new forms. The construction pattern and model of large data centers will continue the current style because large data centers are multi-year investments.
- Due to the improvement of computing density and capacity, small data centers will have the performance of current medium-sized data centers, but are more economical in terms of land occupation and power supply, and will be deployed in various places.
- On-prem data center service will be gradually standardized and templated. One reason is that the public cloud and on-prem big market players are driving changes in user habits, and the other reason is that customers themselves will also experience the benefits of templating. There is no conflict between templating and customization. Templating is on the user interface, while customization is for efficient implementation.
- The trend of moving to the public cloud and moving to on-prem co-exists simultaneously. Customers who need more flexibility and elasticity move to the public cloud, and customers who need more cost and control move to on-prem.

2 CURRENT MARKET TREND OVERVIEW

In the past 5 years, including the COVID-19 pandemic period, the cloud data centers grow 26% year over year. The on-prem side does not have the increase rate at this level but still increased by about 5%.

The so-called “digital tipping point” also arrived. You can observe it on the market in many aspects. Starting in 2020, the investment in cloud infrastructure had surpassed the investment in on-prem data centers according to Synergy Data ($125 billion vs. $85 billion) [1]. If you add them up, it is close to Gartner’s number [2] for data center systems spending, which is $208 billion. In 2022, the SaaS revenue surpassed the traditional software revenue. The numbers mean that “late adopters” accepted the pay-as-you-go model and it becomes the majority.

This trend may raise some concerns but simply shows that on-prem data centers need to improve to attract more customers. The overall amount of on-prem equipment is still larger than the cloud equipment but mostly due to historical reasons. The on-prem DCs need to evolve to sustain the businesses. The existing large business connections can serve as a base to develop next-generation on-prem DCs to meet customer needs.

The next table shows the main logic behind the two evolution paths: software-defined and purpose-built. Software-defined solutions use over-the-counter servers or IaaS VMs, with over-the-counter accelerators. Purpose-built solutions use custom-made hardware, including a different set of chips than over-the-counter servers and accelerators. Some examples include custom-made CPUs, DPUs, FPGAs, etc.

Survey results	76% of customers want vendors to take more day-to-day admin and operation responsibility [3]
&	73% of customers plan to use pay-as-you-go [3]
⬇	In this model, O&M becomes the vendor’s responsibility (e.g., TCO, energy, etc.)
⬇	Requirements to lower TCO by improving performance/cost ratio, reducing operational cost, and improving elasticity to meet SLAs.
&	Software-defined solution (more flexibility, more elasticity for change, multi-cloud friendly) Purpose-built solution (best fit into cost, energy, and customized need for some sets of workload)
⬇	Mix software-defined and purpose-built solutions based on: • Datacenter size • How often does the workload change (elasticity requirement) • Hybrid-cloud and multi-cloud interoperability
Ideal solution	Mixed solution with scalabilities for both software-defined and purpose-built. Mixed solution with on-prem and cloud Mixed ratio based on customer needs and workload types

2.1 ON-PREM DATA CENTER IS EVOLVING AND VARIOUS DC TYPES

In today’s data center, hardware virtualization can turn generally available hardware into computing, network, and storage. Therefore, it’s hard to find a DC center without virtualization. However, the question is whether to use a fully virtualized data center, also known as the software-defined data center (SDDC) or data centers that utilize the advantages of different kinds of hardware-accelerated solutions, sometimes using specially designed hardware. We call the latter kind purpose-built.

When we put different kinds of data centers in the following diagram, you will see that both small-scale and large-scale data centers favor purpose-built more than the general-purpose configuration. The reason differs. The small-scale ones lean towards a special purpose and fixed environment. As we know, many embedded devices are purpose-built. On the other end, large-scale data centers can afford the investment in purpose-built equipment because even 2% of saving can sum up to a big amount. On-prem hyper scalers, such as Meta®, like the cloud vendors are interested in unique ways to improve their data centers.

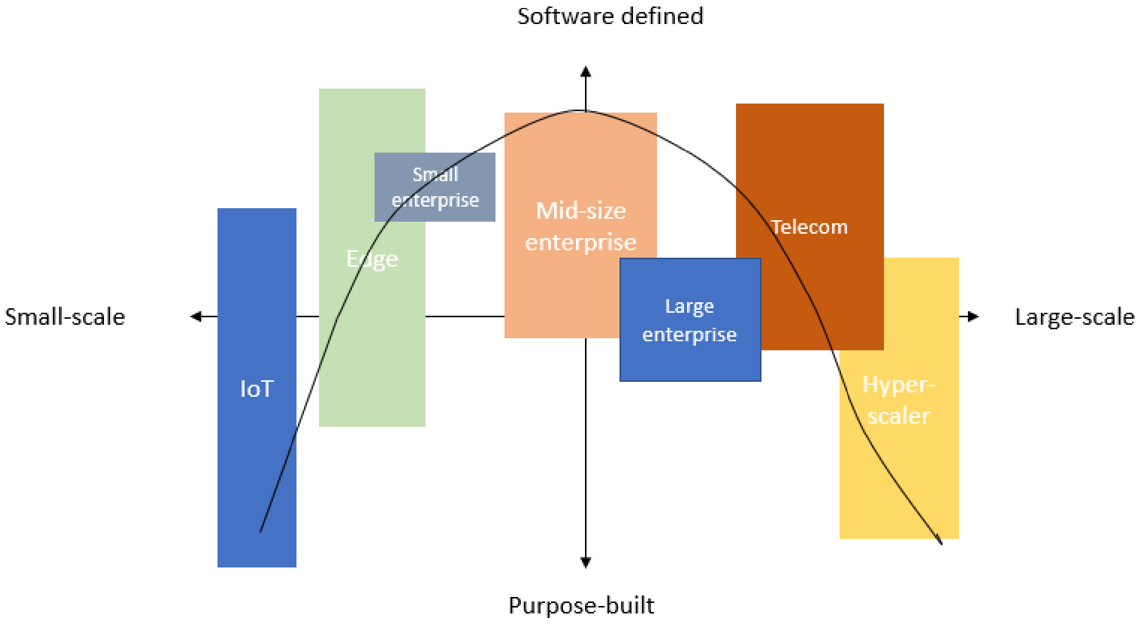
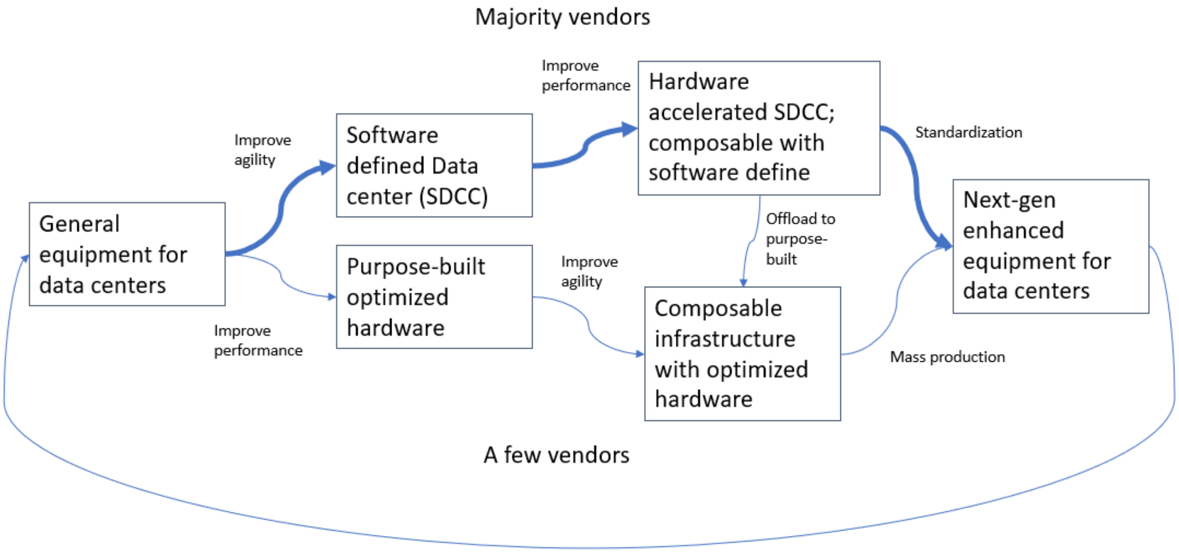


Figure 1 Enterprise preferences under the traditional (non-PAYU) model

A large DC tends to have limited types of hardware, due to maintenance reasons. Too many types of hardware will increase the complexity. For large DCs, even purpose-built equipment will have a large-scale deployment. Mid-size DCs may deploy some purpose-built equipment, based on the features needed by customers.

Note that this diagram is based on the traditional, not pay-as-you-go model. In the traditional model, the users must worry about many details, such as pooling and utilization rate, etc. In the pay-as-you-go model, these issues will be pushed to the vendor side.

We call the initiative to generalize and specialize “two evolution paths”. These two directions are not mutually exclusive. As the hardware providers learn from one another, the purpose-built hardware can emerge in the general-purpose market. The purpose-built system also learns from the SDDC side to offload popular algorithms. Eventually, general-purpose hardware, when become very powerful, will replace some old-generation purpose-built hardware. Then the cycle continues.


Majority vendors
Improve agility
Software defined Data center (SDCC)
Improve performance
Hardware accelerated SDCC; composable with software define
Standardization
General equipment for data centers
Purpose-built optimized hardware
Improve performance
Improve agility
Offload to purpose-built
Composable infrastructure with optimized hardware
Mass production
Next-gen enhanced equipment for data centers
A few vendors

The top path is easier to start and that is why many vendors chose to explore. However, this path depends heavily on the general hardware providers, such as Nvidia®. The bottom path has the advantage of less dependent on other vendors and has the potential to develop the best-fit products, but sometimes lack community support. The purpose-built solutions need to find a balance between cost, performance, and time-to-market. Some vendors choose to pursue both paths and use the top path to provide agility (e.g., running inside public clouds or general servers) and use the bottom path to provide performance-optimized solutions.

2.2 On-prem-as-a-service

As the user requirement changes, such as:

- No CAPEX
- Reduced maintenance needs
- Agility and elasticity
- Reduced IT support

more users are interested in the as-the-service model [4], previously called Pay-As-You-Use (PAYU) mode.

According to IDC [3], 76% of customers want vendors to take more day-to-day admin/operation responsibilities. 73% of customers plan to use pay-as-you-go. As a result, most IT vendors are providing some form of as-a-service model. Examples include HPE® Greenlake and Dell® Apex. Depending on the user's model, the on-prem-as-a-service may incur a lower cost than equivalent public cloud ones.

Public cloud vendors have also joined this market. They provide solutions for the on-prem deployment of a cloud-like system. Such a system is intended to provide the basic need and cloud integration. Many services still reply on the public cloud deployment.

Under this model, the customers offload the admin and operation responsibilities to the service providers. Therefore, the argument of software-defined and purpose-built is invalid in this case.

Customers want to have an industrial-standard user interface, and the rest becomes the vendor's cost. Whoever can provide both agility and a high performance/price ratio eventually will lower the TCO, thereby gaining the advantage in the competition.

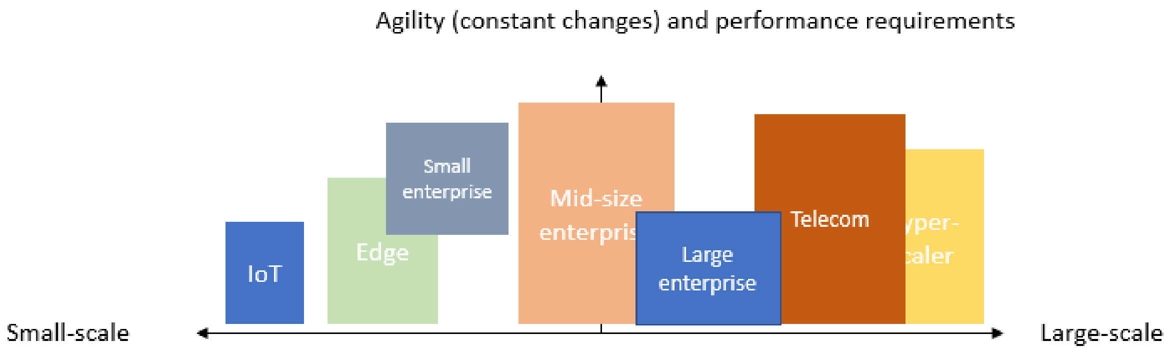


Figure 2 Agility requirement of different sized enterprises in the as-the-service mode

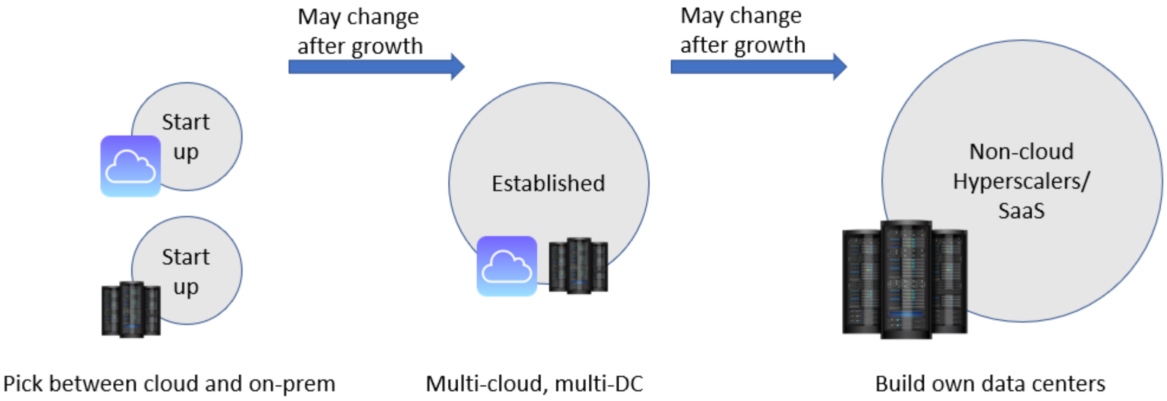
As you can see from the diagram, in the PAYU mode, customers do not care whether the systems are purpose-built. Neither do they care about the efficiency, because they are buying the services. However, they do have the agility (or elasticity) and performance requirements. That will translate into the cost of ownership of the providers, in many cases MSPs or IT vendors themselves. The two paths to improving performance and agility are discussed in the previous subsection.

Some DC vendors, such as Equinix, are trying to shorten the time from purchase to delivery. With the PAYU model, a customer can get the DC resources right away, instead of waiting for 2 months. This quick turnaround time significantly increased the attractiveness of the on-prem offering, which was often criticized for long delivery periods.

2.3 Hybrid IT, hybrid cloud, and multi-cloud

According to this cloud adoption stat report [5], over 90% of enterprises are using public clouds in some form. But many are not exclusive. Over 69% opt for a hybrid solution.

To some extent, small enterprises need the public cloud more than large enterprises, due to their lack of IT support. However, many of them still need data centers and hybrid IT (or hybrid cloud) is adopted by many SMBs, as stated in this Lenovo SMB solution promotion article [6]. Security reasons are often mentioned when on-prem solutions are favored instead of public clouds. However, according to [5], 94% of SMBs appreciate the security upgrade that adopting the cloud brings.



Along the life cycle and size of a company, the data center styles are different. When a company just started, it may choose to use cloud or on-prem, depending on its previous experience. When the business grows to a certain size, many will opt for a hybrid cloud solution. When the business grows too large, some even comparable to public clouds, the enterprise will build its data centers and provide services that fit its needs. Regarding cloud repatriation, A16z [7] raised a paradox "*You're crazy if you don't start in the cloud; you're crazy if you stay on it".* The repatriated systems will require cloud-like services. This is why large data centers are more and more cloud-like from service perspectives.

A mixed multi-cloud IT environment including on-prem DCs and several public clouds will be more common in the following years. Multi-cloud is getting its popularity because an established enterprise will consider vendor lock-in and business continuity issues. Enterprises that can afford to build completely new global infrastructures are not common. Only a few top enterprises have the resources and business needs to do it and eventually execute the plan.

Without any double, adopting hybrid IT and hybrid cloud is a trend today that impacts the next generation of data centers.

2.4 DC IS MORE THAN INFRASTRUCTURE: DATA PLATFORM AND DATA ASSET VALUE

Data is the most valuable asset for the future. Therefore in most scenarios data centers are called "data" centers, not computing centers.

As we know, the total number of bytes generated is increasing rapidly. The trend is reflected in enterprises saving a large amount of data. For example, data lakes are built to keep those data for data analytics. AI/ML tools are used to provide business decisions.

There are many existing data platforms, such as data lakehouse [8] [9]. Many vendors are catching the wave and providing solutions. For example, Databricks® has Delta lake [10] [11] and AWS analytics [12] has its version of the lakehouse. Many more forms of data processing platforms are available and will be developed. So far, the cloud vendors are leading the game but soon on-prem DCs will catch up and provide similar but different solutions.

Deloitte's white paper [13] about data valuation represents the trend of looking at data from a new perspective. When enterprises evaluate data centers, the value of hardware and infrastructure is a portion of the equation, the data stored in the DC could be more valuable than the infrastructure itself. After all, the infrastructure can be rebuilt if lost, but the data may be lost forever. As a result, the industry is finding ways to convert the data into dollar amounts accurately.

2.5 SIMILAR STACK YET DIFFERENT

Because all vendors have a similar technology stack, customers gradually shifted from choosing the stack to better software quality and service level.

In the past 15 years, the industry has been vastly changed by the open-source movement. From operating systems to user applications (e.g., artificial intelligence algorithms and data processing tools), many of the technologies have been the shared asset of human beings and openly provided to all users. In other words, companies have no incentives to build a technology stack from the ground up. Developers are used to the current paradigm. Re-inventing the technology stack out of the open-source base is a huge effort.

However, underneath the cover, the implementations can be drastically different. On each layer, the vendors can have a set of choices, from x86 to ARM® architecture, or different accelerators such as TPU, GPU, etc. Service providers may choose a set of vendors based on their cost, performance, and roadmap. Software is the key component to glue these components into a solution, which matches the customer’s expectations as a service.

2.6 COST SENSITIVENESS

A Gartner report [14] shows that during the COVID pandemic years IT spending is down but will rebound and grow until 2024. The general managers were searching for ways to reduce costs.

In the digital-first era, every industry is undergoing digital transformation. All these strategies need funding to push forward. However, before the new revenue is generated, all these investments are part of the cost. On the other hand, when all enterprises are investing in the digital transformation, the investment becomes a must-have and a way to maintain the current revenue, instead of getting new markets. Ideally, when all enterprises are looking for a solution, the price of such a solution decreases, so that it is affordable to enterprises. Cloud vendors are looking into this trend and attract a large crowd of customers with low entry-point funding.

Another burden on enterprises is the increasing amount of data. As we discussed in the trend “DC is more than data infrastructure”, enterprises invest a large number of dollars for data processing and storing. They are asking for more cost-effective ways to process the data. As an important piece of the puzzle, if not the most important piece, on-prem solutions are used to process the most sensitive and most valuable data.

2.7 ENERGY EFFICIENCY

Today, all large-scale DCs are evaluated by the Power Usage Effectiveness (PUE), which is the ratio of Total power usage and the IT equipment power usage. However, PUE mostly only measures the efficiency outside of the IT equipment (e.g., cooling) without taking the efficiency of IT equipment into account. There are some new metrics to evaluate IT efficiency, but the equipment utilization rate is still the most commonly used one.

Pawlish et al. [15] analyzed the DC energy usage in a university setting and they found that an effective way to increase energy efficiency is to increase the utilization rate by pushing some workloads to the cloud. Hybrid IT and hybrid cloud essentially push the problem to the cloud side. According to AWS® [16], AWS® is five times more energy-efficient than average EU enterprises.

On-prem solutions are improving energy efficiency and utilization by virtualization, hyper-convergence, etc. For a large DC, the on-prem solution can utilize the same techniques as public cloud vendors. For smaller businesses that do not plan to build reserved capacity, a hybrid cloud is a reasonable way to boost efficiency on the on-prem side.

Thanks to the energy efficiency sensitive data centers, the annual actual electricity usage of all US data centers is flat from 2007 to 2014 [17], at roughly 70 billion kWh. In 2017, the number seems to be increased to 90 kWh [18] but the workload may be increased, and the measurement scope may be changed. With the new AI applications and autonomous driving coming, whether energy usage can be kept at a slowly increasing rate is a question.

2.8 Datacenter sustainability

Sustainability is defined as the capability of performing functions without affecting future generations. In many cases, sustainability is measured in the environment, economic, and equity impacts [19]. The sustainability issues are ranked much higher than 10 years ago due to climate change concerns. Governments and politicians are investigating better laws and mandates for DC sustainability. Millennials (born between 1981 and 1996) and Gen Zs (born between 1997 and 2012) grew up in the sustainability discussion. Many millennials are already making purchase decisions, and Gen Zs are the future customers. As a result, 92% of S&P 500 companies issue sustainability reports.

Sustainability includes green energy but also includes other aspects, such as fair treatment of employees, diversity, and inclusion, and other social issues. Initially, enterprises treated sustainability as a business cost, but recent data show that enterprises focused on sustainability lowered the total cost. For example, clean energy may have a lower cost than traditional fossil-based energy providers. In a larger context, sustainable enterprises have other benefits. For example, a sustainable enterprise having a good reputation can recruit talents easier than others.

3 Challenges

There are many challenges in future data centers. On the other hand, users are not satisfied with their current infrastructure. According to Forbes [20], only 29% of leaders and engineers say their infrastructure meets the current needs.

Let us describe the challenges in the following subsections.

3.1 Resource utilization rate

The utilization rate is an important metric to measure how efficiently the systems are used. A low utilization rate generally means that some part of that investment is unnecessary, even though it can be temporary.

The PAYU mode or the as-a-service model cannot avoid the resource utilization problem. Imagine a service provider that receives a request to double the resource but releases the doubled resource after one hour, the resource utilization immediately dropped to below 50% and half of the resources are not generating revenue. The fundamental problem here is that public clouds can attract enough customers so that their usage peaks and valleys can help one another, whereas on-prem vendors often do not have the luxury to build a large enough customer pool.

For on-prem data centers, to achieve a high utilization rate, the data center can either (1) reserve a small capacity and keep the resources fully utilized, and use the public cloud as a pool of resources, or (2) build an on-prem pool of resources by mixing different kinds of workloads. This pool will have enough elasticity to prioritize important workloads and make some workloads idle on demand.

If multiple data centers are pooled together, a distributed workload scheduling may be helpful to transfer workloads among these DCs. So the overall data centers can have a high resource utilization rate. In this case, if enough DCs are pooled together, the resources can be built as an economical model, such as banking loans. However, the workloads are often not easy to move. Or under some circumstances, not allowed to move due to security and regulation reasons. That will increase the cost of the elasticity.

3.2 Energy efficiency rate

Energy utilization rate is related to the utilization rate because when the utilization rate is high the overall energy efficiency rate is higher. However, these two concepts are different. By energy efficiency, we mean the energy consumed to finish a fixed amount of workload.

In most cases, a hardware-accelerated solution is more energy-efficient than a solution built on general servers with CPUs. Many vendors use ASICs, FPGAs, GPUs, etc. to accelerate and improve energy efficiency. In theory, if a workload requires only one type of computation then offloading it to energy-efficient hardware provides the best efficiency. A purpose-built system can optimize energy consumption.

The challenge is to find a balance between performance, cost, and time triangles. Developing the acceleration solution takes a long time and increases the cost in the beginning.

3.3 System complexity and self-driving data centers

The systems are getting more and more complicated, but at the same time, the goal is to reduce the staff numbers. First, the complicated systems are not manageable by humans. A mid-size data center has thousands of servers and by virtualization, the total number of VMs is skyrocketing. Second, the progress in automation and AIOps has made self-driving data centers possible.

The effort of building self-driving data centers has been an effort for many years. The result is a mix. On the provisioning side, automation with "infrastructure as code" has effectively released the burden from data center owners. On the health monitoring side, progress has been made to better observe and understand the data center systems. However, the goal of an overall self-driving data center is not reached. Many of the efforts are individually made by different vendors. An overall framework or standard is needed to manage the full infrastructure. Many of the old frameworks are not developed for modern-day data centers. The new front of self-driving data centers is controlled by several cloud companies, and not shared with the public. In a way, the cloud vendors perfectly avoided the integration problems, because they built the infrastructure from the bottom up. The need to integrate with AIOps tools of different vendors is not there.

For on-prem data centers, the challenge of system complexity and self-driving data centers is going to be more severe. For customers, the dilemma is that some level of self-driving is more realistic if a single vendor is used, but there is a risk of running into "vendor lock-in". Again, the as-a-service and PAYU model will avoid this dilemma and customers only use the services. The complexity and self-driving data center problem are pushed to the vendors.

3.4 LOWERING THE TCO

The TCO includes many aspects, including hardware, software, management costs, and more. In older days, the concept is vague and hard to predict. However, with the cloud vendors joining the game, the service-based market has been putting a clear price tag on the services, from IaaS to SaaS. These numbers are often referenced as a basic starting point to compare with on-prem data center solutions.

Today, lowering the TCO via hardware cost is a challenge, partly because the need for computation and storage resources are growing exponentially but the CPU and memory development slowed down. The CPU used to be 10 times faster every 4 years. But between 2006 and 2020, the CPUs are 10 times faster in 16 years [21]. The memory price depends heavily on other economic factors and you can see the price go up and down in the price trend chart [22]. It is far from the ideal chart in which the memory price goes down continuously. The result of these factors is that even though systems are faster each year, enterprises are still spending more than before on IT budgets.

The cloud vendors are using ODM products to lower the costs. The quality of each server or component is not emphasized as much as in the on-prem solutions. The ideas are simple. If in a large deployment environment individual component or server failures are expected to happen, why would anyone spend too much for the guarantee of a single component? A large on-prem DC solution shares the same vision. However, in smaller deployment settings, a better-quality component not only improves the availability but also saves the maintenance cost in terms of time and effort.

On the software side, automation can be used to lower the TCO. Self-driving data centers not only solve the complexity issue but are also used for a lower TCO, which we discussed in detail in the previous section.

The TCO challenge is closely related to the discussion on different types of data centers. As we discussed previously, different DCs require different ways to minimize cost. Some prefer to use software-defined data centers, and some prefer to use purpose-built solutions to create the biggest value for a dollar.

3.5 DATA RESILIENCE

Data resilience means that It includes both the data security and data protection/DR aspects.

Data security is a worldwide issue for data centers. The security covers a wider range of attacks, for example, information theft, overflow exploits, and ransomware attacks. The data security issue has been particularly sensitive as several high-profile ransomware attacks are in the news. When activated, ransomware encrypts the data on disks and asks for a ransom to have the data decrypted. Thus, the attacks are profitable worldwide. The ransoms are paid by cryptocurrency so the tracing is deemed impossible.

Data protection and DR are old subjects. But while the computing environment evolves, the data protection and DR needs to be evolved to support new virtualization environments, such as containers. The cost of data protection and DR also needs to go down because the cost becomes a burden when the amount of data goes up quickly.

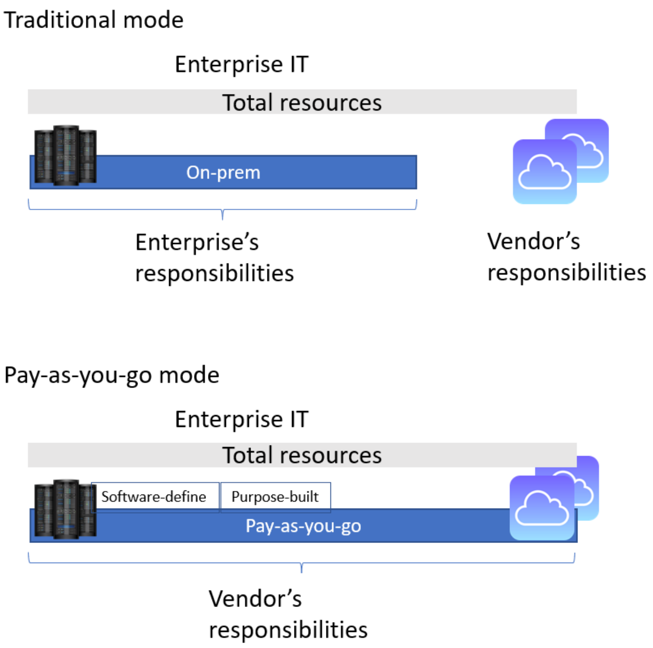
How to reach data resilience is an art that requires the effort of both the vendor and the data owner to work together. In many attack cases, it is the data owner who ignored best practices. It will interesting to see whether data resilience as a service could shift the burden to the vendor side.

4 PREDICTIONS FOR 2025--2030

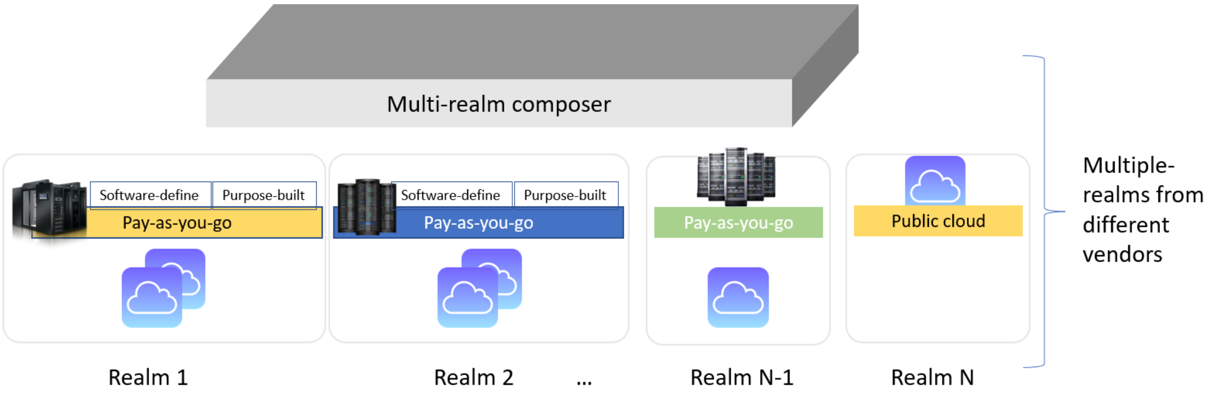
4.1 ON-PREM DATA CENTER CONSTRUCTION MODEL

We predict that the data centers between 2025—2030 adopt the following guidelines:

- More on-prem DCs adopt the Pay-as-you-go model or rental model (DC-as-a-service)
- Day-to-day DC administration and operation managed by vendors, including upgrading and tech refreshing.
- Low-cost elasticity satisfied by hardware-level composition and virtualization-based workload management
- Mix software-defined components for elastic workload and purpose-built hardware for best performance. Some software-defined components and a purpose-built component may converge.
- Fast scale-out and delivery for on-prem DCs, ideally comparable to the public clouds
- Elasticity to clouds as a requirement may be imposed on vendors
- Energy efficiency requirements imposed on vendors for sustainability
- Self-driving DC management
- Data resilience is a requirement imposed on vendors



In the next years, more responsibilities will be shifted to the on-prem vendor. The enterprise IT may not care about the percentage of on-prem and the percentage of public clouds. The public cloud vendors are also trying to capture this market. In the next years, the competition may decide the actual market percentage between on-prem IT vendors and public cloud vendors.



Pay-as-you-go mode with multiple vendors

Furthermore, an issue of combining multiple vendors into an integrated DC environment will arise. Unlike a public cloud, where all zones are under one company's control, and on-prem data center may consist of multiple realms, which are provided by multiple vendors. Each realm may have its ways of optimization and administration. The enterprise IT department has to use a "multi-real composer" to combine these realms as a composable infrastructure. The process is more complicated than the multi-cloud problem because there are much more vendors in the on-prem vendor list. Hopefully, the industry can propose a standard when the market is mature.

4.2 FROM APPLICATION-CENTRIC TO APP-DATA-CENTRIC

In the past 15 years, the number of applications increased rapidly. The amount of data is also increased exponentially. Due to the DevOps, container, and microservices, modern applications are more agile than previous generations of applications. The data processing mechanism also went through revolutions by different big data processing engines.

In enterprises, the centralized IT organizations gradually give more responsibilities to the line of business (LOBs), and the business architecture is shifting the IT architecture. These two revolutions will be combined in the next 10 years. From application-centric IT is shifting to app-data-centric because the data of each LOB need to be processed efficiently. Each LOB has the challenge of processing a massive amount of data.

A data center must provide the infrastructure needed by these LOBs. For small DCs, LOBs can deploy purpose-built equipment for the features that they need. For mid-size or large DCs, LOBs can only deploy software-defined solutions because larger DCs generally limit the hardware to a few models. In recent years, with virtualization software-defined systems are getting more and more powerful.

Any vendors who can capture the data needs of the enterprises will have a better chance of being accepted.

4.3 VIRTUALIZATION PROGRESS

The VM virtualization and the cloud-native platform are already a fact in on-prem data centers. In the next five years, more users will adopt them.

However, the fate of serverless computation in on-prem data centers is unclear. The serverless platform is a success in the public cloud already. But it takes years to be expanded to on-prem infrastructure. There are several reasons why serverless is not yet popular in on-prem settings:

1. The ecosystem for serverless-based applications is not mature.
2. Many serverless-based applications are not optimized in performance.
3. On-premise vendors are not eager to provide serverless PAYU modes. The profit is not proved to be high with the limited amount of usage.
4. For traditional on-prem models, serverless is only cost-related. The current business model does not trigger the necessity to adopt serverless.

The revolution in data processing might change some of that. The current infrastructure of serverless computing may not be needed on-prem, because the same virtualization idea can be implemented in different layers. For example, A SaaS platform accepting function-as-a-service requests may dominate some scenarios, and be more favorable than a general-purpose serverless infrastructure.

4.4 EDGE REVOLUTION

Edge data centers are driven by IoT and new edge applications. With 10 billion devices connected in 2021 [23], the amount of data needed to be processed in edge data centers. The revolution is ongoing. The direction and period of this revolution are still unclear. But in the industry, most analysts believe the IoT revolution will start when enough devices are connected. AI and cloud are two trends that will be combined with IoT on edge.

From a data center perspective, edge data centers are deployed near the edge. The granularities can be different. Some edge data centers will be deployed very close to devices. But some edge data centers are relatively centralized. Therefore, edge data centers have different sizes.

The new trends of PAYU, agility, elasticity and efficient data processing are making progress in the edge computing world. The coming new applications very likely will adopt the latest application trends and require infrastructure support. The balancing of software-defined and purpose-built is also applicable here.

4.5 Real-time/near real-time processing

Numbers show that real-time and near real-time workloads are increasing. The speed of finding information in data is a key metric in today's business world. A large percentage of computing infrastructure in future data centers is optimized for real-time processing.

The infrastructure users have different requirements in terms of cost, performance, and scalability. Some use cases need to have ultra-short latency – for example, financial data processing. Many users would like to pay for the short latency. Some use cases need a large-scale configuration but are less stringent on latency. The more stringent the real-time, the more expensive the solution is.

Academia and industry are very interested in this area. There is a niche market for purpose-built hardware for extra processing capability. However, the majority of customers are not bound to an extreme latency and hence existing solutions such as open-source-based solutions are sufficient.

4.6 Multi-clouds

Multi-cloud is already an overused term in the industry. There are at least two types of multi-cloud integrations:

1. Use other clouds as an extension (for backup, DR, migration, etc.)
2. Use other clouds as peers

The second type will have a tighter coupling than the first one. An enterprise may establish a multi-cloud solution among different realms, which may not be geo-physically separated. Due to the solution from multiple vendors, even in one data center, the multi-cloud capability becomes useful.

In future data center solutions, interoperability to other clouds (or other vendors) is a critical feature, unless the data center is designed to be standalone, or only communicate within the same vendor.

5 Possible changes

Due to the development of hyper-convergence and SDS, the three traditional pieces of a data center -- compute, storage, and network – cannot fully describe a modern data center anymore. We describe the new computing paradigms which cover both computing and storage, then add a section for storage alone.

5.1 NEW DATA CENTER PARADIGMS

5.1.1 Mixed pools and the composable infrastructure

The composable infrastructure is to flexibly compose a “machine” that is equivalent to today’s bare metal resources from different pools of CPUs, DPUs, GPUs, memory, disks, etc. With the new infrastructure of high-speed links, the composable concept becomes more feasible.

Gartner® defines composable infrastructure as a “highly flexible software architecture that leverages standard hardware building blocks that can be easily scaled-up, scaled-out, scaled-down, or scaled-in as business needs dictate”. The composable infrastructure does not limit the architecture to support the building blocks or pooling. Another term “Composable disaggregated infrastructure” (CDI) is used to describe the nature that the infrastructure is often disaggregated, which contains different types of resource pools.

Composable infrastructure provides the capability to “compose” an environment that fits user needs within a boundary. The benefit is that an application does not need to care about the distributed nature of the underlying environment. The “machines” have virtualized traditional components of CPU/Network/GPU/Storage. Today’s applications are already taking the distributed system into account with partitioning, etc. Whether new types of applications will benefit from a large resource pool, such as near “unlimited memory”, is a question.

Composable infrastructure has two evolution paths as well. One is purpose-built with unique hardware and is more “pure-bred”. HPE® Synergy and Fungible® are two examples. Another path is a “soft” composable vision, which uses today’s hardware as bases and composes systems from “virtual” pools.

In the 2025-2030 time frame, we predict that the software-defined path may be more popular in the on-prem world when there are already many hardware choices on the market. The reason is that on-prem customers are still interested in integration and migration issues and reuse of existing infrastructure, even in the pay-as-you-go model. The purpose-built composable infrastructure takes a long time to evolve and be accepted by customers. On the other hand, private/public clouds use the black-box model and can pursue the more “pure-bred” version without affecting the customer user experience, and these companies have a higher chance of experimenting with new computing architectures. On-prem large data centers share many characteristics with clouds and therefore may deploy more purpose-built composable infrastructure.

5.1.2 Bundle, various forms of convergence, and xPUs

Compared to the composable infrastructure idea, which only defines the goal without limiting the implementation, the various forms of convergence architecture clearly define the implementation. Bundled systems are welcome in the on-prem configuration because they provide similar benefits as cloud providers. They define clear boundaries so that deployment is easy. Very little hardware integration with existing systems is needed. These converged systems tend to have their independent management interfaces as well.

The xPU movement impacts the co-processing market deeply. Using co-processors or accelerators can provide better ROI for customers. The value is easy to justify when the workload is unified and stable. The flexibility could become an issue when more types of workloads are sharing the same infrastructure.

Some kind of software-defined co-processors might change the market when they are available. Agility is another metric to measure how service providers can increase their ROI. DPUs can fully function as servers and can be configured for both management and data processing purposes.

In the PAYU model, customers choose to ignore the difference between CI, HCI, and dHCI. What vendors used to attract customers are now part of the internal O&M burden of the vendor. Similarly, the benefits of software-defined or purpose-built are internal to service providers. Therefore, the freedom could cultivate a group of new exciting solutions having new tradeoff standards under new criteria.

5.1.3 Silicon customization and the merge of processing units

From general-purpose computation to purpose-built, many vendors in the industry are trying to find a unified way. For example, chiplet [24] is a proposed way to embed pre-designed dies into a foundation. With multiple chiplets combined, a vendor can quickly design a system on a chip (SoC) [25] for the end customers.

This is an example of the endless cycle -- software-defined solutions and purpose-built solutions will merge and generate new general-purpose solutions. Silicon customization used to be in the purpose-built area. Soon the flexibility issue limits the speed of integration. Hence, the industry is using the concept of chiplet to overcome flexibility issues. Eventually, with the joining of vendors like Intel® and AMD®. CPUs will have chiplets embedded and thus the different processing units are merged into one.

We predict the software-defined functionalities may gradually be moved to SoC. For large DCs, the wave is predicted to happen and a software-defined method.

5.1.4 Serverless and function-as-a-service

Serverless is another paradigm shift that fits the pay-as-you-go. Unlike composable infrastructure, serverless can be implemented on the current infrastructure. However, as we pointed out in previous sections, how to optimize the cost and solve the elasticity problems become the vendors' problem. Serverless is also an extreme form of software-defined infrastructure.

Note that function-as-a-service is not going to be mainstream until the applications become serverless. The trend may take years, but function-as-a-service is very friendly to data analytics and AI workloads.

Some of the serverless applications are possibly be adopted first:

- Big data processing applications
- ML applications
- ETL pipelines
- Batch jobs

Note that in an on-prem setting function-as-a-service is only meaningful when the vendors find a win-win way to obtain enough profits from providing such a service. The risk is that the ROI is so small or even negative when only a few people adopt this service.

5.1.5 High density

The density of computation is a metric to save rack spaces, which are limited by physical spaces. This metric has a higher value for large DCs than smaller DCs because a small percentage of improvement has a big impact.

It is interesting to see whether this metric will become a vendor’s responsibility in the PAYU model. After all, the end-users would like to care less about the infrastructure side, such as rental and electricity bills.

5.1.6 High-speed network

The network speed is quite fast today, but the adoption rate of a fast network is still rare. The price needs to be dropped to attract more users. The topology of the web-based application today limits the response latency. Any speed up in the backend only improves the overall time by a small percentage. Therefore, the users are not investing in speeding up the network. Maybe when the applications have dedicated and low latency networks, such as 5G, the backend network will need an upgrade.

5.2 Traditional Storage Scope

IDC’s market forecast [26] predicts that while growth in the Worldwide External Enterprise Storage Systems market will slow from 7% to 2% in 2022-2025, revenue will reach $36.9 Billion in 2025. The worldwide enterprise storage market (Worldwide Enterprise Storage Systems) grew at a higher CAGR of 7.5%. A total of $177.1 Billion by 2025, which also includes OEM, and ODM markets, such as server built-in storage. In 2022, ODM share, a large portion of which is storage supplied to cloud vendors, will surpass the external enterprise storage market for the first time.

The SAN market is fairly stable. With better hardware and co-processers, the performance is expected to get faster. But customers seem to be reluctant to adopt faster systems, such as NVMeoF. In the next five years, the upgrade could be slow but will certainly happen. In the following years, distributed storage has a higher growth rate than primary storage.

5.2.1 File and object (FOBS)

Distributed storage includes Virtual SAN (HCI appliance), files, and objects. On-prem files/objects have limited growth in 2022-2025, from 8 Billion to 8.4 Billion. It means that a large number of manufacturers still have to strive for share growth in the file and object market of the same size. Among them, private cloud files are much more competitive than public cloud vendors.

Open network Ethernet-based files and objects are projected to grow at a CAGR of 4.5% through 2020-2025. And the growth from 8.0 Billion in 2022 to 8.4 Billion in 2025 is very low. IDC's growth forecast for this segment of the market is very conservative. This also means that the current manufacturers in this market have to compete with each other and strive to increase their share in the same size file and object market. At this time, how to provide more competitive products in a homogeneous market becomes an interesting question.

5.2.2 On-prem or public cloud

By 2025, IDC predicts that private cloud FOBS will have a capacity of 139.2EB, while public cloud will have a 1ZB capacity. About 15% of the public cloud. Although objects are larger than files, they account for 86% of the capacity and 60% of the profit. But a closer look at the status of files and objects in the private cloud is quite different, where files and objects show opposite trends.

- Object public cloud accounts for 80% of capacity, and on-prem accounts for only 6%. Almost 7% of OBS revenue, if the unit price is similar to that of the public cloud, it is less than 2 Billion. However, the market unit price is higher than that of the public cloud, so it is more revenue than 3 Billion.
- File public cloud accounts for only 8.4% of the total FOBS capacity, and on-prem also accounts for 5.6% of the total capacity. Almost 40% of FOBS revenue. Taking the scale-out file system as an example, it is a revenue market space of 4.8 Billion.

That is to say, in the field of files, on-prem and public cloud has a match, but in the field of objects, a large number of users choose the public cloud, and on-prem must provide reasons to attract customers.

5.2.3 Scale-out NAS

Sale-up NAS revenue is at a slightly declining -1.2% CAGR through 2025. The scale-out NAS will grow at a CAGR of 11%, and by 2025, it will become three times the revenue of Sale-up.

5.2.4 Virtual SAN and HCI appliance

Correspondingly, IDC predicts that the VSAN market is growing rapidly. The CAGR is expected to be 11.6% through 2025. From 6.7 to 8.8 Billion in 2025, more than the market for files and objects. There are plenty of opportunities to accommodate new products and vendors.

5.2.5 New Media

The industry is not making much progress in the area of new media. Storage class memory has been on the market but is facing some difficulties in finding its position. The adoption rate is not as high as previously expected. However, if the price of it is reduced in the coming years, it may find its usage in many places.

Archiving media is another area where progress is stagnated. Blue-ray has been on the market for years now. Other media, such as DNA storage, is still in the pre-infant stage. In the next 5 years, there is no real marketable new media on the horizon.

6 SUMMARY

We reviewed the market conditions today for on-prem data centers and summarized many trends, challenges, and predictions. The outlook of on-prem data center market is very bright in the 2025-2030 time frame. Customer survey shows that there are many market requests for on-prem DC as a service. Under this model, the ROI and TCO become the vendor’s responsibility. A vendor may adopt a mix of software-defined infrastructure and purpose-built infrastructure to provide both agility and optimization. The purpose-built infrastructure may be the differentiator among different vendors in terms of user experiences. Multi-cloud support is a feature that users would like to have, and critical for some users.

Another important piece of the puzzle is the services provided. We believe data services will become a basic service provided along with the basic storage services. The data services include both the data protection and data analytic capabilities. Services that can help users process data efficiently in real-time or near real time would be provided as a service.

In other words, users would like to offload the complexity of management to the provider, while enjoying the services with elasticity and agility as a user. Based on the size of the DC, a vendor would like to have a spectrum of products to satisfy user needs.

7 REFERENCES

[1] R. Miller, "Cloud infrastructure spending passed on-prem data centers in 2020," 19 Mar 2021. [Online]. Available: https://techcrunch.com/2021/03/19/cloud-infrastructure-spending-passed-on-prem-data-centers-in-2020/.

[2] Gartner, "Gartner Says Worldwide IT Spending to Grow 4% in 2021," [Online]. Available: https://www.gartner.com/en/newsroom/press-releases/2020-10-20-gartner-says-worldwide-it-spending-to-grow-4-percent-in-2021.

[3] IDC, "IDC connects talk 2022," [Online].

[4] M. Korolov, "On-prem-as-a-Service Comes Into Its Own During Pandemic," Data Center Knowledge, 22 Nov 2021. [Online]. Available: https://www.datacenterknowledge.com/cloud/prem-service-comes-its-own-during-pandemic.

[5] N. Galov, "Cloud Adoption Statistics for 2021," 9 Aug 2021. [Online]. Available: https://hostingtribunal.com/blog/cloud-adoption-statistics/.

[6] Lenovo, "Bridging the Small Business IT Gap between On-Prem and Cloud, with Lenovo," 26 Jan 2022. [Online]. Available: https://techhq.com/2022/01/bridging-the-small-business-it-gap-between-on-prem-and-cloud-with-lenovo/.

[7] S. Wang and M. Casado, "The Cost of Cloud, a Trillion Dollar Paradox," [Online]. Available: https://a16z.com/2021/05/27/cost-of-cloud-paradox-market-cap-cloud-lifecycle-scale-growth-repatriation-optimization/.

[8] S&P Market Intelligence, "So, the data lakehouse is now officially a ‘thing’ what is it and why should you care," [Online]. Available: https://www.spglobal.com/marketintelligence/en/news-insights/blog/so-the-data-lakehouse-is-now-officially-a-thing-what-is-it-and-why-should-you-care.

[9] Databricks, "What is a lakehouse?," [Online]. Available: https://databricks.com/blog/2020/01/30/what-is-a-data-lakehouse.html.

[10] Databricks, "Delta Lake: Reliability, security and performance," [Online]. Available: https://databricks.com/product/delta-lake-on-databricks.

[11] M. Armbrust, A. Ghodsi, R. Xin and M. Zaharia, "Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics," in *11th Annual Conference on Innovative Data Systems Research (CIDR '21)*, 2021.

[12] R. Pathak, "Harness the power of your data with AWS Analytics," Amazon, 09 Dec 2020. [Online]. Available: https://aws.amazon.com/blogs/big-data/harness-the-power-of-your-data-with-aws-analytics/.

[13] Deloitte, "Data valuation: understanding the value of your data assets," [Online]. Available: https://www2.deloitte.com/content/dam/Deloitte/global/Documents/Finance/Valuation-Data-Digital.pdf.

[14] Gartner, "Gartner Says Worldwide Data Center Infrastructure Spending to Grow 6% in 2021," [Online]. Available: https://www.gartner.com/en/newsroom/press-releases/2020-10-07-gartner-says-worldwide-data-center-infrastructure-spending-to-grow-6-percent-in-2021.

[15] M. Pawlish, A. S. Varde, S. A. Robila and A. Ranganathan, "A Call for Energy Efficiency in Data Centers," *SIGMOD Record,* vol. 43, no. 1, 2014.

[16] Amazon Web Services, "EU businesses that move to AWS Cloud can improve energy efficiency and reduce carbon emissions," [Online]. Available: https://blog.aboutamazon.eu/aws/eu-businesses-that-move-to-aws-cloud-improve-energy-efficiency-and-reduce-carbon-emissions.

[17] ABB, "How data centers can minimize their energy use," [Online]. Available: https://new.abb.com/news/detail/66580/how-data-centers-can-minimize-their-energy-use.

[18] Forbes, "Why Energy Is A Big And Rapidly Growing Problem For Data Centers," 5 Dec 2017. [Online]. Available: https://www.forbes.com/sites/forbestechcouncil/2017/12/15/why-energy-is-a-big-and-rapidly-growing-problem-for-data-centers/?sh=c681b8a5a307.

[19] UCLA, "What is Sustainability?," [Online]. Available: https://www.sustain.ucla.edu/what-is-sustainability/.

[20] Forbes, "The data center of the future," [Online]. Available: https://www.forbes.com/sites/insights-vertiv/2020/01/22/the-data-center-of-the-future/?sh=7522534a5a3a.

[21] AI Impacts, "2019 recent trends in Geekbench score per CPU price," [Online]. Available: https://aiimpacts.org/2019-recent-trends-in-geekbench-score-per-cpu-price/.

[22] PC Part picker, "Price Trends: Memory," [Online]. Available: https://pcpartpicker.com/trends/memory/.

[23] B. Jovanovic, "Internet of Things statistics for 2022 - Taking Things Apart," [Online]. Available: https://dataprot.net/statistics/iot-

statistics/#:~:text=In%202021%2C%20there%20were%20more,in%20economic%20value%20by%202025..

[24] Wikipedia, "Chiplet," [Online]. Available: https://en.wikipedia.org/wiki/Chiplet.

[25] Wikipedia, "System on a chip," [Online]. Available: https://en.wikipedia.org/wiki/System_on_a_chip.