SDIA

Public Cloud Providers

How and why to regulate digital resource utility companies

Public Cloud Providers: How and why to regulate digital resource utility companies

Authors

Thomas Moran
Advisor
Sustainable Digital Infrastructure Alliance e.V.

Max Schulze
Director
Sustainable Digital Infrastructure Alliance e.V.

About the SDIA

Established in 2019, the Sustainable Digital Infrastructure Alliance e.V. (SDIA) is a non-profit network of more than 900 members and partners working to build a sustainable digital economy. The SDIA brings together stakeholders from across industries and fields, both public and private, to realize its Roadmap to Sustainable Digital Infrastructure by 2030 (sdia.io/roadmap). As an independent organisation, the SDIA offers a holistic, systems-thinking approach to solving the challenges facing ICT sustainability, ranging from energy supply and data centres to fibre-optic networks and software.

About the Paper

This position paper comprises publicly available information from white papers, academic studies, industrial publications, and news articles to formulate the most observed challenges, and potential opportunities for a fair and equitable market. Feedback was provided by industry experts before the final draft was released. The options stated in this report are neither exhaustive nor exclusive.

Table of Contents

1. Summary

The cloud market is perceived as complex, especially due to the overcomplicated language found in the Information and Communication Technology (ICT) sector. To understand the cloud market, its business models and infrastructure, it is helpful to compare it to another critical infrastructure sector – the energy sector. The similarities are striking. We can apply many of the regulatory and policy lessons-learned from the development of the energy sector to the cloud infrastructure industry, as the economic issues are so similar.

In 2020, the Sustainable Digital Infrastructure Alliance (SDIA) wrote extensively about the premise of digital resources as a utility, drawing parallels with the development of the energy sector in: The Utility of the Future Report.

Authors Thomas Moran and Max Schulze, representing the SDIA community, build on the work contained within that report to further extrapolate the urgent need to regulate digital resource utility companies.

2. Defining the digital resource utility

Digital resources enable digital products and services to operate. For example, digital resources become useful when 'powering' software applications to deliver digital products and services to customers. As such:

- A digital resource is the commodity produced by (any) IT infrastructure.
- The utility of a digital resource is the **storage and processing** of data. A digital resource is produced using a **digital generator** (a computer or network device) which converts electricity into digital resources.
- These generators are housed in facilities that provide energy and cooling for them to run efficiently — digital resource generation plants. (we often refer to them as **data centres**, but if anything they are data storage and processing centres).

A digital generator is no more than a computer, albeit one specifically designed to maximize the output of digital resources per unit of energy being used.

The analogy to the energy sector is clear: A digital resource plant is the same as a power plant, converting one commodity (e.g., natural gas, coal) into another (electrical power) using large-scale generators. One can have many smaller power plants (decentralized) or large-scale ones that have higher efficiency due to economies of scale.

This is the same with Hyperscale digital resource generation plants (very large-scale, very efficient), Enterprise, or Edge digital resource generation plants (small to medium scale, sometimes on-site). The latter can be compared to having a power plant at the site of a factory.

The transfer of data to digital resource plants for storage, processing, and back is enabled by a network. There are two basic types of networks:

1. **Internet Access networks** connect citizens to remote digital resource generators.
2. **High bandwidth transport networks** connect medium-and-large-scale digital resource generators to each other and sites of large digital resource consumption, such as offices and manufacturing plants.

Comparative to country roads and interstate highways - High-volume connectivity meets at network exchanges, where network and data centre facility companies provide what are typically known as cross connects, to form the interconnections that are the foundation of both transport networks and Internet Access networks.

The analogy to the energy sector is again clear. The internet access networks for citizens are equivalent to a District System Operator (DSO) and the transport networks are transmission lines, Transmission System Operators (TSOs).

Unlike the energy sector, however, network connectivity is not universally available to everyone, and can be provisioned in different ways and at different speeds. Hence most customers, both enterprise and private, have personal digital generators (devices in the form of home computers, gaming consoles, Smart TVs, smart-phones, or laptop computers). Each device is capable of generating its own digital resources to provide a service or can utilise digital resources produced by a remote digital resource generation plant (data centre) when connected to a high-speed network.

Therefore, a digital resource utility is an infrastructure function which provides the storage and processing of data to businesses and customers alike via transport networks.

3. Traditional digital resource utilities

Digital resource utilities are not new, nor have they emerged from cloud providers.

Before what we now refer to as the public cloud - a global supercharged digital resource utility, there were already many local, regional, and multinational operators offering similar resources and services. In addition, many businesses provided internal services with their own employees and facilities. These IT infrastructure businesses and teams grew organically with demand, supporting both enterprise and consumer services with solutions such as websites, e-commerce hosting, etc. They were akin to municipal utilities offering power, water, and **network access** within the region.

Local and regional digital resource utilities are now being driven to consolidate to remain competitive. Many still exist, though their role is diminishing with the force of the supercharged global digital resource utility emerging (more on that below).

It is important to highlight that these local utilities, however, represent a resilient infrastructure that responds to the needs of a region, and provides local IT expertise that can help a region along in its digital transformation. This is a value that can never be generated by large-scale centralised infrastructure, especially when digital transformation is still in its infancy and every region needs to be taken along on the journey.

With a global shortage of skilled labour within the IT sector, especially in areas such as cyber-security, the economic leverage of large corporations is enhanced through their ability to pay higher wages, which drains talent out of smaller companies and less prosperous regions, increasing risk and weakening local economies.

Because local and regional operators can provide a broader variety of services, which are tailored in response to local needs, they also provide greater flexibility and resilience than cloud providers, which offer only standardised, and fully integrated services that are designed for scale and not interoperability.

4. It's not a Public Utility, yet we call it so

There is caution, and a public duty not to mislead people within the vast complexity of the digital economy and the ICT sector. What the global digital resource utility companies have accomplished is quite a feat, as we refer to them as **Public Cloud Providers**, creating the illusion that the infrastructure they provide is somehow public, a public good, or publicly owned.

Among the general public, there is a common misconception that the Internet is a public utility, or form of common ownership. In fact, the Internet is comprised of private networks owned by corporations that would be better understood as a set of standards for interoperability rather than an infrastructure platform. The term “Public Cloud” was clearly intended to build on this misconception and position the large IT infrastructure platforms provided by AWS, Microsoft, and Google as another form of public utility.

The IT community adopted the terms largely because their use of the term “public” serves to distinguish these resources from those that are “in-house / private” (only accessible over a private corporate network).

Creating the illusion of a shared “public” utility serves to obscure the entirely private ownership of virtually all ICT infrastructure. This broad misconception of the Internet and now the digital resource generators that support it as a public utility, serves a very specific strategic objective — which is to associate it with the perceived public benefits of “digitalisation”.

Making the two synonymous creates an aura of progress around all things digital, that serves the needs of the technology industry, and its biggest players the most. Because the new infrastructure platforms of the digital age are perceived to be public, they are trusted.

Because they are almost entirely in the hands of private companies and unregulated, this trust based on a lack of knowledge is becoming increasingly dangerous. The danger is compounded by the fact that private companies controlling a key portion of critical IT infrastructure underpinning the digital world, are rapidly being consolidated into the hands of just three American owned mega-corporations, with economic might that dwarfs most countries.

5. The supercharged global digital resource utility: Multinational Public Cloud Providers

Now let's look at how the supercharged digital resource utility emerged.

To build a digital resource utility, one needs to invest in digital resource generation plants and digital generators.

Digital generators (computing or ICT Equipment) have a life cycle of 3–5 years (depreciation period), while the facilities that houses them (data centre facilities) typically have a life cycle of 15–20 years, making digital generators an unattractive asset to invest in. As a result, investors are pouring billions (estimated to reach $432 billion per year by 2025) into building new facilities (also investing in them as an asset class)[1], but only a few investors are open to investing in ICT equipment, resulting in a lack of access to capital for competitors of hyper-scale cloud providers.

To finance digital resource generators through banks or investors, one needs to secure long-term customer contracts (over the whole depreciation period). However, companies are used to buying digital resources on-demand from Public Cloud Providers with no commitment. This "pay as you go" model further strengthens the position of supercharged cloud providers, even though for many companies the costs are much higher using hyper-scale cloud infrastructure in comparison to traditional IT infrastructure and operation models[2].

To offer on-demand digital resources requires high-level investments into building digital resource generators, and to assume the potential risk of under-utilisation. That is where it becomes useful to be vertically integrated. Each of the Public Cloud Providers has a cash-generating business at its core:

- Google:
    - Main Cash Generator: Advertising
    - Free Cash Flow: $60.01 Billion (2022)

---

[1] https://www.prnewswire.com/news-releases/global-data-center-market-report-2021-growth-opportunities-in-edge-computing-investmentma-new-capabilities-geographic-expansion-partnerships-artificial-intelligence-liquid-cooling-renewables-301231802.html

[2] https://www.vpsbenchmarks.com/compare/azure_vs_hetzner

- Microsoft:
  - Main Cash Generator: Software Licensing
  - Free Cash Flow: $65.14 Billion (2022)
- Amazon / AWS:
  - Main Cash Generator: E-Commerce
  - Free Cash Flow: $-11.56 Billion (2022), $31.02 Billion in 2019

These cash-generating core activities enable hyperscale Public Cloud Providers to invest in digital resource generators at an unprecedented scale — accepting the risk of under-utilisation. And, because all their ancillary business activities are digital as well, it means they are also a large digital resource consumer. This means they can mitigate the risk of under-utilization simply by using the digital resource generators themselves.

This creates a perfect circle: Generate cash using core-activity, invest in digital resource generators and build a supercharged utility, which generates even more revenue.

In Q1 2022, Amazon Web Services (AWS), the supercharged digital resource utility of Amazon, generated nearly all the corporations $6.5 Billion in profit.

To reiterate: Without the cash-generating core business, without the vertical integration, it would be economically unfeasible to build a supercharged digital resource utility of the scale we have today. And since very few businesses generate as much free cash flow as the big tech companies do, it’s unlikely we will see any significant competition in the future.

This consolidation of hardware and infrastructure in the hands of mega-corporations that also provide software, services such as enterprise applications, and advertising, compounds the fair market risks associated with the size and economic leverage of these companies by reducing choice and competition.

6. The internal market design of the supercharged utility

In the energy sector, we have created marketplaces to a) ensure that all our power generating infrastructure is efficiently utilised and b) create transparency on price and demand signals. The supercharged digital resource utilities have replicated those markets, without transparency on price, demand, or efficiency of utilization. (For example, one can buy from a hyperscale public cloud provider as a customer, but that same customer as the owner of a digital resource generation plant, cannot sell back into their market.)

This market design can easily be observed in the pricing of their digital resource offerings:

On Demand: Highest price, no upfront costs, pay-as-you-go with guaranteed supply availability. Equivalent to Day Ahead markets in the energy sector.

- Reserved: Lowest price when committing to 1, 3, or 5 year contracts with fixed digital resource capacity included. Essentially, a futures market.
- Spot: Buy under-utilised capacity at low prices when available (fluctuating supply). The energy market equivalent is the Intraday market.

It is clear that this market is designed to maximise the utilisation rate of the digital resource generators, which is the main profitability driver in any infrastructure sector.

When the supercharged utility first emerged, it started with the on-demand model, putting their digital resource generators at risk. This disrupted the existing digital resource market, which before was built on at least monthly or annual commitments (Reserved or Futures) offered by regional and national digital resource utilities. It enabled the supercharged utilities to take market share quickly, especially in the start-up sector, where upfront costs are an issue for early-stage companies.

The ability to invest in digital resource generators at risk, and the willingness to buy market share is evident in a scheme offered by AWS to start-ups (the same is available from Google and Microsoft) who also offer up to $100.000 of free digital resource credits.

Few national or regional actors can offer equivalent financial incentives, which require large capital investments, without the surety of customer contracts. For these smaller companies, debt is not an option because they are not vertically integrated, and do not have a built-in revenue generator.

7. Packaging digital resources with services, for customer lock-ins and building demand

In the early days of the energy sector, it was common that utility companies (for example, General Electric in the USA) would also sell fridges, electric ovens, and other electricity consuming appliances to their customers.

This practice created a predictable base-load of utilisation for the electricity generators they had invested in. This practice is known as ‘load building’.

Today, the supercharged digital resource utilities use the same ‘as-a-service’ strategy to sell digital resources. These cloud services are essentially a prefabricated, convenient way for IT departments and developers to build an application like building blocks - picking prefabricated components as services, and then programming their specific logic on top of it.

Cloud services are even more effective at load building than fridges, as the latter was not combined with a power contract.

When purchasing a fridge, customers were able to change their electricity provider — with Cloud services the digital resource utility and ‘as-a-service’ are bundled together as a package.

Specifically, a cloud service is usually made of two components:

- An existing piece of software, such as a database, most of them open-source (public goods, developed for free by the open-source community).

- An instruction (written by the cloud provider) on how the software can operate using the digital resources of the cloud provider.

The open-source software is free, (for example Kubernetes), and packaged by Amazon as 'EKS' and by Microsoft as 'AKS'.

This bundling is very favourable for the providers, as cloud services are not portable, meaning they cannot be run using other provider's digital resources. For example, a regional IT infrastructure provider cannot offer the digital resources it has generated to run the cloud provider's services.

In contrast, anyone originating software who is willing to let cloud providers bundle it with their digital resources, can do so within their marketplaces, under terms set by the provider. The terms and conditions of these "walled garden" marketplaces have come under increasing scrutiny from regulators.

Cloud providers have become so proficient at taking open-source software, and packaging it together with their resources, that they have launched more than 300 such services. AWS alone launches 50-100 new services per year.

Subtle differences in their approaches are worth noting:

- Microsoft: Will incentivise enterprises to move email and office services to Microsoft Azure (their digital resource generator) to increase the utilisation. Additional "value add" enterprise services such as Microsoft Teams are only available to customers who have moved their Office environments to Azure.
- Amazon AWS: Focuses more on supplying services that can be assembled into new IT applications - e.g., targeting tech start-ups, software development agencies, and even governments who want to be able to build new applications quickly, using the prebuilt services (components) that AWS provides.
- Google Cloud: Focuses on making its data analytics and AI capabilities available as cloud services (e.g., for text-to-speech, speech processing, translations, etc.) to increase the utilisation of their digital resource generators.

What these providers have in common is that revenue generated from operating their digital resource infrastructure is re-deployed to build more services and create more bundles (packaged products and services) that are designed to "lock-in" customers to that providers platform.

8. Incentivising IT consultants

There is another path that enables rapid-scale load building, and that is to incentivise the people who are helping companies build IT applications. Many small and medium enterprises (SMEs), corporations, and governments rely on external expertise for building new applications, especially as part of their digital transformation.

From the onset, the Public Cloud Providers made sure that all large IT consultancies around the world, for example Atos, Cap Gemini and Accenture are strongly incentivised to do two things:

1. Build all new applications for clients using cloud services (bundled with digital resources and lock-ins).
2. Heavily promote that all 'legacy' applications should be rebuilt or migrated into the public cloud so that everything is using only digital resources produced by the large providers.

It is normal practice to build 'channel' or 'reseller programs', yet the scale at which the incentives are offered, and the lack of transparency is unique. From the perspective of the cloud provider, higher customer acquisition costs are offset with a model that locks customers into the cloud providers 'grid' for the longest duration possible.

Based on the potential for future revenue from a "locked-in" customer, AWS, Microsoft, and Google all offer $20,000 to $100,000 of service credits to new customers, which one can assume they believe will be offset by future revenue.

Smaller regional and national digital resource utility companies cannot compete with cash rich companies who can 'start the growth machine' with revenues from vertically integrated business models that are the financial basis for public cloud infrastructure — advertising, licensing, and e-commerce.

9. By owning the majority of digital resource generation capacity, any form of digitalisation is perceived as a good thing

Critical infrastructure industries such as energy and water, are generally either large "public" utilities or large enterprises that can manage mission critical operations at scale in a highly regulated environment. The first wave of ICT infrastructure comprised telecommunications networks that started out as public utilities which have since been privatised and deregulated to varying degrees.

During this time, ICT infrastructure has become increasingly critical to society, especially since the introduction of computers (digital generators), data centres (digital resource generation plants), and the Internet.

While the regulatory environment for telecommunications is still mostly intact, there is no comparable regulatory framework for the rapidly advancing networking, compute, and storage functions of ICT infrastructure. Multinational private enterprises that provide "public" cloud at scale have used this absence of regulation to build massive operations and leverage their economic power to make competition increasingly difficult for smaller companies.

While the benefits of digitalisation are significant, the anti-competitive behaviours of large enterprises that are increasingly dominating digital resource utilities are stifling innovation and creating ever greater risks for society.

When private companies like AWS have an outage the economic impact can be massive, as well as the knock on effect and even greater impact to public services, that are increasingly dependent on these infrastructure platforms.

Many government entities are now migrating mission critical systems on which their citizens depend on, to these public cloud platforms without sufficient oversight. This combination of reduced competition and increased risk highlights the need for a more robust regulatory framework.

For example, government departments (digital generators) that produce digital resources are required to run the digital economy, digital applications, digital processes, and so on. Therefore, Public Cloud Providers that hold a de facto oligopoly are quickly extending their influence into the public sector. One could easily compare the position of power for the new ICT industrial complex to that of the US military industrial complex.

While the potential benefits of digitalisation are significant, the risk of unregulated platforms such as social media are also readily apparent. Disinformation campaigns have influenced elections across the world. Whilst the risks associated with the underlying platforms of the digital era are less visible, they every bit as real in terms of their impact.

There is a well understood set of material issues that impact the **sustainability** of ICT companies, but virtually no regulation around these issues as politicians struggle to keep up with the pace of technological change. The economic might of these same companies also has profound implications on political decision making, as evidenced by the fact that Google **spends** more money on lobbying than any other US company.

The ICT industrial complex is sponsoring, financing, and supporting any initiative from governments or industries that is pushing digitalisation. All in the spirit of innovation, yet with a simple motive: The more digitalisation achieved or aspired to, the more digital resources are needed, the more digital resources that are made, bundled, simplified, and locked-in to platforms, intentionally designed to make them difficult to move.

10. Then they build a moat around it

Supercharged digital resource utilities have established market dominance in most of the developed world and are quickly establishing similar positions in the developing world by leveraging their resources to "subsidise" services for cash strapped governments. Extended public procurement contracts and the challenges of moving large, complex, and highly integrated platforms make these investments very lucrative in the long term.

Energy companies have similar positions, but there are fundamental differences in how the energy markets operate that make the costs and risks for citizens much lower. Transparency around price, supply, and demand as well as mature and strong regulatory frameworks enable regulators to control energy markets in ways that serve the public interest. It is vital that we develop and implement similar schemes in the ICT sector before the economic might of a few multinational companies makes both regulation and competition impossible.

The regulatory framework of the telecommunications sector provides a foundation for ICT regulation in the same way that the highly regulated energy distribution grid ensures safe and sustainable operation. The most pressing example of the need for (network related) regulation of ICT infrastructure use by cloud providers is in the area of egress charges.

For telecommunications providers the unit cost of moving data is unrelated to the direction of travel, even in cases where bandwidth availability is asynchronous (e.g., upstream, and downstream bandwidth capacities are different). With Public Cloud Providers there is a massive differential between free "inbound" traffic and expensive "outbound" traffic that serves as an export tax.

The obvious intention of what are commonly known as "egress charges" is to discourage to the point of economic infeasibility the removal of data from the providers platform. This is despite the fact that the cost of transporting data out of the platform is no different than transporting data onto the platform for the provider.

One could get the impression that Public Cloud Providers would like all the data to stay within its own infrastructure — where it can only be stored and processed by its digital resource generators.

For example, based on an analysis performed by Cloudflare, the "export tax" mark-up for AWS outbound network traffic ('egress charges') is between 357% to 7959% (as shown in the Table 1)

Estimated average AWS utilization per month	20%			
AWS Region	**AWS Cost at 1TB/month***	**Implied Mbps $**	**Estimated Regional Cost Per Mbps**	**Markup**
US/Canada	$92.07	$6.37	$0.08	**7959%**
Europe	$92.07	$6.37	$0.08	**7959%**
India (Mumbai)	$111.82	$7.73	$1.00	**773%**
Singapore	$122.76	$8.49	$0.50	**1698%**
Korea (Seoul)	$128.90	$8.91	$2.50	**357%**
Japan (Tokyo)	$116.63	$8.07	$0.50	**1613%**
Australia (Sydney)	$116.63	$8.07	$1.00	**807%**
Brazil (Sao Paulo)	$153.45	$10.61	$0.50	**2122%**
	* Taken from the official AWS "Simple Monthly Calculator" as of July 21, 2021			

Table 1: Report by Cloudflare on AWS egress charges (Source: Cloudflare)

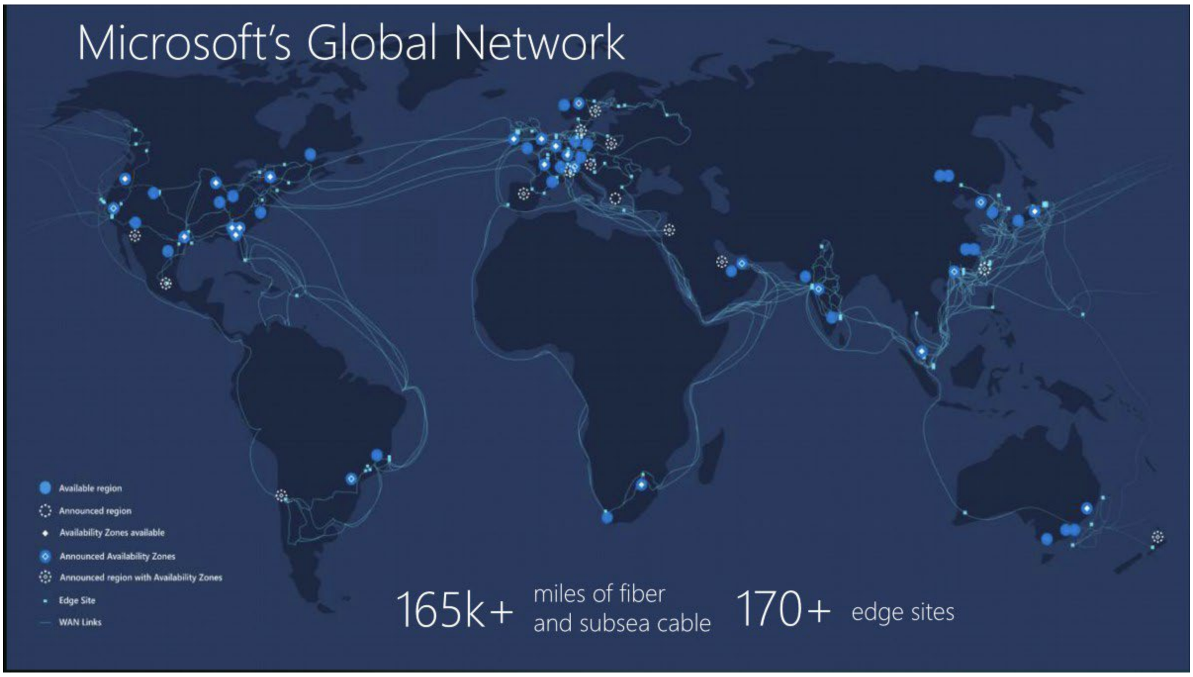
This export tax serves as a moat that is often overlooked when 'migrating' into the Public Cloud Providers infrastructure or when building 'a digitally transformed business' using cloud services from digital resource utilities.

Alternatively, these pricing policies are a transparent attempt to create a "moat" around the providers platform, the cost and risk of which is often overlooked when services are migrated onto the platform.

This is clearly anticompetitive behaviour intentionally designed to block customers from using the digital resources of a competitor as the cost of the data transport would make any other digital resource offering orders of magnitude more expensive.

The intent of the data transport pricing policies is further evidenced by the fact that movement of data between locations on the providers platform does not incur these same egress charges.

In short, the economic model for data transport in public cloud pricing policy is the same "walled garden" approach used in software application marketplaces and other areas of ICT. In such instances, open standards and interoperability have clear and well documented benefits that would better serve the public interest.



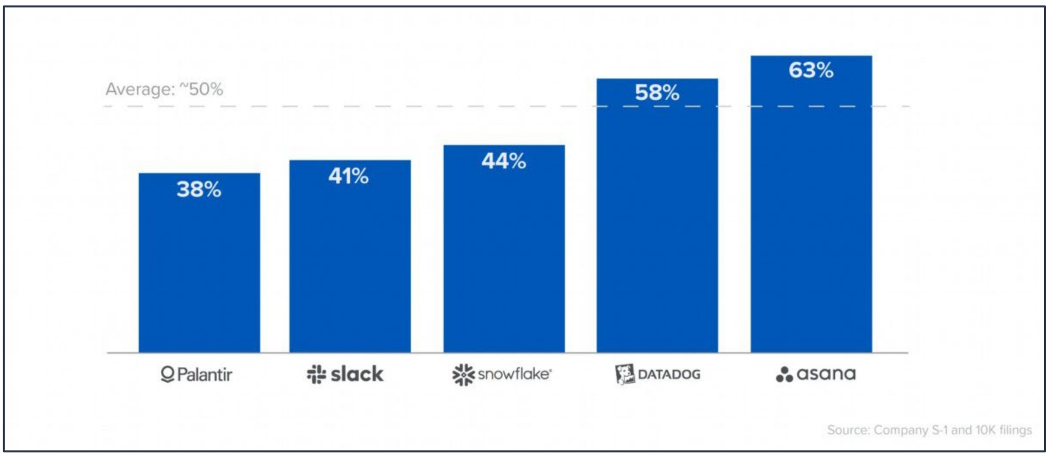
Presentation by Microsoft at the European Commission, Wednesday 14 April 2021

11. The Public Cloud Tax

As the public cloud creates the illusion of being public infrastructure, it is also collecting taxes from different countries, and nations.

It is reasonable to assume that technology savvy companies such as software start-ups have the capabilities to optimise their usage and spend related to IT infrastructure. Despite this fact many of these companies spend 50% or more of their operating costs on public cloud platforms. The operating cost model for digital infrastructure services and platforms (digital resource utilities), creates larger revenue streams for the platform providers (Public Cloud Providers), than for the software companies. Which makes those companies perpetual targets for acquisition by Public Cloud Providers looking to further strengthen their market position through reduced competition and increased economies of scale.

As their services mature so that more of the component parts and the software of acquired companies are integrated into the core platforms, the revenue, the lock in related export taxes, and the cost of migrations will increase even further.



Estimated Annualized Committed Cloud Spend as % of Cost of Revenue (Source: A16Z)

Based on the current unregulated model for cloud services the most likely outcome is that virtually all private and public ICT platforms for compute, storage, and processing as well as most of the software applications, will end up with just one of three companies. It is also worth noting that despite their economic might these companies have avoided acquiring telecommunications providers, most likely because this would open them up to more regulation and more intense scrutiny from governments.

Taking into consideration the opaque corporate setups of most of the vertically integrated digital platform companies such as Google, Microsoft, or Amazon AWS it is questionable if the profits from operating the Public Infrastructure are taxed in the countries where the infrastructure is used or in the United States.

12. Options to act:

1. Forcing the split of the digital resource utility from the digital platform business (e.g., from advertising, search engines or, e-commerce) – OR – Building a regulatory framework for the data processing, storage and transport functions of ICT infrastructure that extends concepts already well established in telecommunications regulation such as ownership restrictions, neutrality and price/tariff based controls.
2. Creating a public market for digital resources — which enables any owner of digital resource generators to sell into it, and any provider of services to buy digital resources from it — with transparency, and fair competition on price and efficiency.
3. Restricting the bundling/ packaging of cloud services with digital resources or enforcing that all produced digital resources are sold to a public market, and services must buy from the market (analogous to the energy market).
4. Eliminating or restricting the use of "export tax" egress charges for data leaving a cloud platform — for example, by requiring all data transport to be priced the same regardless of direction.
5. Restricting the use of anti-competitive discounting and credits (e.g., giving $100.000 incentives), or repurposing the schemes to increase rather than decrease competition by incentivising the use of local or regional providers.
6. Matching the transparency of ownership, usage and tariffs that exists for most telecommunications networks for all ICT infrastructure that is or has the potential to be "Critical Information Infrastructure" (CII).