Data Centre

Week 3

Formal Definition:

- A data centre is a building (or self contained unit within a building) used to house computing equipment such as servers along with associated components such as telecommunications, network and storage systems.

- A data centre is equipped with a guaranteed power supply and high bandwidth connectivity.

REDUNDANCY - NETWORK, POWER [CRITICAL day to day operations]
• stock markets [transaction is in miliseconds], Air asia promotion [zero fares] -> waiting room

- Resilience is critical so redundancy (duplication) of networks, power and other infrastructure is common to ensure continuity.

Network: Unifi, Maxis, TIMES, Digi, Celcom
power -> TNB, USP, Generator

Redundancy === $$$$

- Other facilities include building management controls such as air conditioning to maintain the environmental conditions for the equipment within a specified envelope of temperature and humidity, and security systems to ensure that the facility and its data remain secure.

What does it actually mean?

- Data centres house computers so that they can
  - do stuff like process, manage, store and transmit data and
  - talk to each other or to digital equipment in other data centres or in offices, homes, vehicles, in satellites orbiting the Earth, on the moon or in fact anywhere you can think of.
- Business processes, government services, telecommunications, transport infrastructures all depend on computers interacting in this way, exchanging digital information.
- Many of your everyday activities also rely on data centre processes, including obvious things like using your smartphone or the internet or sending email, and less obvious things like doing your shopping or catching a train.

(Data Centres: a Day In Your Life).

Malaysian Genome Institute - Bangi Lama

Data centers with 100,000+ servers

power == heat generated
-> cool it down?
-> use what to cool it down?
-> use air-conditioner
-> generate heat

SDG [sustainable]
-> green energy
-> innovative technology
-> use water to cool it down
-> water shortage issue

DRAM -> Dynamic Randam Access Memory

What is a Data Center?

- The kind of computers you find in data centres are known as servers (computers that are usually assigned to specific roles as opposed to personal computers which are more generic).

- Servers perform computing functions remotely from their operators and therefore don't have keyboards or screens or people hunched over them in the way that desktops would.

Typical server:
- 8-128 cores
- DRAM: 32-512 GB

Typical storage:
- 30 drive

Typical Network:
- 72 10GB

What is a Data Center?

- Servers run on electricity so data centres need to have a good power supply that will not get disconnected.
- This is called UPS or uninterruptible power supply.
- Back-up power supply is usually provided to make sure that even in the event of a power cut to the grid, the data centre is not left without electricity.
- Huge batteries provide instantaneous backup and there are usually generators on site with guaranteed fuel supply to provide for longer term outages.
- This is because so many of the critical functions of a modern digital economy rely on data centres that a power failure could have catastrophic consequences.

What is a Data Center?

- Servers also need to talk to other digital devices so data centres need to have excellent communications infrastructure in the form of high bandwidth connections (super-ultra-mega-fast-broadband), often from more than one telecommunications provider to ensure that if one option fails, connectivity is not lost.

- Essentially a digital economy is a networked economy and data centres can be seen as network nodes.

What is a Data Center?

- It might help to think of a data centre as a bit like a brain, which stores lots of information and runs both cognitive functions - the things we realise we are doing, and unconscious functions – the invisible things that we take for granted like breathing and keeping our hearts beating.
- If the oxygen supply to the brain is cut off then the consequences are rather significant. In the same way a data centre needs its own form of oxygen - a continuous, reliable power supply.
- And just like a brain has a nervous system to tell heart and lungs and arms and legs and eyeballs what to do, data centres need their own nervous system – telecommunications connectivity - to keep things like ATMs, street lights, telephones, hospital computers and air traffic control running smoothly (to name just a few).
- This is also the reason why data centres have to be highly resilient to external threats whether from terrorist attack or environmental cataclysm.

Why do we have data centres – What are they for?

- Data centres consolidate any number of separate IT functions within a single operating unit.
- Twenty years ago there were no data centres – or at least none as we know them today.
- That's probably because there wasn't enough digital data to create a requirement for specialist facilities in which to house and process it, and the data that existed did not underpin enough critical government, business or social functions to make its integrity and security the key priority that it is today.
- The growing presence of data centres is the result of our increasing reliance on computing and on digital technology generally.
- More and more of our everyday processes, including government service delivery, business processes and individual activities rely on computing to function.
- The growth in data centres is also the result of changes in the way that we approach computing and how we choose to handle our computing resource.

Why do we have data centres – What are they for?

- Twenty years ago, big, complex computing jobs tended to be carried out by big, complex computers called mainframes that were housed in special facilities, usually in office buildings.
- Then as computing became more and more accessible, and individual processors became more and more powerful and capable, the model of “distributed computing” developed.
- Distributed computing just means that the computing functions are spread around according to where they are needed instead of being concentrated in one place.

Why do we have data centres – What are they for?

- More recently, as computing – and society – has become increasingly networked (see jargon buster on LANs and WANs) we need additional computers to manage these networks, collate and back up all the data from individual computers in one place, and manage applications like email, databases and internet presence.
- The computers that do these specialist functions are called servers and tend to be located away from people in server rooms or cupboards.

Why do we have data centres – What are they for?

- During the last ten years two things have happened: Firstly the demand for digital data has grown explosively, (especially the demand for moving it around) and secondly, as computing has become more pervasive, we are increasingly reliant on it.
- If those computing processes fail for some reason or if data is destroyed or stolen the consequences could be catastrophic for business, government or individuals.
- Moreover there is the threat of punitive damages if this failure interrupts the provision of services to third parties.

Why do we have data centres – What are they for?

- These two factors, (rapid growth in the volume of data we need to deal with and our reliance on it) have made us reconsider the way we handle our computing functions.
- Locating servers in offices along with other corporate activities has the appearance of keeping all our eggs in one basket and could make a company very vulnerable to mishap.
- Unlike laptops or PCs, servers perform specialised functions so don't have to be close to the people who rely on them provided they are connected by a telecoms network.
- So it makes sense to consolidate and protect these critical business functions in separate, specialised, secure facilities.

Why do we have data centres – What are they for?

- Another benefit of consolidating computing in this way is that it is much more energy efficient than distributed computing.
- The result of all this has been the rapid growth in data centres. Some companies have built their own, others use third-party providers. In fact a range of business models exists in this rapidly emerging market.

Expense and variable demand

- Building a data centre from scratch is exceptionally expensive.
- Moreover, customer requirements vary enormously.
- Some customers just want a shell within which they can install all their own equipment.
- Some want a fully fitted space and others want a complete IT service.
- The expense and the variability in demand are two factors that have influenced the way that the data centre market has evolved.

Space

- Data centres are described by size in terms of their gross space, their technical space and their core IT space (also known as white space or raised floor or net technical space).
- Gross space is the total footprint of the building in which the data centre is housed (sometimes even the total footprint of the whole site, car park and all, or the part of the site for which planning permission has been obtained), white space is the space reserved just for the IT equipment and technical space includes white space plus the supporting technical infrastructure (chillers, generators, network rooms and controls).
- Some operators view technical space and white space as the same thing and some use the term "data centre space" which could probably mean any of the above. So the definitions are not rigid.
- White space is the most important measurement for a data centre because this is the space available to house IT equipment in the form of racks of servers. This is essentially the area that is available for lease or hire in colos so you can see why it is the most useful measure. In a data centre with 2,000 M2 of gross space, only about 50% of that is likely to be white space. White space in turn is only about 70% of technical space.
- As if this isn't confusing enough, data centre operators in the UK tend to mix units when describing data centres and use both metres and feet almost interchangeably. So a site might be described as 20,000 ft2 of gross space with 10,000ft2 of white space and with a power supply of 1.4 KW per square metre.

Power

- Data centres are starting to describe their offerings not in terms of space but in terms of total power available to a site, which is known as power provisioning.
- Power availability is also frequently described in terms of the power available per square metre or per rack (expressed as KW per M^2 or rack or sometimes in Watts per square foot) or the total amount of power available to a site (expressed as MW or MVA, or for small sites, KW or KVA – see jargon buster for KVA/MVA).
- A site with 10MW of power provisioning means that the operator has a guaranteed power supply and back-up provision of 10MW.
- The proportion of this power that they actually consume varies considerably from site to site depending on occupancy levels and the way it is operated.
- Power in W – Watts - represents instantaneous power demand. Total power consumption is measured in Watt hours, Wh, or KWh, or MWh. One Wh represents a continuous demand of one Watt for an hour. 1000Wh is equal to 1KWh and so forth.

kVA - Kilo-Volt-Amperes: measure of apparent power; the total amount of power in use in a system. In a 100% efficient system, kW = kVA

kW = amount of power that is converted into a useful output; known as actual power or working power

https://bit.ly/3NGTw6l

Power

- You might think that you could calculate the total power demand of a site by multiplying its power feed in MW by 8760 (24x365) to get an annual electricity consumption figure in MWh (MegaWatt hours).
- However, you would be wrong. The total power demand will never get anywhere near this figure because data centres are always heavily over-provisioned.
- When installing (provisioning) power, operators obtain the maximum power capability that they can. Firstly they have to provide power for cooling the servers as well as powering them and have to provide enough on the basis that the facility is full and that every server is working flat out and that it is the hottest day of the year and all the cooling equipment is flat out too.
- They will also be thinking about growth up to 20 years ahead. So you can safely assume that power consumption will always be much lower than provisioned power.

Resilience and security

- Data centres are designed to help ensure business continuity.
- However, some data is more mission critical than other data so there is a grading system to indicate the level of resilience a data centre can offer using categories called tiers.
- This system has its drawbacks because it is based on American criteria for electricity supply which do not match those found in the UK.
- Tier 1 is the simplest and may just be a server room.
- The highest level is Tier 4, equipped to host mission-critical data and systems.

Resilience and security

- Tiers 3 and 4 require top level security (usually with biometric entry controls) and uninterruptable power supplies – in which the back-up power supplies have their own back-up systems, and multiple broadband connectivity options involving a variety of routes and suppliers.
- Duplication of power supplies and connectivity to improve resilience is known as redundancy.
- Such data centres have very high levels of redundancy and there may also be a duplicate data centre facility at a separate geographic location to guarantee continuity in the case of an extreme weather event or other catastrophe.
- Operators also try to identify and minimise what they call SPoFs – single points of failure – to ensure that data centre function cannot be compromised by the failure of just one supporting system.

24/7/365

Data Center Tiers Compared

The table below offers a high-level overview of the four data center tiers and shows what different models provide to clients.

PARAMETERS	TIER 1	TIER 2	TIER 3	TIER 4
Uptime guarantee	99.671%	99.741%	99.982%	99.995%
Downtime per year	<28.8 hours	<22 hours	<1.6 hours	<26.3 minutes
Component redundancy	None	Partial power and cooling redundancy (partial N+1)	Full N+1	Fault tolerant (2N or 2N+1)
Concurrently maintainable	No	No	Partially	Yes
Price	$	$$	$$$	$$$$
Compartmentalization	No	No	No	Yes
Staffing	None	1 shift	1+ shift	24/7/365
Typical customer	Small companies and start-ups with simple requirements	SMBs	Growing and large businesses	Government entities and large enterprises
The main reason why companies select this tier	The most affordable data center tier	A good cost-to-performance ratio	A fine line between high performance and affordability	A fault-tolerant facility ideal for consistently high levels of traffic or processing demands

1. Concurrently maintainable: does not require a total shutdown during maintenance or equipment replacement

2. compartmentalization: limiting of access to information to persons or other entities who have a need to know it in order to perform certain tasks.

https://bit.ly/3qZvvha

Tier 4 data centers

Features	Characteristic
Fault tolerance	designed with **multiple active power** and **cooling distribution paths**
Redundancy	have **redundant components**, such as backup generators, UPS systems, and cooling equipment
Security	have multiple layers of physical and digital security measures, including **biometric access controls, video surveillance, fire suppression systems, and intrusion detection systems**
High availability	have a **guaranteed uptime of 99.995%**; designed to have **no more than 26.3 ~~seconds~~ *minutes* of downtime per year**.
Scalability	designed to be easily scalable, allowing them to **quickly adapt to changing business requirements and growth**
Energy efficiency	designed with energy-efficient technologies and practices, such as **free cooling, hot and cold aisle containment, and virtualization, to reduce energy consumption and costs**

Resilience and security

- Anonymity and a general vagueness over location also contribute to resilience because companies may be keen to ensure that the precise coordinates of their mission critical data remain unknown.
- Data centre contracts between freeholders and operators or between operators and enterprise customers typically include strict non-disclosure clauses, especially if the customers are from the financial services sector.

Key Facts

- Data centres underpin and enable the digital economy: they are its physical manifestation
- Data centres are the only geographical hook connecting the digital economy to a physical location
- A single data centre generates multiple levels of economic activity
- A single large data centre can house the IT function for thousands of businesses
- The UK currently dominates the European data centre market with around 60% of market share, spread between 250-300 sites with a combined power demand of 2-3TWh per year

Why data centres matter?

- Data should be regarded as one of the key utilities of the 21st century, along with energy and water.
- In the same way that governments prioritise the efficiency and security of their energy supply and distribution, they must recognise in data centres as a key technology, resource and skill set that should be retained within national and economic borders.
- All high tech manufacturing and knowledge economies are dependent upon ICT and on data centres in particular.

China and Hong Kong Data Center Market to Reach $35.11 Billion by 2027. Amazon Web Services (AWS), Tencent, Huawei, and Alibaba are the Major Hyperscale Investors– Arizton

Scope of the China and Hong Kong Data Center Market Report

Report Coverage	Details
Market Size (Investment)	$35.11 Billion
Market Size (Area)	7.40 Million Sq.ft
Market Size (Power Capacity)	1505 MW
CAGR	2.27%
Base Year	2021
Forecast Period	2022-2027
Geography Analysis	China and Hong Kong

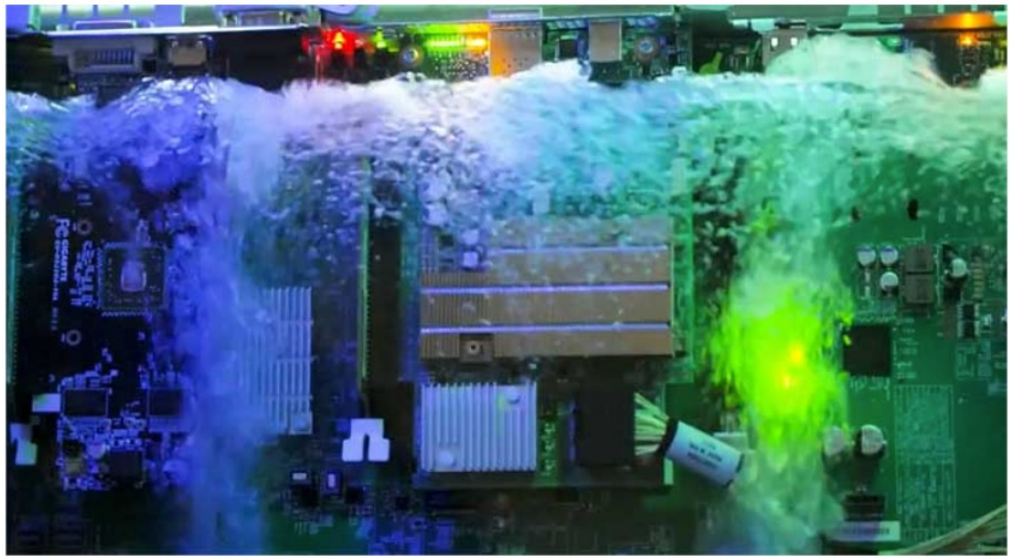
https://bit.ly/3uVJEgd

Water to cool it down -> not feasible

mineral oil -> heat resistant > 100 degree

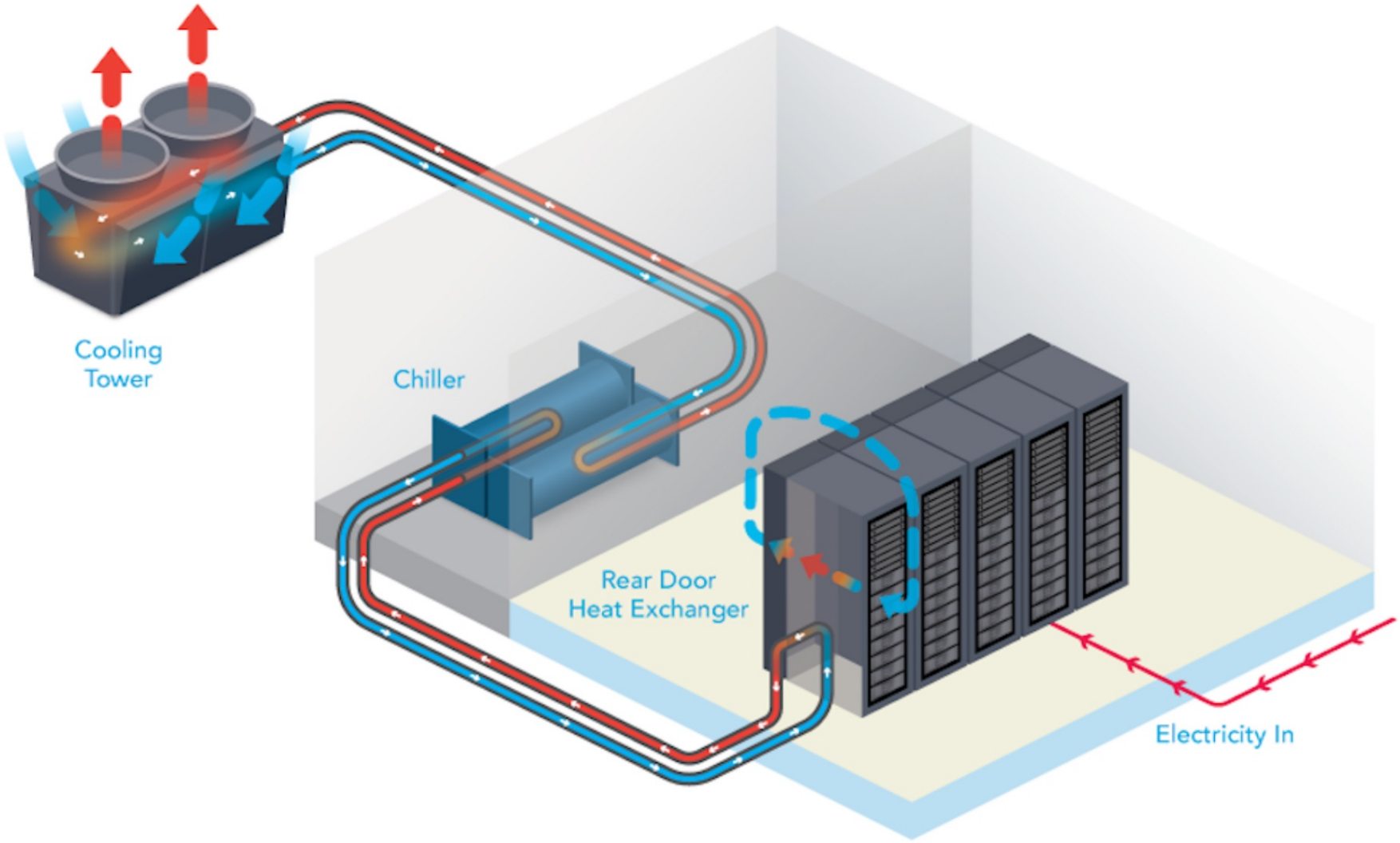
Liquid Cooling Gains Wider Adoption (such as mineral oil)

- Several factors are prompting a reassessment of cooling
- powerful new hardware for AI workloads, growing pressure to eliminate water use in cooling servers, and major progress in enabling liquid cooling from cloud and colocation providers, as well as server and chip vendors. non-conductive coolant
- Air cooling systems simply can't cool these high-density racks in an efficient, sustainable manner



Microsoft immersion-cooled servers

Liquid cooling in data centre



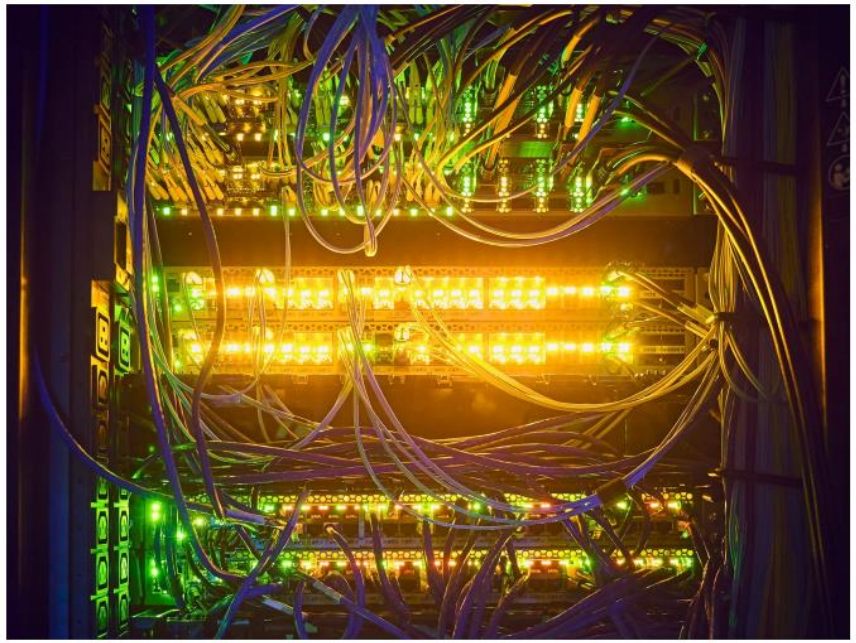
https://bit.ly/3LEN6CO

Article • Sustainability

Microsoft's Zero-Water Solution for Data Centre Cooling

By James Darley
December 12, 2024 • 3 mins

SHARE



Microsoft has announced plans to implement water-free cooling systems across its new data centre developments from August 2024 | Credit: Getty

COMPANIES
Microsoft ›

https://tinyurl.com/mwubhdtw

Data and Energy Forge Deeper Connections

- driven by demand for renewably-powered data centers and the deep pockets of global investors.
- develop renewable energy projects to support its data center campuses, a move that will help its customers reduce their environmental impact. *[https://bit.ly/3r3hVZT]*
- Microgrids are shaping up as a key enabler of on-site generation and energy diversification. A microgrid provides access to multiple energy sources – the utility grid, energy storage, generators, solar or wind energy – and allow data centers can operate independently of the grid during outages and times of crisis.
- Data center microgrids can also integrate renewable energy sources, which can reduce carbon emissions.

Carbon footprint of a data center

- refers to the **amount of greenhouse gas emissions**, primarily carbon dioxide (CO_2), that is produced as a result of its operations.
- Data centers consume a significant amount of energy to power and cool their servers and other equipment, which results in a large amount of carbon emissions.
- Strategies for reducing the carbon footprint of data centers
  - increasing energy efficiency through the use of advanced cooling systems and energy-efficient hardware
  - deploying renewable energy sources
  - implementing virtualization and cloud computing technologies to reduce the overall number of physical servers required.

Malaysia courts new data centers in bid to be next Southeast Asia hub *[2023-10-14]*

TECHNOLOGY

Malaysia courts new data centers in bid to be next Southeast Asia hub

Cheap prices, tax breaks aid push to replicate neighboring Singapore's success

https://tinyurl.com/59axb6wp

Sedenak Tech Park, Biggest Data Centre in Malaysia *[March 2024]*

- https://www.youtube.com/watch?v=n2b0ljMXyEg

NTT Global Data Centers – Cyberjaya 5

- https://www.youtube.com/watch?v=sRHxxq07UQw

Data Centres: Malaysia's Next Frontier *[2025-01-01]*

- **https://tinyurl.com/yc23mvak**