

Modern Data Center Network Architecture

Spine-and-Leaf Topologies, Virtualisation Technologies,
and Socioeconomic Implications for Digital Infrastructure



Dr. A. Shaji George



Modern Data Center Network Architecture Spine-and-Leaf Topologies, Virtualization Technologies, and Socioeconomic Implications for Digital Infrastructure

Dr.A.Shaji George

Independent Researcher, Chennai, Tamil Nadu, India.

Abstract – In today's digital age, data center networks are the backbone of the digital infrastructure that facilitates the storage, processing, and transfer of data supporting global commerce, scientific research, healthcare systems, and everyday communication. This article takes a look at how things have changed since the old days of hardware centric traditional architectures to software-defined, virtualized networking systems using spine-and-leaf architectures. The analysis is based on published technical literature, industry reports and verified operational data to describe the role and importance of modern data center networks and the economic and social value they offer. The investigation shows that today's methods with Ethernet Virtual Private Network and Virtual Extensible LAN overlays and intent-based networking automation overcome longstanding issues of scalability, latency, human error and vendor dependency. The business model for such networks is one of capital investment in the physical network and operational savings from automation and minimisation of downtime. The advantages of societal benefits range from enhanced capabilities of digital services for citizens to the national level benefits on economic competitiveness and infrastructure resilience. The article outlines areas that require further research, such as optimizing energy use, workforce transition plans, and access to modern network infrastructure in a global context. Knowing about data center networking is essential for policy makers, technical professionals and informed citizens to engage in decisions that impact digital society.

Keywords: Data center networking, Spine-and-leaf architecture, Network virtualization, Intent-based networking, EVPN-VXLAN, Distributed processing units, Network automation, Digital infrastructure economics.

1. INTRODUCTION

Data centers are the backbone of the digital world, handling all sorts of data, from emails to video streams, financial transactions, and even medical records. These facilities are where the servers, storage systems and network equipment that makes digital life possible are housed. But outside the technical world, the networks that join these components are largely unknown, and have an impact on everything from the time it takes to search the web to the effectiveness of emergency services. A data centre network is a collection of switches, routers and firewalls, along with cabling, that connects servers to one another and to the external network. The design of this network will affect the speed at which data moves, the reliability of service, and the ability to easily expand the system. These systems have been designed for 30 years by network architects who mostly design networks using proprietary hardware in a hierarchical manner. A user's request could go through access switches, aggregation switches, core routers and back down again, introducing delays and creating single points of failure.

This proved to be problematic as data volumes increased at a geometric rate. The growing need for video streaming, cloud computing, artificial intelligence training and the ever-expanding number of devices to

connect required a new architecture that could not be achieved with ever-expensive hardware upgrades using the traditional architecture. In a hierarchical network, when a large switch failed, sections of the network were blacked out. Scaling involved getting larger and more complicated equipment that was more expensive and was prone to catastrophic failure. Starting in about 2010, network architects started to use other principles. Instead of using big, costly devices and a rigid hierarchy, they constructed networks using numerous smaller, identical switches in mesh configurations. Functions formerly performed by dedicated hardware were replaced by software. Virtualization enabled multiple virtual networks to utilize the same physical network. Manual configuration was eliminated and deployment was accelerated through automation, minimizing the risk of errors.

These changes are significant not just within the confines of data center buildings. Data center networks form the backbone of financial institutions, telemedicine, competitive businesses relying on cloud services, and government agencies providing digital services. Investing in modern networking infrastructure offers benefits to countries in attracting technology companies, and innovation within their own borders. People gain from quicker and more reliable access to digital services, integrated into everyday life.

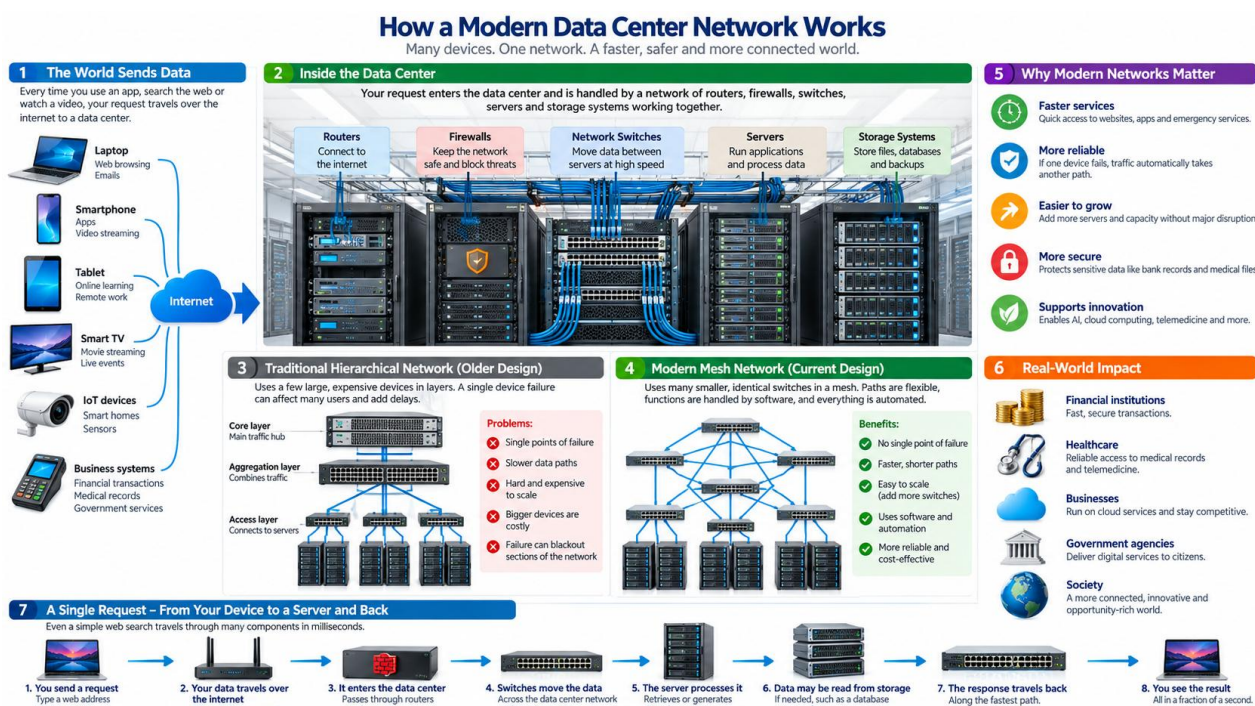


Fig -1: How a Modern Data Center Network Works

This article will dive deep into modern data center networking for those readers who don't have a network engineering background but still need to understand this infrastructure. It demystifies the technical concepts in easy to understand terms, examines the economic models that support these systems, and explores the consequences for societies, organizations and individuals. The aim is to provide readers with enough information to be able to engage in the discussion and decision making regarding digital infrastructure.

2. OBJECTIVE



The purpose of this article is to describe to the educated general public the technical, economic, and social aspects of the modern data center networking. It is intended to give enough technical detail for readers to understand, but not so much that they have to have networking expertise. The article aims to identify the gaps in the current knowledge, practice and policy of this infrastructure.

3. METHODOLOGY

The information for this article was obtained from a systematic literature search of published technical literature, industry standards documents and verified operational reports. Specifications from standards organizations, such as the Internet Engineering Task Force and the Institute of Electrical and Electronics Engineers, that specify protocols such as Ethernet Virtual Private Network and Virtual Extensible LAN were used as primary sources. Technical detail was obtained from industry white papers from several vendors and no specific vendor was endorsed. The literature review on network architecture, automation and economics was carried out using the well-known databases. Data on market activity and characteristics of operations have been checked against several independent sources and only reported if there was consistency between sources. If there is any doubt as to individual metrics, this is clearly expressed. No statistics, quotes or facts were invented for this article.

4. PHYSICAL FOUNDATION SWITCHES, ROUTERS, AND CABLING

All data center networks start with hardware. Switches are the most abundant, and function as traffic directors, receiving data packets on one port and sending them on to another port, depending on the addressing information. There could be hundreds or thousands of switches in a single data center. Routers are used to link various networks together and to determine the optimum path for long-haul traffic. Firewalls analyze traffic and apply security policies to prevent unauthorized traffic and allow authorized traffic.

Data is transmitted as electrical signals in copper or as light pulses in fiber optic glass along the cables that connect these devices. Copper cables usually twisted pair, Category 6 or 6a are suitable for up to 100 meters and are still cheaper than fiber. Fiber optic connections extend much further single-mode fiber can be used to carry signals tens of kilometers and provide more bandwidth and are immune to electromagnetic interference. Modern data centers employ both, with the former copper usually connecting servers to local switches, and the latter fiber connecting switches to one another throughout the data center.

How this equipment is set up is of critical importance. The traditional design was to have servers plugged into access-layer switches that were plugged into aggregation-layer switches that were plugged into core routers. This three tiered hierarchy was sufficient when there was a lot of traffic between servers and outside users. In an office building, a user would ask for data from a server and it would go out through the tiers to get to the user.

Traffic patterns changed. With the advent of cloud computing, servers were required to communicate more with one another than with outside users. Machine learning training tasks may include dozens of servers communicating with each other using huge amounts of data. One of the previous requests had to make one trip out and back, but the new requests made many trips between servers and each trip through the hierarchical layers added delay. The architecture that was good for north-south traffic (users to servers) was not good for east-west traffic (servers to servers).

The Physical Foundation: Switches, Routers, and Cabling

Real Devices. Real Connections. Real Data.

From servers to the internet, hardware devices and cables move data, keep it safe, and connect many networks.

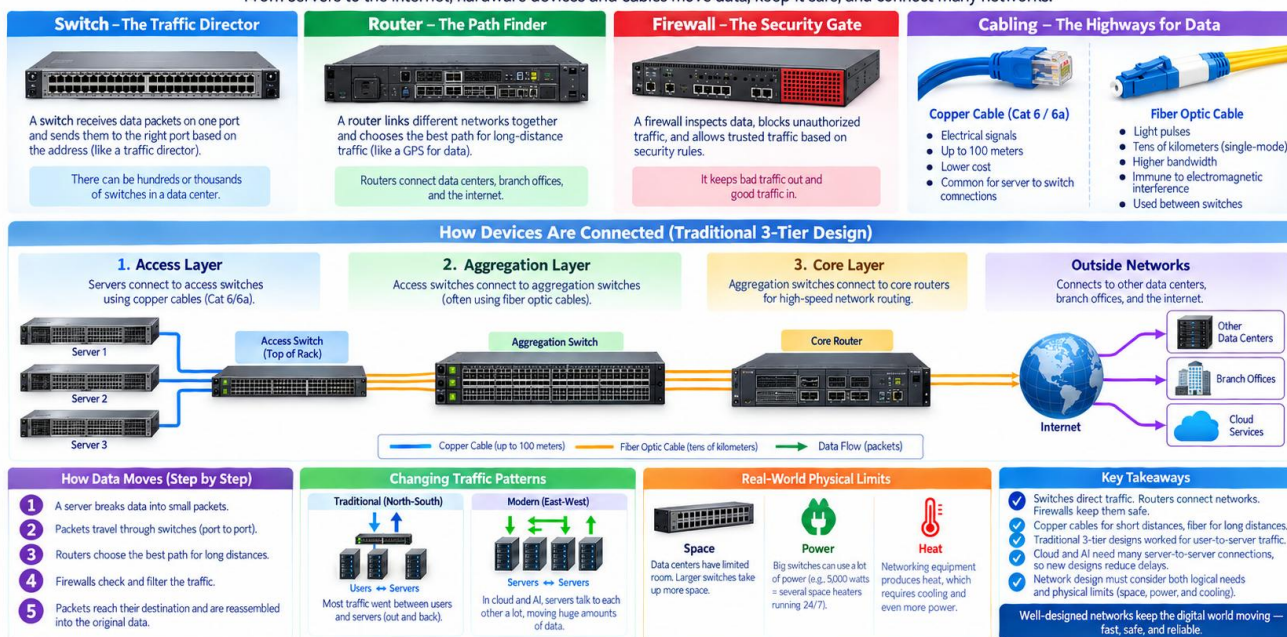


Fig -2: The Physical Foundation Switches Routers and Cabling

These architectural constraints are exacerbated by physical constraints. Data center buildings are limited in terms of space and power capacity. Larger switches to support more traffic use more of both. Networking equipment also produces heat which needs cooling, thus increasing power consumption. One large switch that uses 5,000 watts produces the heat of several space heaters operating around the clock. These facts make it necessary for architects to take into account not only logical network design, but physical engineering constraints as well.

5. THE SPINE-AND-LEAF TOPOLOGY FLATTENING THE NETWORK

The spine-and-leaf design, also known as a “Clos” topology after Charles Clos, the engineer who described similar designs in 1953, is used to overcome some of the drawbacks of hierarchical designs. This configuration is a servers-to-leaf switches arrangement with each leaf switch connected to each spine switch. No switch is connected to another switch of the same type, that is, leaf switches don't connect directly to other leaf switches, and spine switches don't connect directly to other spine switches.

With this design, any server can get to any other server in at most three hops one leaf switch, one spine switch, one leaf switch. This is in contrast to a traditional three-tiered hierarchy where traffic between servers in different areas may go through five or more devices. The fewer hops the less latency, which is important for latency-sensitive applications.

Spine-and-leaf designs provide significant redundancy, increasing reliability. In the event of a spine switch failure, each of the remaining spine switches has connectivity with all of the leaf switches. Traffic is rerouted to

alternate routes. The "blast radius" of any one failure is reduced as no one device is a single point of failure for a significant portion of the network.

The Spine-and-Leaf Topology: Flattening the Network

A simpler, faster, more reliable way to connect servers in a modern data center.

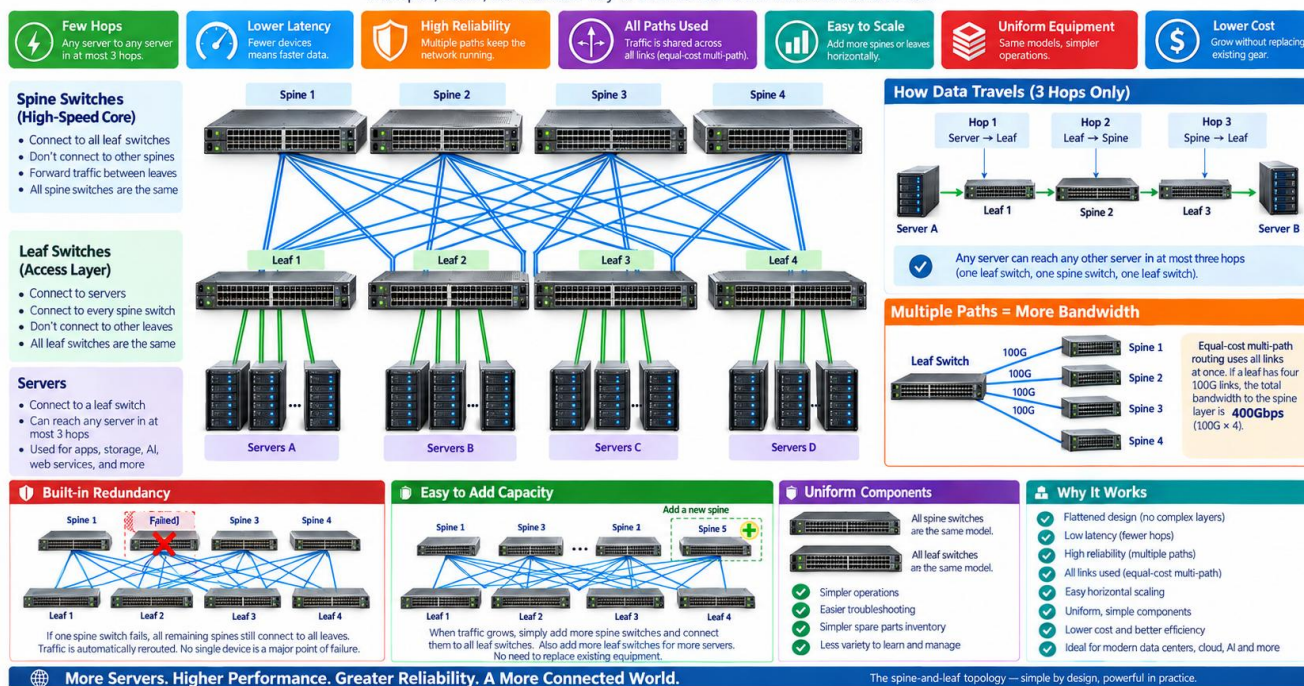


Fig –3: The Spine-and-Leaf Topology Flattening the Network

Equal-cost multi-path routing takes advantage of the multiple parallel paths between any two points. The network does not make one path the primary and others the backups, but rather all paths are used at once, with traffic spread across them. This is done to boost the total bandwidth between leaf switches to the sum of all the links connecting them, not the capacity of any one of the links. Assume that the leaf switch has four 100G links to four spine switches, then the effective bandwidth to the spine layer is 400Gbps.

It's easy to add capacity. As traffic increases, when the spine switches cannot handle the traffic, the architects add more spine switches and connect all the leaf switches to the new spine switches. As more servers are required, architects add leaf switches and then add connections between all the leaf switches and all the spine switches. This horizontal scaling eliminates the need to replace existing equipment with larger and more costly equipment a major economic gain. Spine-and-leaf networks have the benefit of uniformity of components, which makes operations simpler. All leaf switches are equivalent and can be configured and have the same hardware. All spine switches also have the same specifications. Technicians must be knowledgeable with less variety of devices. Spare parts inventory management is easier. Troubleshooting is done in a similar manner irrespective of the particular devices involved.

6. VIRTUALIZATION AND OVERLAY NETWORKS

While physical network topology gives data center networks a foundation, there are layers of abstraction in modern data center networks that are provided by virtualization. This separation enables multiple virtual

networks to coexist on the same physical network without being able to affect one another. Different networks for various departments, customers or applications can be accommodated by a single set of switches without any physical rewiring.

Virtualization and Overlay Networks

Many Virtual Networks. One Physical Network. No Rewiring Needed.

Virtualization lets multiple isolated networks share the same physical switches using overlay technology (VXLAN + EVPN).

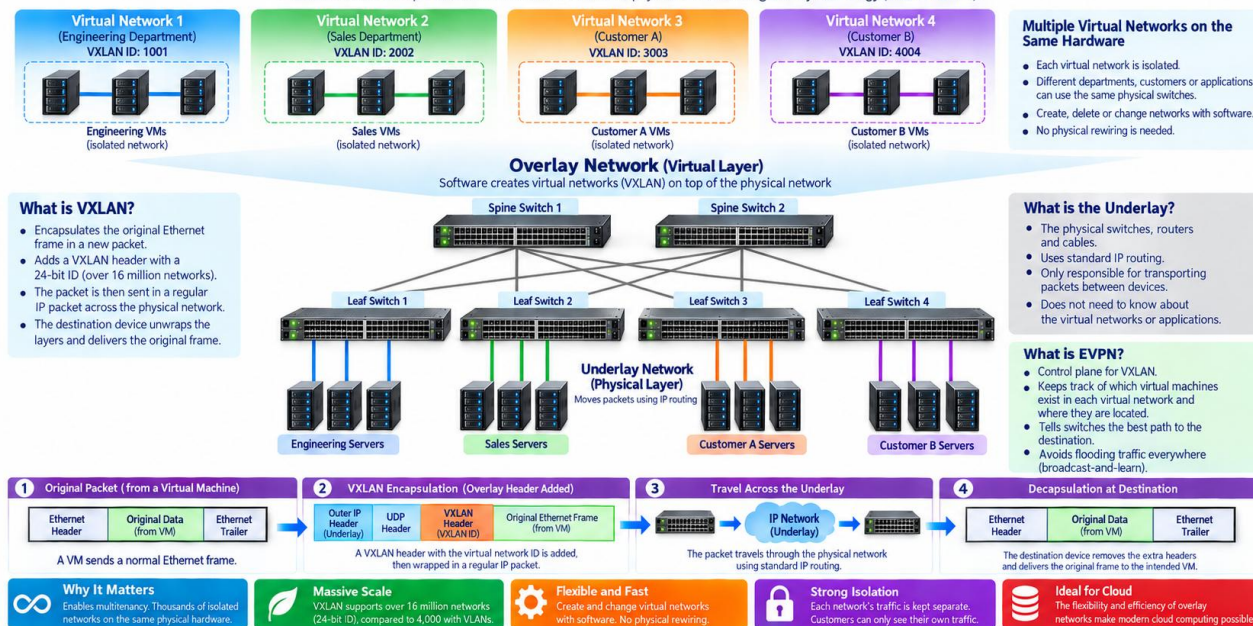


Fig -4: Virtualization and Overlay Networks

This virtualization is modeled with the overlay-underlay model. The underlay is the physical switches, routers and cables that are connected to provide basic connectivity via standard Internet Protocol (IP) routing. This layer is only responsible for transporting packets between physical devices with respect to their addresses. The underlay does not have to be aware of the applications using it, or the organizational boundaries between the applications.

This physical infrastructure is overlaid with the software that forms virtual networks. The original data packets are encapsulated in new packets that are sent across the underlay, which is known as Virtual Extensible LAN (VXLAN). An Ethernet frame for a virtual machine in one virtual network is encapsulated with a VXLAN header containing information about the virtual network and then encapsulated in a regular IP packet and transmitted over the physical network. The destination device unwraps the layers, and passes the original frame on to the intended recipient.

This encapsulation addresses a number of issues. Traditional virtual local area networks (VLANs) were only capable of identifying up to 4,000 isolated networks, which is not enough for large cloud providers serving thousands of customers, each requiring an isolated network segment. VXLAN supports 24-bit identifier, which enables more than 16 million virtual networks to run on the same physical infrastructure.

VXLAN relies on Ethernet Virtual Private Network for the control plane to support its practicality at scale. EVPN provides information on the endpoints available in each virtual network, and their physical location. If a server from one rack needs to communicate with a server on another rack, EVPN will make sure that the sending

switch knows where to find the destination without sending the traffic everywhere to find it. The distributed learning is an alternative to the inefficient broadcast-and-learn model of traditional Ethernet switching. These technologies make multitenancy possible. A cloud provider can provide isolated virtual networks to hundreds of customers, all on the same physical switches. The VXLAN identifiers keep each customer's traffic isolated, so that each set of servers can only see their own traffic. Creating and changing virtual networks involve software changes only no physical re-cabling is necessary. This flexibility is the basis of the economics of cloud computing.

7. INTENT-BASED NETWORKING AND AUTOMATION

Network device configuration used to be a manual process that involved logging into each switch and typing in commands. A network engineer would take these business rules e.g., the finance department shouldn't be able to access the engineering servers and turn them into technical configuration, and then implement the configuration on all the applicable devices. This was a very time consuming process in a network of hundreds of switches, and prone to many errors. If a command is mistyped on one switch, it could result in outages for thousands of users.

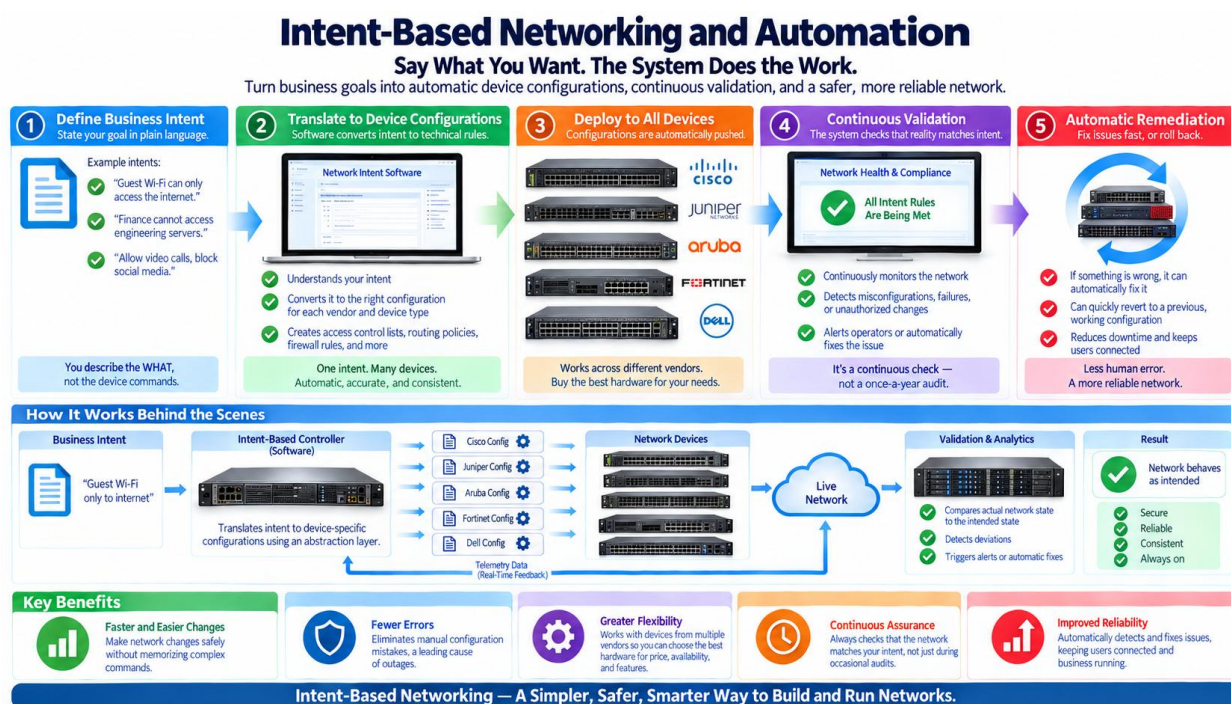


Fig -5: Intent-Based Networking and Automation

Intent-based networking turns that on its head. Network operators state their desires as business goals, instead of device level instructions. The software converts these high-level intentions to configurations for each device type and vendor. Statements such as traffic from the guest wireless network should be allowed only to the internet, not to internal servers are automatically translated into access control lists, routing policies and firewall rules that are pushed to all appropriate devices.



Validation is the round trip. Once configurations are deployed, intent-based systems constantly check the network to ensure that the network behavior is in line with the intent. Should there be any variation such as misconfiguration, failure of a device or unauthorized modification, the system will identify the deviation and notify operators or automatically take corrective action. This is a continuous assurance that is different from point-in-time audits which may identify issues weeks or months after.

The advantages multiply over the lifespan of the network. Traditional methods involved different skills and tools for the initial deployment, changes and troubleshooting. All phases were carried out manually, which is prone to new errors. Intent-based systems keep the model consistent across the board, minimizing the knowledge required to run the network and enabling staff to safely make changes to the network without having to be an expert in all the vendor's command syntax.

The benefit of vendor independence comes into play. With intent-based software providing configuration generation for devices from various manufacturers, organizations can purchase hardware according to price, availability and features, instead of compatibility with current management software. Each vendor has its own implementation, which is translated by the abstraction layer to the organization's needs. This flexibility gives more leverage when it comes to negotiating with suppliers. The decrease in human error equals an increase in reliability. Manual misconfiguration is a top reason for network outages, according to studies of outages. A class of failures is eliminated by removing the direct human interaction with device configurations. If the changes do cause issues, automated systems can revert to previous configurations much quicker than any human operator could diagnose and manually correct the issue.

8. SECURITY ARCHITECTURE IN MODERN NETWORKS

The zero-trust security model presumes that none of the traffic should be allowed by default, no matter where it comes from. In the past, security was based on the perimeter and everything in the internal network was considered safe. This proved unsuccessful many times when attackers who managed to get in, were able to move laterally through the network with little resistance.

Microsegmentation enables security policies to be enforced at a micro level, between individual workloads, not just between network zones. If the security policy does not explicitly allow the traffic between the two servers, they may not be able to communicate directly with each other. By default, unauthorized transfers from a compromised server to neighbouring servers are blocked.

Microsegmentation is now possible without the choke points of centralized firewalls, thanks to the distributed processing units that are now available in both servers and network switches. These specialized processors filter and analyze traffic at thousands of locations across the network, instead of all traffic going through a handful of big security appliances. The inspection takes place near the workloads, thus minimizing latency and maximizing total inspection capacity available.

Policy enforcement becomes more and more workload-centric. Security controls can be placed around a database server, even if it is located in a different physical location, that stores sensitive customer data. Security infrastructure can be moved with the workload to a different rack or a different building without the need for reconfiguration. This mobility is akin to the one used by today's applications, which dynamically create and destroy themselves according to the demand.

Data in transit is protected by encryption between two points in the network. Historically, traffic between web servers in the data center was not encrypted because of performance considerations, although traffic

between end users and web servers was encrypted via HTTPS. Encryption at line rate also does not introduce any noticeable latency, so organizations can encrypt internal traffic and prevent it from being eavesdropped on by others with physical or logical access to the network.

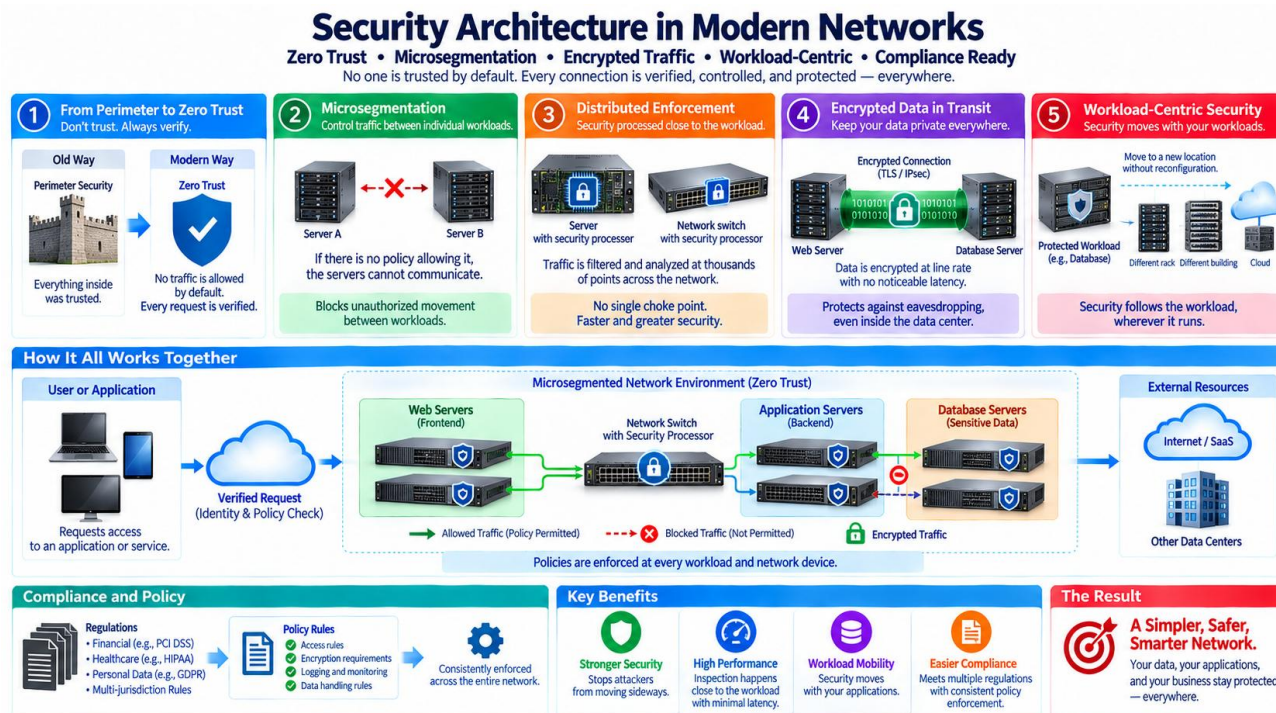


Fig -6: Security Architecture in Modern Networks

These security measures are becoming more and more required in compliance. Financial data, healthcare records, and personal information have regulations that outline controls for network traffic. Organizations with multiple jurisdictions need to meet multiple requirements. Compliance requirements can be expressed as rules that are consistently enforced by the network in the policy-based approach of modern security architecture.

9. DISTRIBUTED PROCESSING UNITS AND HARDWARE EVOLUTION

For decades, the CPU central processing unit has been the star of computing, processing all kinds of different calculations that applications need. Graphics processing units, which were initially developed to process graphics, turned out to be very appropriate for parallel processing required for machine learning and scientific computing. The new category of data processing units are chips specifically built to handle infrastructure workloads that would otherwise be using resources required by applications. A data processing unit is a unit that processes data and helps to offload networking, security, and storage from the main server processors. Dedicated silicon for parsing packets, encrypting data, load balancing, and collecting telemetry data, for example. The main processors are not tied up on overhead infrastructure tasks, but can be used for application workloads.

The economics are simple. The servers processors are expensive and require a lot of electricity. When used for such purposes as network traffic encryption, those costly resources are wasted and could be used for running

revenue-generating applications. A data processing unit that performs those tasks while using less power and less money enhances ROI in server infrastructure.

DISTRIBUTED PROCESSING UNITS AND HARDWARE EVOLUTION

Different chips. Different jobs. A faster, safer, more efficient infrastructure for everyone.

Offload infrastructure work, free the main processors, and do more with less power and cost.

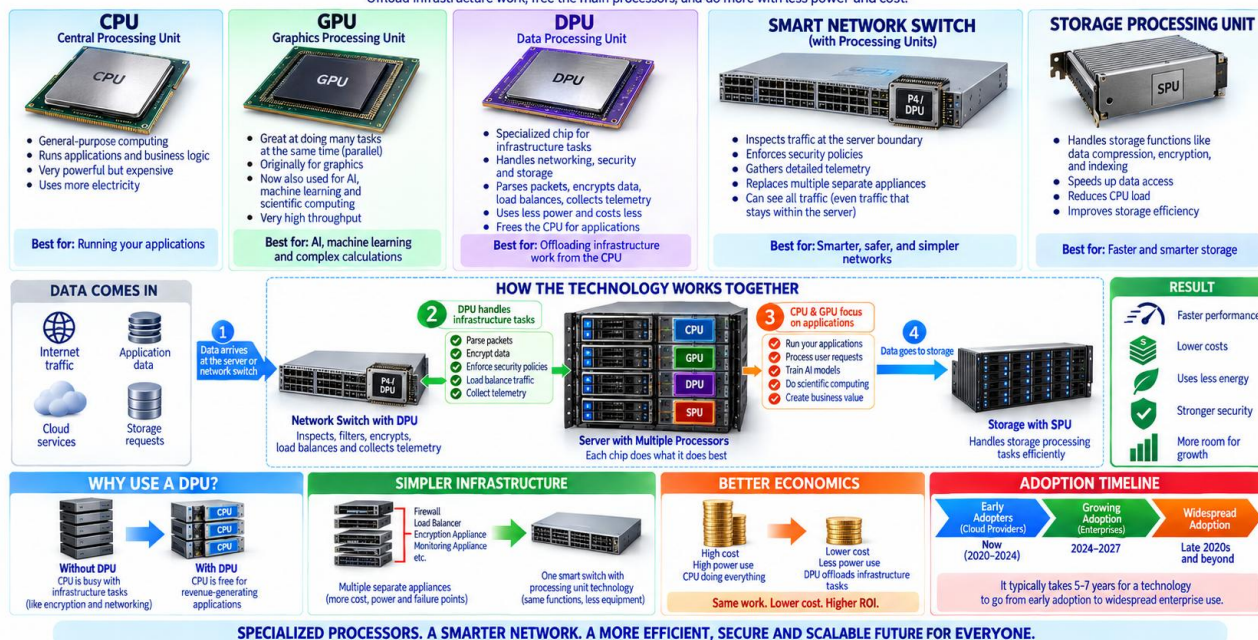


Fig -7: Distributed Processing Units and Hardware Evolution

Today, network switches have similar features. Switches with processing unit technology can be used to inspect traffic, enforce security policies, and gather comprehensive telemetry data, rather than just forwarding traffic based on its destination address. These functions used to be handled by separate appliances, which added up to extra expense, extra power use and extra failure rates. These capabilities are located at the network server boundary and alter the architectural possibilities. Security inspection at each server connection point provides security enforcement that can't be circumvented by traffic remaining within a single physical server. The same points can be used for telemetry, which can show traffic that never touches the central monitoring infrastructure. Capabilities of the architecture that would be impractical or impossible in the centralized models.

It's being deployed in both cloud providers and enterprises. Some of the biggest cloud operators were early adopters because the savings from small efficiency improvements can add up to significant savings at the scale of their operations. Enterprise adoption follows as equipment costs drop and vendors make the technology a part of their products. It takes about 5-7 years for a technology to go from being used by early adopters to being widely adopted by enterprises, so we can expect to see it being widely adopted by the late 2020s.

10. THE ECONOMICS OF DATA CENTER NETWORK INFRASTRUCTURE

The capital investment for a Data Center network depends on the size of the network, but it exhibits certain trends. These costs include switches, routers, firewalls and cabling. The price for a leaf switch for a modern

spine-and-leaf deployment could be in the 15,000 to 50,000 dollar range, depending on the number of ports and speed. A network with 500 servers may need 15 leaf switches and four spine switches, at a cost of several hundred thousand dollars without installation, cabling and optics.

Opex is an ongoing cost of the network. The cost of power varies with the local price of electricity, but is usually 10–20% of the equipment cost per year. Cooling costs are additional costs as a function of heat generation. Staffing is the biggest operating expense for most organisations, with network engineers earning salaries commensurate with the expertise they have. Equipment vendors' maintenance contracts are usually 10–15% of the purchase price per year.

THE ECONOMICS OF DATA CENTER NETWORK INFRASTRUCTURE

Invest today. Operate efficiently. Stay reliable. Grow for tomorrow.
Understanding the costs, savings, and value behind a modern data center network.

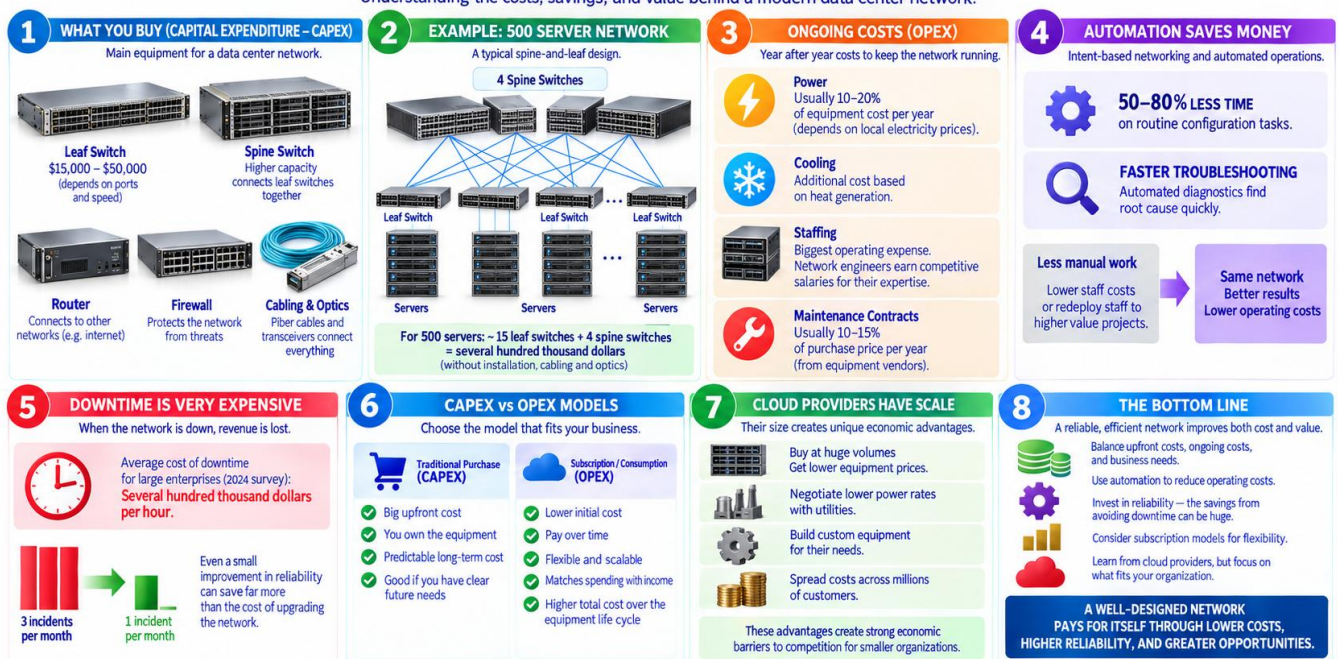


Fig –8: The Economics of Data Center Network Infrastructure

Automation saves a significant amount on the operating costs. Those that have reported on their intent-based networking deployments say they are saving 50 to 80 percent of time on routine configuration tasks. Troubleshooting time is reduced in a similar manner as automated diagnostics can pinpoint root cause more quickly than manual troubleshooting. These time savings equate to less staff or redeployment of staff to higher value projects.

For organizations where network availability is directly related to revenue, the cost of downtime outweighs equipment and operational costs. According to a 2024 industry survey, the average cost of downtime for large enterprises is several hundred thousand dollars an hour. Even if reliability is only improved a little, from three incidents per month to one incident per month, the savings over the life of the network far outweigh the cost of upgrading the network infrastructure. Once organizations come to an honest assessment of their downtime costs, the financial case for reliability investments is easy to make.

The move towards operational expenditure models alters purchasing decisions. Network infrastructure has become more of a subscription service from vendors, with costs spread out over time, and no big upfront capital expenditure. This model is based on consumption, and is good for companies that are not sure about their future needs or want to match their spending with their income. The compromise is increased life cycle costs of the equipment, but flexibility and lower initial equipment costs.

Cloud providers have other economy altogether. They rely on charging customers more for the computing and networking resources they provide than it costs to provide the resources. Their size allows them to buy things that smaller buyers can't, such as equipment at volume discounts, negotiate lower power rates with utilities, and build custom equipment to suit their needs. These benefits add up, and can provide economic barriers to competition to incumbents.

11. NATIONAL-LEVEL INFRASTRUCTURE IMPLICATIONS

The availability of reliable, high-capacity data center networks is a key factor in drawing businesses that rely on digital infrastructure to countries. Cloud service providers site their facilities according to cost of power, network connectivity, physical security, regulatory environment. Such facilities foster direct jobs and indirectly encourage technology ecosystems. Nations that do not have the necessary infrastructure see these investments go somewhere else. For applications such as financial trading to video conferencing, latency is an important consideration. Even at the speed of light, data takes some 70 milliseconds to make the journey from New York to London via undersea cables. The closer data is processed to users, the better the experience will be for latency-sensitive applications. The location of countries along the main network routes is an advantage others are not so well located and have to overcome this disadvantage through investment in local infrastructure.

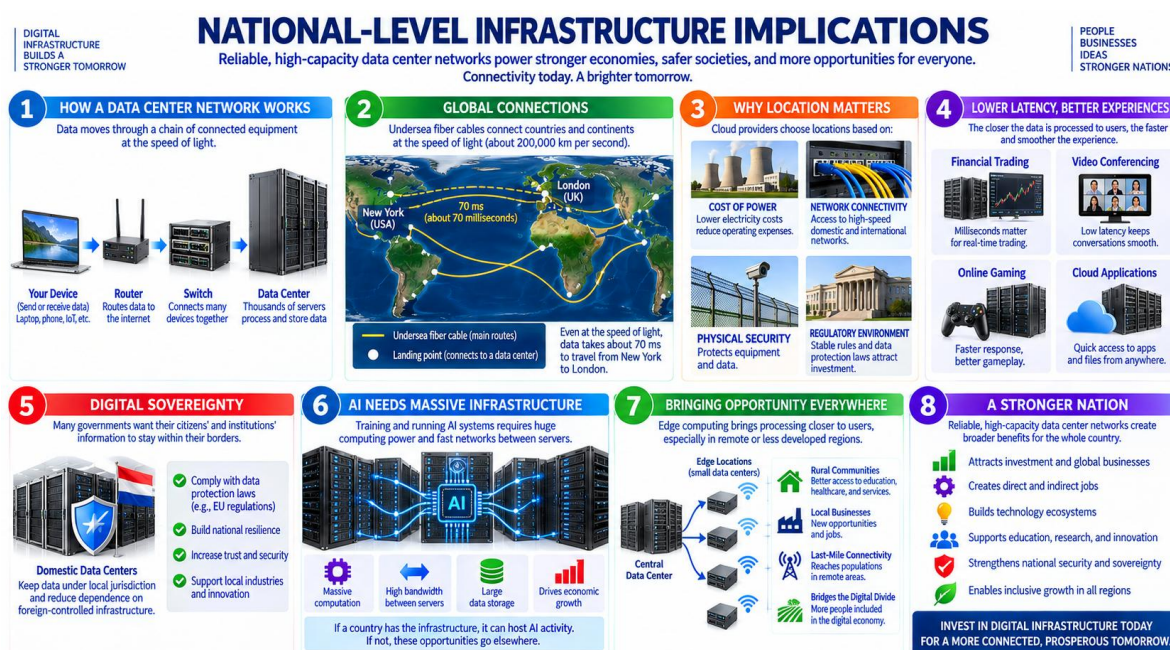


Fig -9: National Level Infrastructure Implications

Digital sovereignty is a driving force behind some national infrastructure initiatives. Governments are increasingly wanting their citizens and institutions information to stay within their borders, under local jurisdiction. This preference is behind the building of domestic data centers and networks, which are not dependent on foreign controlled infrastructure. Such investments have been driven by the European Union's data protection regulations which have speeded up the investments in member states that want to comply. The infrastructure that is needed for training and deployment of artificial intelligence has special economic importance. Large language models, and other AI systems that require massive amounts of computation, have a lot of bandwidth requirements between parallel processing units. If a country has a sufficient data center infrastructure, then it can host this activity, otherwise not. As AI is projected to have a significant impact on the economy, investments in network infrastructure serve as a strategic move to stay ahead in the technological race.

In less developed and remote regions of countries, there are infrastructure deficits that could be mitigated by modern architectures. Edge computing brings processing nearer to the users, lessening reliance on faraway centralized facilities. This, along with better connectivity at the last mile, can help reach populations hitherto cut-off by geographic barriers with digital services. These possibilities can only be realized through policy decisions on infrastructure investments.

12. PRACTICAL APPLICATIONS FOR ORDINARY CITIZENS

The typical user is unaware of the protocol that their streaming video goes through, and doesn't need to be. However, the reliability, speed and cost of digital services is directly tied to decisions made about the networking infrastructure of data centers. Network architecture plays a big role when a video buffers, a payment fails or a telemedicine appointment freezes.



Fig -10: From Your Device to The World



Networked systems are becoming more and more essential for healthcare delivery. EHRs are databases that can be accessed from anywhere with appropriate authorization and contain patient information. Video and exchange of diagnostic images during telemedicine consultation. The medical devices in the hospitals are networked to monitoring devices and to other devices. Patient safety is at stake with failures in any of these systems. Modern network architecture affects healthcare quality with its redundancy and reliability.

The education has undergone a significant transformation towards digital delivery. Course materials, assignments and discussions are provided and accessed via networked applications. The COVID-19 pandemic of 2020–2021 has revealed infrastructure gaps for millions of families, as they have been forced to rely on remote learning. Educational institutions' capacity to provide education to students is partly dependent on the network infrastructure available to the educational institutions and students.

Almost all financial services are networked. Data center infrastructure is critical for bank accounts, investment portfolios, payment processing and fraud detection. The security architecture of networks safeguarding financial institutions plays a role in part whether customers' money and information are safe. The regulatory demands on financial networks are a reflection of the impacts of failure.

Digital connectivity is becoming more essential to access to employment. Network infrastructure is essential for job applications, remote working and freelance platforms. Low connectivity in certain locations creates barriers to access to employment opportunities for those in well-connected locations. The geographical distribution of data center networks does not match the geographical distribution of population, resulting in unequal access.

There has been a significant shift of government services to the online environment. Tax filing, benefit applications, license renewals and requests for public records are increasingly conducted via web applications that rely on government data center infrastructure. For citizens who cannot get these services, or whose access to them is unreliable, the only way to obtain services is to go through a bureaucratic alternative that requires more time and effort. Digital government infrastructure has a real impact on the citizens' relationship with government.

13. ENVIRONMENTAL CONSIDERATIONS AND ENERGY CONSUMPTION

It is estimated that data centers use about 1–2% of the world's electricity, and this percentage is expected to rise with the growing demand for computing. This consumption includes a significant portion of consumption by network equipment. One large switch could consume 500 to 2000 watts all the time. The data centre with hundreds of such switches consumes significant amounts of electricity, resulting in carbon emissions if electricity is generated from fossil fuels.

Direct power use is exacerbated by cooling needs. Electronic equipment produces heat according to its power consumption. This heat must be removed to keep the operating temperatures, and thus needs extra energy. Traditional computer room air conditioning methods have power usage effectiveness ratios of ~1.5 to 2.0, which indicates that 50 to 100 percent more power is used to cool the equipment than to power the equipment. These ratios can be significantly improved by advanced cooling technologies such as liquid cooling or cooling with outside air economization.

The impact of modern network architectures on energy consumption is complicated. Spine-and-leaf designs require more switches than hierarchical designs, which can result in a higher overall power consumption. But little switches are more efficient than the very large ones that are needed in the traditional designs. This

allows for a gradual scaling up to add capacity as it is required, rather than having overprovisioned infrastructure operating at less than capacity.

Virtualization increases utilization, which decreases the number of equipment required to support the same workloads. If two or more virtual networks are running on the same physical hardware, and the hardware is highly utilized, then less hardware is required than if the networks were running on separate hardware. This consolidation makes it possible to lower capital costs and energy consumption.

ENVIRONMENTAL CONSIDERATIONS AND ENERGY CONSUMPTION

Powering the digital world responsibly for a cleaner, brighter tomorrow

Data centers and networks use electricity to keep you connected. Here's how the energy is used, why it matters, and what can make it greener.

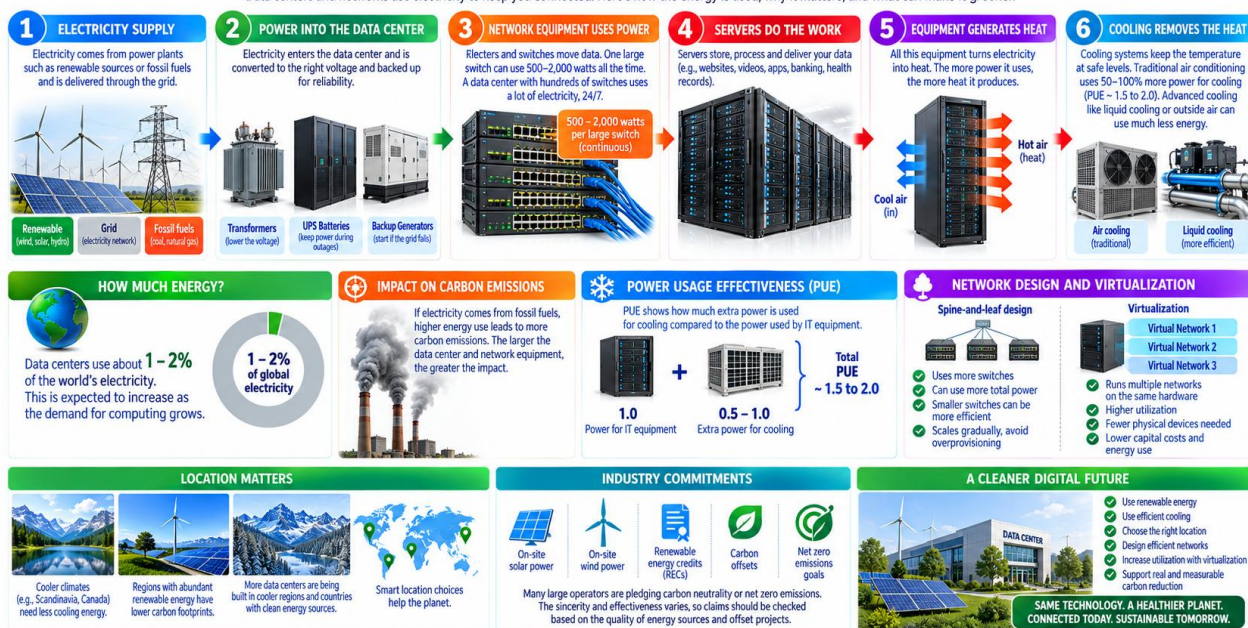


Fig -11: Environmental Considerations And Energy Consumption

Where you choose to locate has an impact on the environment. Data centers in areas with plentiful renewable electrical power have lower carbon footprints than coal or natural gas-powered data centers. Cooler climate areas need less cooling energy. With the environment becoming more of a factor, some of the data center building is moving to cooler climates and countries with renewable power sources, such as Scandinavia, Canada, and other countries.

More and more, large operators are making pledges toward carbon neutrality or achieving net zero emissions. These include renewable energy credits, on-site solar or wind power investments, or carbon offset investments. The sincerity and effectiveness of such pledges varies, some are genuine changes and some are mostly publicity. Specific claims must be evaluated in terms of the quality of energy sourcing and offset quality.

14. WORKFORCE AND SKILLS REQUIREMENTS

The skills needed to operate modern data center networks are different from those required for traditional networks. Knowing how to use specific vendors equipment in the command line is not as important as

knowing network architecture concepts, automation tools and software based management systems. This change is a challenge and opportunity to network practitioners.

For the current workforce, there is pressure for them to update or get obsolete. As more and more technologies are automated, certifications in specific vendors become less valuable. Knowledge of programming concepts, version control and infrastructure-as-code practices grows in value. Those who have learned and become proficient with a specific way to do things over the years need to invest in learning new ways.



Fig -12: Workforce And Skills Requirements

Entry points to the profession change accordingly. With software development experience, new entrants can help with network operations in ways that were never possible before. The necessity to know about distributed systems concepts goes up, the barrier to memorizing arcane command syntax goes down. Educational programs are making curriculum changes to accommodate these changes, but there is still not a perfect match between academic preparation and employer needs.

Even with automation, the need for networking skills is on the rise. Automated systems need to be overseen by humans, designed, managed for exceptions, and continually improved. The job moves from doing the same thing over and over to creating systems, setting policies and dealing with new situations. As automation has progressed, there has been no decrease in employment in networking, and wages for those who have the necessary skills have risen.

The spatial distribution of knowledge results in economic structures. Technology-rich areas attract networking talent and then other talent and investment. Areas without this density find it difficult to develop data center businesses without having to import their expertise at high expense. These imbalances can be overcome over time through investments in education in networking and related areas. The networking profession is still a fairly homogenous field. Women, racial minorities and other underrepresented groups are

underrepresented in network jobs relative to the workforce as a whole. The reasons are similar to those that affect technology employment in general lack of access to education and mentorship, workplace cultures that don't keep diverse employees, and continued hiring bias. These patterns can only be dealt with through deliberate action.

15. RESILIENCE PLANNING AND GEOGRAPHIC DISTRIBUTION

Redundancy within one data center guards against hardware failures, but not against occurrences that impact the entire data center. Even with the best site construction, fires, floods, power grid failures and natural disasters can make an entire site unavailable. Continuous availability is therefore an important consideration for organizations that rely on it, as they need to think about geographic distribution of their infrastructure.

RESILIENCE PLANNING AND GEOGRAPHIC DISTRIBUTION

Keep Your Data and Applications Available — Anywhere, Anytime

Multiple locations. Connected by reliable networks. Designed for any situation.

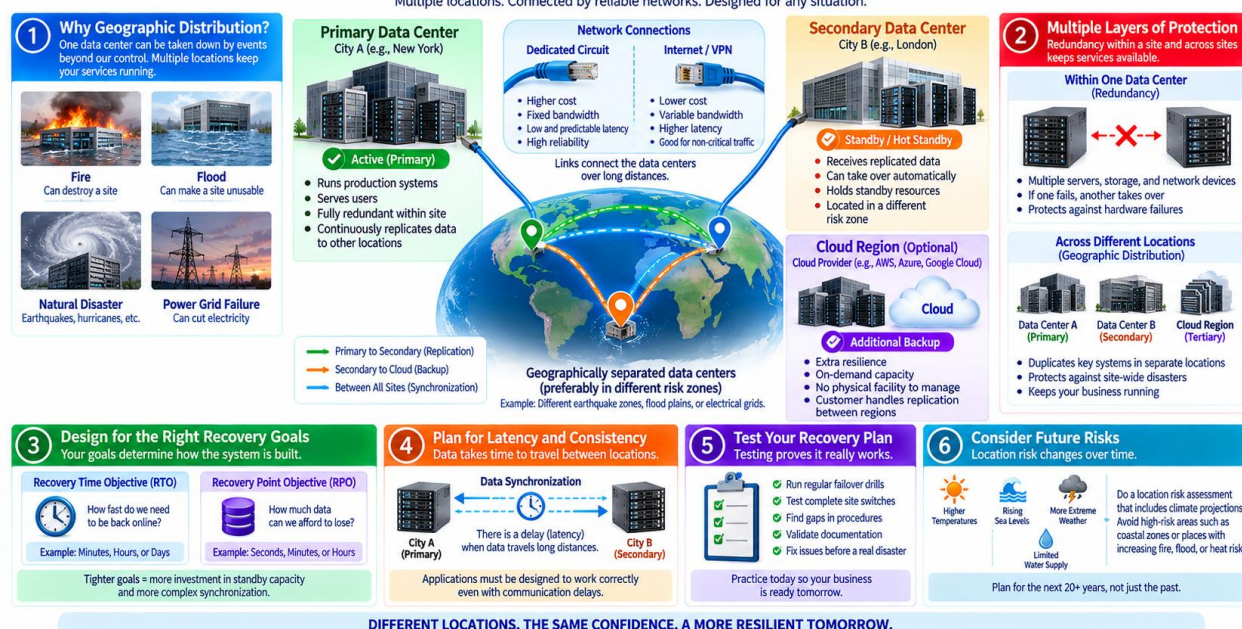


Fig -13: Resilience Planning And Geographic Distribution

Multi-site designs duplicate key systems between data centers in various locations, preferably in separate risk zones. A financial services company could have primary systems in one city and replicated systems hundreds of kilometers away, in a different earthquake fault zone, flood plain or electrical grid area. This latency between sites needs to be taken into account when designing applications so that they can be implemented in a consistent manner while still accommodating for communication delays.

These distributed sites are linked together via a network architecture and it is this architecture that dictates recovery capabilities. Fixed bandwidth and latency, but dedicated circuits are fixed costs, no matter if they are used. Internet connections are cheaper, but do not provide the same assurance of performance and availability. Most organizations will utilize combinations based on the criticality of the workloads.

Recovery time objectives are a statement of the time allowed to recover from failures. Recovery point objectives define the amount of data that can be lost. These metrics are used to make design decisions



regarding how often to replicate, how to automate failover, and how much resources to keep at backup sites. The more tightly the objectives are set, the more investment is needed in standby capacity and more complex the synchronization between sites.

Testing verifies recovery capabilities that they work. Groups that spend a lot of effort building comprehensive disaster recovery systems often find that they don't work out until disaster strikes. Practice exercises including full failovers to backup sites show up gaps in procedures, documentation and technical implementation. The difference between organizations with real resilience and those with only theoretical protection lies in the discipline of operation that they have to do to be effective at testing.

Cloud Services alter the economic and logistical aspects of geographic distribution. Organizations can buy capacity in various regions run by cloud providers without owning or operating any physical facilities. The physical resilience of the cloud provider is provided within a region and the customer is responsible for the replication process between regions. This model is a transfer of responsibility and not a denial of the need for careful consideration of geographic distribution.

Data center location risk assessments change with climate change. Previously safe facilities now are at greater risk for flooding, fire or extreme heat. For organizations that have a 20-year or longer infrastructure planning period, it is important to take into account the impact of climate projections on site selection. Coastal sites, sites with a water supply that is stretched for cooling, and areas with more extreme weather events have higher long-term risk that may not be captured by historic data.

16. RESEARCH GAP

There are key gaps in the current knowledge about data center networking that impact practice and policy. Academic literature on energy optimization is still insufficient. Although there are a number of documented efficiency techniques, there are no detailed frameworks to assess the total environmental impact of the network throughout its lifecycle, including manufacture, operation and disposal. Lack of guidance for organisations making infrastructure decisions on how to minimise environmental damage.

The long term economic consequences of architectural changes are not well known. Spine-and-leaf architectures and intent-based automation are embraced by organizations, but there are few longitudinal studies that track real-life costs and benefits over network lifecycles based on vendors' claims and early adopter reports. The 5-10 year time horizons required for such studies have not yet passed since many of the modern approaches were widely put into use.

There is a need to pay more attention to workforce transition impacts. Workers whose skills are impacted by automation are either displaced or must pay for retraining. There is little research on how to make these transitions equitably, by supporting the transfer of experienced workers instead of replacing them with workers with different skills. Development of best practices for individual organizations and policy approaches for broader labor market adjustment are needed.

Security architectures are under attack at a much faster rate than security research. Zero-trust models and microsegmentation are considered best practice today, and the adversary's capabilities continue to evolve. Their effectiveness in withstanding future threats, such as those that could be introduced by quantum computing or advanced AI-powered attacks, is an area that still needs to be explored.

The issue of access equality in infrastructure is not given enough attention at the global level. The majority of research and practice is undertaken in affluent organisations in developed economies. The issues that are



encountered by developing countries, small organization and under-served communities in acquiring modern networking capabilities vary from the ones that dominate the literature. Research needs to target resource-poor settings in order to fill these gaps. There is a lack of standardization processes that keep up with technological change. Years-old protocols are used to operate current infrastructure, and newer protocols are not yet formally standardized to ensure interoperability and minimize vendor lock-in. An unresolved challenge is to speed up the standards development process without compromising the thorough review process that makes standards reliable.

17. CONCLUSION

Data center networking has come a long way from being a technical specialty of interest only to engineers to infrastructure that impacts virtually every aspect of modern life. The shift from hierarchical architectures to spine-and-leaf, from hardware to software-defined, and from manual to automated operations is the result of learning that has been gained on how to build systems that are reliable, scalable, and economically sustainable. The reason these changes are significant is that the services that people rely on healthcare, education, finance, employment, government are all now dependent on this infrastructure.

These are not just engineering issues the technical details discussed here such as EVPN-VXLAN overlays, intent-based automation, microsegmentation security, and distributed processing units. They decide if digital services will be available in times of crisis, if organisations can safeguard sensitive information, if regions can be a place where technology is invested in, and if environmental effects can be controlled. Technical decisions have social and economic consequences.

Data center networks are fraught with opportunities and tensions arising from economic models. Modern infrastructure is capital and knowledge intensive, placing power in the hands of large institutions and affluent places. Consumption-based and cloud service models may lead to a potential democratisation of access, but also to a dependency on the dominant providers. Competitive positioning is set by national infrastructure investments for decades. Knowing these dynamics is useful to make organizational and policy decisions.

The knowledge gaps highlighted here environmental impact assessment, long-term economic assessment, workforce transition, security evolution, global equity and standardization are real obstacles to making informed decisions. Studies that would fill these gaps would be useful, not just for the sake of research. The risk is worth the money to fill them. This examination puts into perspective the claims made about digital infrastructure, engaging in policy debates, and comprehending the systems that make up our everyday digital experiences for readers who are not networking professionals. While it may not be visible when working well, citizens are better served to know that the infrastructure exists, how it functions, and what options influence its evolution.

REFERENCES

- [1] Abednarz. (2016, November 18). 3 leading reasons for network downtime. Network World. <https://www.networkworld.com/article/957779/3-leading-reasons-for-network-downtime.html>
- [2] Ashley, M. (2026a, March 2). People, process and pitfalls: Tackling human error and skill gaps in the data center network operations. Nokia.com. <https://www.nokia.com/blog/people-process-and-pitfalls-tackling-human-error-and-skill-gaps-in-the-data-center-network-operations/>



-
- [3] Bamberger, A., Shashidhar, A., Kolobov, S., & Gounder, A. A. (2023, March 1). IMET Route Filtering for Ethernet VPN (EVPN) Ethernet Tree (E-Tree) with VXLAN Encapsulation. <https://www.ietf.org/archive/id/draft-bamberger-bess-imet-filter-evpn-etree-vxlan-00.html>
- [4] Bitar, N., Shekhar, R., Uttaro, J., & Henderickx, W. (2018a). A network virtualization overlay solution using ethernet VPN (EVPN). <https://doi.org/10.17487/rfc8365>
- [5] Client challenge. (n.d.). <https://www.scribd.com/document/482632813/Cisco-Data-Center-Spine-And-Leaf-Architecture-Design-Overview-White-Paper>
- [6] Clos network architectures. (n.d.). CS-652. <https://pantelis.github.io/cs652/docs/cloud/lectures/clos/>
- [7] Data centres & networks - IEA. (n.d.). IEA. <https://www.iea.org/energy-system/electricity/data-centres-and-data-transmission-networks>
- [8] George, D. (2026a). Healthcare Monopoly and Patient Welfare: Economic analysis of essential dialysis services in underserved markets. Zenodo (CERN European Organization for Nuclear Research). <https://doi.org/10.5281/zenodo.19165911>
- [9] Deep dive into CLOS architecture: understanding leaf-spine, folded-clos, and fat-tree topologies. (2026, March 13). Asterfusion Data Technologies. <https://cloudswit.ch/blogs/deep-dive-into-clos-architecture-leaf-spine/>
- [10] George, D. (2025b). Data centers and water crisis in India: Why digital infrastructure could drain our wells dry by 2030. Zenodo (CERN European Organization for Nuclear Research). <https://doi.org/10.5281/zenodo.17920097>
- [11] Energy demand from AI – Energy and AI – Analysis - IEA. (n.d.). IEA. <https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai?>
- [12] George, D. (2025a). The Evolution of Data Center Networks: Strategies for Modern Infrastructure design. Zenodo (CERN European Organization for Nuclear Research). <https://doi.org/10.5281/zenodo.15450624>
- [13] Executive summary – Energy and AI – Analysis - IEA. (n.d.). IEA. <https://www.iea.org/reports/energy-and-ai/executive-summary>
- [14] George, D. (2026d). Digital Inclusion vs. Fiscal Revenue Assessing India's Per-Gigabyte Data Tax. Zenodo (CERN European Organization for Nuclear Research). <https://doi.org/10.5281/zenodo.20922074>
- [15] Human error causes outages. (2026a, August 4). Light Reading. <https://www.lightreading.com/cable-technology/human-error-causes-outages>
- [16] George, D., Dr.T.Baskar, & Dr.M.M.Karthikeyan. (2026b). From concrete to compute the Middle East's bid for global AI infrastructure. Zenodo (CERN European Organization for Nuclear Research). <https://doi.org/10.5281/zenodo.20600016>
- [17] IEEE SA. (n.d.). IEEE Standards Association. IEEE Standards Association. <https://standards.ieee.org/ieee/802.1Q/10323/>
- [18] George, D. (2026g). Killer acquisitions and the blind spot of merger control: economic logic, contested evidence, and implications for innovation policy. Zenodo (CERN European Organization for Nuclear Research). <https://doi.org/10.5281/zenodo.22250777>
- [19] Intent-driven Autonomous network and Service Management in future Networks: A Structured Literature review. (n.d.-b). <https://arxiv.org/html/2108.04560v1>
- [20] George, D. (2026f). Color-Coding Conventions in Structured Cabling Administration evaluating the divergence between ANSI/TIA-606 specifications and field practice. Zenodo (CERN European Organization for Nuclear Research). <https://doi.org/10.5281/zenodo.22165533>
- [21] Ivanov, I., & Ivanov, I. (n.d.). Leaf-Spine architecture. NetworkAcademy.IO. <https://www.networkacademy.io/ccna/network-fundamentals/leaf-spine-architecture>
- [22] George, D. (2026e). Generative Artificial Intelligence and Expertise Development among Generation Z: A Critical Review of Cognitive Offloading, Critical Thinking, and AI Literacy. Zenodo (CERN European Organization for Nuclear Research). <https://doi.org/10.5281/zenodo.21971812>
- [23] Krattiger, L. & Cisco. (2023). VXLAN BGP EVPN Multi-Site. In Cisco Live. <https://www.ciscolive.com/c/dam/r/ciscolive/apjc/docs/2023/pdf/BRKDCN-2913.pdf>
- [24] George, D., Dr.T.Baskar, & Dr.M.M.Karthikeyan. (2026a). India's M.A.N.A.V vision: Redefining global AI governance through Human-Centric principles and strategic sovereignty. Zenodo (CERN European Organization for Nuclear Research). <https://doi.org/10.5281/zenodo.18730381>
- [25] Leaf-centric Logical Topology design for OCS-based GPU clusters. (n.d.). <https://arxiv.org/html/2603.28168v2>
- [26] George, D. (2026b). Is India running out of water - An analysis of water bankruptcy, AI data centers, and big finance. Zenodo (CERN European Organization for Nuclear Research). <https://doi.org/10.5281/zenodo.19259949>
-



- [27] Leaf-Spine architecture explained for modern data centers. (n.d.). The Network DNA: Networking, Cloud, and Security Technology Blog. <https://www.thenetworkdna.com/2026/03/leaf-spine-architecture-explained-for.html>
- [28] George, D. (2026c). Data centers: critical infrastructure, global risk & geopolitical power. Zenodo (CERN European Organization for Nuclear Research). <https://doi.org/10.5281/zenodo.20443237>
- [29] Li, X., Lung, C., & Majumdar, S. (2016a). Green spine switch management for datacenter networks. *Journal of Cloud Computing Advances Systems and Applications*, 5(1). <https://doi.org/10.1186/s13677-016-0058-8>
- [30] LLM-Based intent processing and network optimization using Attention-Based Hierarchical Reinforcement learning. (n.d.). <https://arxiv.org/html/2406.06059v2>
- [31] Mushy, M., & Dutta, R. (2017a). Human factors in network Reliability engineering. *Journal of Network and Systems Management*, 26(3), 686–722. <https://doi.org/10.1007/s10922-017-9440-1>
- [32] NetIntent: Leveraging large language models for End-to-End Intent-Based SDN automation. (n.d.). <https://arxiv.org/html/2507.14398>
- [33] Nobles, C. (2022). Investigating Cloud Computing Misconfiguration Errors using the Human Factors Analysis and Classification System. *Scientific Bulletin*, 27(1), 59–66. <https://doi.org/10.2478/bsaft-2022-0007>
- [34] RFC 7209: Requirements for Ethernet VPN (EVPN). (n.d.-a). IETF Datatracker. <https://datatracker.ietf.org/doc/html/rfc7209>
- [35] Ritchie, H. (2026b, July 20). How much energy do data centers and artificial intelligence use? Our World in Data. <https://ourworldindata.org/how-much-energy-do-data-centers-and-artificial-intelligence-use>
- [36] Sharma, D. a. K. (2026). AI-Powered Cybersecurity for OT and the IoT: A review of convergence, risk, and resilience in industrial environments. Zenodo (CERN European Organization for Nuclear Research). <https://doi.org/10.5281/zenodo.20156819>
- [37] Towards End-to-End Network Intent Management with Large Language Models. (n.d.). <https://arxiv.org/html/2504.13589v1>
- [38] Trantzas, K., Brodimas, D., Agko, B., Tziavas, G. C., Tranoris, C., Denazis, S., & Birbas, A. (2025). Intent-driven network automation through sustainable multimodal generative AI. *EURASIP Journal on Wireless Communications and Networking*, 2025(1). <https://doi.org/10.1186/s13638-025-02472-x>
- [39] UpTime announces Annual Outage Analysis Report 2025. (n.d.-a). Uptime Institute. <https://uptimeinstitute.com/about-ui/press-releases/uptime-announces-annual-outage-analysis-report-2025>
- [40] What the data centre and AI boom could mean for the energy sector – Analysis – IEA. (2024, October 18). IEA. <https://www.iea.org/commentaries/what-the-data-centre-and-ai-boom-could-mean-for-the-energy-sector>
- [41] Wikipedia contributors. (2026, August 12). VXLAN. Wikipedia. <https://en.wikipedia.org/wiki/VXLAN>
- [42] Yin, Z., Ma, X., Zheng, J., Zhou, Y., Univ. of Illinois at Urbana-Champaign, Bairavasundaram, L. N., Pasupathy, S., Univ. of California, San Diego, & NetApp, Inc. (2011). An empirical study on configuration errors in commercial and open source systems [Journal-article]. <https://www.sigops.org/s/conferences/sosp/2011/current/2011-Cascais/printable/12-yin.pdf>
- [43] AI (2025a). Energy supply for AI – energy and AI – analysis – IEA. [online] IEA. Available at: <https://www.iea.org/reports/energy-and-ai/energy-supply-for-ai>
- [44] Alhazmi, A.S., Mohamed, S.H., El-Gorashi, T.E.H. and Elmirghani, J.M.H. (2022a). OWC-enabled spine and leaf architecture towards energy efficient data center networks. arXiv. Available at: <https://arxiv.org/abs/2203.10892v1>
- [45] Asisi, B.F., Galdon, M.A.C., Lai, C.-K., Cheng, R.-G. and Chong, E.K.P. (2025). Intent-based network for RAN management with large language models. arXiv. Available at: <https://arxiv.org/abs/2507.14230v2>
- [46] Bitar, N., Shekhar, R., Uttaro, J. and Henderickx, W. (2018a). A network virtualization overlay solution using ethernet VPN (evpn). [online] doi:10.17487/rfc8365.
- [47] Cao, Z. (2018a). A survey on intent-based networking. [online] Wustl.edu. Available at: <https://www.cs.wustl.edu/~jain/cse5700-25/ftp/ibn/index.html>
- [48] Cisco. (2026). Cisco nexus 7000 series switches. [online] Available at: <https://www.cisco.com/site/us/en/products/networking/cloud-networking-switches/nexus-7000-series-switches/index.html>
- [49] Data (2025). Data centres & networks – IEA. [online] IEA. Available at: <https://www.iea.org/energy-system/electricity/data-centres-and-data-transmission-networks?>
- [50] IEA (2025). Energy demand from AI – energy and AI – analysis – IEA. [online] IEA. Available at: <https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai>



- [51] International Energy Agency (2024). Executive summary – electricity 2024 – analysis. [online] IEA. Available at: <https://www.iea.org/reports/electricity-2024/executive-summary>
- [52] Mahalingam, M., Dutt, D., Duda, K., Agarwal, P., Kreeger, L., Sridhar, T., Bursell, M. and Wright, C. (2014). Virtual eXtensible local area network (VXLAN): A framework for overlaying virtualized layer 2 networks over layer 3 networks. [online] IETF. Available at: <https://datatracker.ietf.org/doc/html/rfc7348>
- [53] O'Callaghan, K. and Mariappanadar, S. (2008). Restoring service after an unplanned IT outage. IT Professional, [online] 10(3), pp.40–45. doi:10.1109/mitp.2008.56.
- [54] Rfc-editor.org. (2026a). [online] Available at: <https://www.rfc-editor.org/rfc/rfc7432.html>
- [55] Sajassi, A., Aggarwal, R., Uttaro, J., Bitar, D.N.N., Wim Henderickx and Isaac, A. (n.d.). RFC 7209: Requirements for ethernet VPN (evpn). [online] IETF Datatracker. Available at: <https://www.rfc-editor.org/info/rfc7209/>