Module 3: Compute Power

Video Transcripts

Module 3 Video Segment Break Down

Video 1: Introduction

Welcome to this module on Computing Power, arguably the engine room of advanced technologies such as AI. While data is often held to be like oil supplying energy to the digital economy, data means nothing without the ability to process it. The proverbial power we are able to extract from computers has been rising substantially. And that, in conjunction with the evolution of algorithms, as John Zysmann, a colleague of mine here at UC Berkeley, has remarked, is one of the foundations of the platform economy.

One key source of compute power is semiconductors, which have become a core commodity of the digital age. China, for example, imports more semiconductors than crude oil. Computing power's strategic and economic importance was recognized in the very early days of the microchips and continues to rise through today.

To start, let's examine the ups and downs of the industry. The first large scale production of silicon based semiconductors began in the 1970s, driven mainly by players in Texas and Silicon Valley California. However, the U.S. semiconductor industry was nearing collapse in the mid-1980s. Japan then led the industry after a multi-year effort to become a force in memory chips.

That raised fears that the U.S. could lose out on not only a highly innovative industry in and of itself, but one that produced the components crucial for everything from computers to weapons systems.

Yet by the early 1990s, U.S. chip makers regained the lead. This success can be attributed to Sematech, a consortium of 14 American chip makers with an ambitious goal: To revitalize the U.S. semiconductor industry by finding ways to reduce manufacturing costs and product defects. European actors also joined the competition via the Joint European Submicron Silicon Initiative. The formation of these international partnerships, even among rivaling companies, underlines the strategic importance of securing capabilities to harness the best technologies for their computing ecosystems.

As it stands today, U.S. companies are still at the forefront of this industry. But they are being challenged by disruptions in the supply chain, new actors and innovation from outside the traditional chip industry. China, for example, is determined to become more technologically independent with a move into more value-added functions and higher margin products. Chip manufacturer make about 40% margin, versus computer and other related hardware producers who make only 20%. To achieve this ambitious goal, China is bolstering its technology expertise, and investing 150 billion USD into becoming semiconductor independent. Their goal is to achieve 70% domestic production by 2025.

From an angle of strategic foresight, compute power is a major disruption force with a long term impact on businesses, politics, and society. As these technologies evolve, we need to be prepared for possible future scenarios. We also need to understand how market forces currently shape the global industry around semiconductors. Only then can we anticipate strategic opportunities, and proactively evolve our understanding of how different technologies, market actors, and industries move in relation to one another. Considering the importance of that, let's dig deeper and strengthen our understanding of compute power.

Video 2: What is Compute Power? The Early Days

We all have seen pictures of early electronic computers that relied on vacuum tubes and rewiring to create circuits. The Electronic Numerical Integrator and Computer, or ENIAC, was the first programmable, electronic, general-purpose digital computer. It stood at 2.25 MHz compute power, and could calculate additions in 200 microseconds.

By 1954, IBM’s Naval Ordnance Research Calculator, or NORC, was the fastest vacuum tube-based computer at the time. NORC could do the same work as the ENIAC, but in 15 microseconds, a 13x improvement. If we now look back at the technology’s S-Curve the technology moved through its lifecycle in about 24 years and then quickly retired once the microchip reached market maturity.

Today, there are seven broad categories of microchips, which include memory, logic, micro, analog, optoelectronics, discrete, and sensor chips. The first four are the so-called Integrated Circuits, or ICs: Memory stores data. Logic refers to binary gates controlling the CPUs. Micro refers to the central processing units. Analog Semiconductors interface with the physical world regulating sound, temperature and the like.

The power of a computing chip is defined by a number of components. However, the number of *transistors* it holds is the most important characteristic. Gordon Moore, Intel’s Co-Founder and former CEO, predicted that the amount of transistors would double every two years, which became known as Moore’s Law.

Moore’s Law, which is actually less of a law than it is a forecast, says that through continuous innovation driven by fierce international competition, the number of transistors and the resulting computing power have doubled every two years. However, there is a physical limit to how many transistors can fit onto a square millimeter of silicon, and we’re fast approaching it.

Therefore, this puts a big burden on IT managers, constantly having to keep up with faster, expensive hardware. Luckily, you now have a choice: you can either continuously purchase more of the latest chip hardware technology yourself, or you can purchase capacity by outsourcing it into the cloud.

The cloud outsources two things: data storage and computing power. By outsourcing, organizations can focus on core business processes. They can now transition Capital Expenditure investments in their infrastructure to Operating Expenditures by outsourcing and renting space, while saving on continuous upgrades as technology evolves.

Video 3: Innovation in the Semiconductor Industry

The supply chain ecosystem of semiconductors is getting increasingly complex, consisting and being dominated by only a handful of players internationally. Why? Because while the raw material for the semiconductors is cheap, making them is very expensive. TSMC, the world’s biggest semiconductor manufacturer, expects its new 3nm chip factory to cost around 19.5 billion USD. This would be the most expensive manufacturing plant ever at 5x the cost of Tesla's state-of-the-art factory in Shanghai. As a result, the large investments and extensive experience required for such complex precision processes have consolidated the semiconductor manufacturing space into just a handful of global players.

So if new entrants in the industry are unlikely to appear on the scene, what are the drivers of innovation in it? The first driver is ever larger amounts of data from which we can derive ever more valuable insights. To do so we need more capable systems. The second driver is physics. Even with the latest microchips being 3nm in size, there is a limit to how many transistors fit onto a piece of silicon. So, R&D labs at big multinationals and universities alike are innovating toward new substrates and architectures of chips, finding entirely new ways to harness the laws of physics.

Aside from opportunities for chip designers to push the limits of Moore’s Law further in terms of hardware, we’ve also seen a strong focus on more efficient computing processes. This is referred to as process optimization, both in an organizational sense and in information science sense. For instance, energy-optimized computing, most recently driven by the popularization of low-power edge devices, increasingly puts organizations on the spot in 2 ways. First, it causes them to rethink their approach of collecting data, which often happens without apparent use cases in mind. Second, it pushes us to not “boil the ocean” of available data, but rather implement specialized and optimized data pipelines and data processing procedures.

Video 4: Edge Computing, Neuromorphic and Quantum Computing

A fascinating development that is emerging from the cloud is the birth of edge computing. Data centers host and hoard enormous amounts of data. They require equally tremendous computing power on the real-time data flows: between user devices that generate data and the data center hubs that process and store them. Since end-user devices can be hundreds if not thousands of miles away, this causes critical latency issues, most notably in the healthcare and financial industries. Edge Computing on the other hand, places certain computing capabilities directly in the device, substantially decreasing latency and increasing data security and privacy. Pre-selecting and locally processing data allows us to reduce the transitions of data between devices and the storage amounts they require in the cloud.

This all means that Edge Computing is an important counterforce against the aggregation of all data into a few data centers. When timely analysis is important and latency requirements are tight, data that is processed as it gets recorded on the Edge saves important bandwidth in data centers. It also gives data operators more autonomy over data handling.

However, this is a double-edged sword. To be able to apply machine learning on the edge, for example, one still needs to train the algorithms of data operators on centrally stored data in the cloud. This allows them to supply enough data points for meaningful training exercises, which then results in actionable and accurate predictions on the edge. Furthermore, while minimizing centrally stored data requirements, Edge Computing puts pressure on available data infrastructure to send thousands of real time analysis results from edge to storage, and to interface devices elsewhere on the planet.

This is a new trend, but there are even bigger ones coming at us. Currently, we are bundling together warehouses full of servers to get the most bang for the buck, as well as to evade the boundaries of Moore's Law. Longer term, entirely new ways of thinking about and architecting compute power are required to meet the demands of the growing cognitive and data economy. Two of those technologies are Neuromorphic and Quantum Computing.

When we seek to maximize computing power, there are generally two possibilities: compute smarter, or compute harder. Neuromorphic computing is challenging the current approach of chip architecture by bridging the gap between machine architecture and our brain's architecture. It's based on the assumption that if we're able to recreate the layout of the human brain, we can augment its functioning and processing in turn. Especially in the search for more capable AI, neuromorphic computing proposes that a computer built like the brain will allow software algorithms designed like the brain to achieve new breakthroughs in computing performance.

However, neuromorphic computing is currently still a concept rather than a mature solution. Hence, instead of computing smarter, we're also looking into computing harder. Enter Quantum computing.

What is quantum computing? Quantum computing is based on quantum physics on a subatomic level which rests on two mind-boggling phenomena: quantum superposition and quantum entanglement. In simple terms, quantum superposition allows for a quantum bit, or qubit, to be in more than one state at a time. Traditional bits can be either "0" or "1", but qubits can be "0 and 1" as a simultaneous state in itself.

Quantum entanglement is the characteristic that two particles, or photons, can be in the same state independent from their physical distance from each other. The current record lies at 1200km between a Chinese lab and a satellite. This serves as the foundation for the quantum internet. Imagine the state of a photon in space flipping to a 1, and then observing that another photon in your lab flips to a 1 at the same time. We can easily see the potential of long-distance parallel computing, or remote system management and optimization enabled by this capability.

Given these characteristics, quantum computing operates outside of our current understanding of the binary computing world. This allows us to conduct very complex calculations simultaneously at a much greater speed. It also allows us to tackle computing problems of previously impossible scope. This achievement is called “quantum supremacy” over any other computing technology available, as initially claimed by both Google and IBM in 2019.

Its effects for the consumer market are minimal at the moment. In the near term, quantum’s benefits in enterprise or consumer markets will lie in solving highly specialized computing problems in R&D, medical, and scientific fields. In the medium term, as we engage in experimental research with quantum computing providers to elucidate operational or environmental uncertainties, more and more use cases will crystallize. In the long term, quantum’s implications for information security and research and development will impact across all industries. Wherever we’re facing significant complexity and are exposed to a great number of unknowns, quantum computing can potentially help.

Video 5: Capabilities & Applications: Autonomous Vehicles

Semiconductors can be found in almost all electronic devices – from laptops and TVs to thermostats and fridges – but new chip architectures can open up these traditional industries to new players. The emergence of autonomous cars and the computerization of vehicles, for example, has allowed graphic card producer NVIDIA to capitalize heavily on its technologies’ capabilities in artificial intelligence. Now, they are in the pole position in autonomous driving as NVIDIA’s technology powers all Teslas, for example.

The computerization of cars through specialized chip designs has wide-ranging implications for the car industry in turn. The contribution of electronics to the total cost of a car had increased from 20% in 2007 to 40% in 2017. It’s expected to rise to 50% in 2030. Such an increase in the share of the cost signals a high dependency on electronics and technology companies, and allows cash-rich technology companies to enter the automotive market. Hence, most of the really exciting innovation is expected to come from tech companies, not car manufacturers.

The impact of this development is not to be understated. While true autonomous driving, Level 5 autonomy, will take some time to become a reality, its precursors already have stark implications for the industry. Level 4 autonomy is already here, meaning that vehicles not only support the driver by autonomously holding lines or changing speed. Rather, they are smart enough to override human control, for example with emergency braking. This means that Level 4 already requires a broad array of vision sensors, lidar, and radar, all enabled by integrated circuits - mostly GPUs - to power the advanced driving assistance systems, known as ADAS.

The automotive semiconductor market is projected to increase globally from 40 billion USD in 2018 to 60 billion USD in 2022. Within that, the Asia Pacific region is expected to grow its demand for automotive semiconductors by 41%, in part due to China being the biggest automotive market.

Aside from autonomous driving, the smartification of the car opens up two trends: On one hand, automotive companies grow into service platform operators; on the other hand, these companies can offer unique value propositions for consumers in motion. Let’s name a few new strategic opportunities emerging from the platformization of the car. These include: location based, driving-specific infotainment, ad serving, utilization-based or driver behavior specific insurance, and automatic peer-to-peer based traffic updates.

Video 6: Capabilities & Applications: Quantum Use Cases & Information Security

Quantum computing uses qubits that can be in multiple states at once, allowing them to run calculations in parallel, rather than in sequence as with classical binary computers. In other words: Because they are able to handle multiple calculations at once, they are best positioned to handle optimization and discovery challenges. What does that mean?

In the pharmaceutical industry, for example, researchers can create new drugs through the development of new proteins or the assembly of different pairs of enzymes. In an attempt to cut down on lab time, many possible combinations are run on computers to see if two enzymes can interact, given their complex structure. However, binary computers are ill equipped to handle many variables in a multidimensional, unknown space. Quantum computers are capable of solving these problems in hours rather than years.

In terms of business, a vast array of optimization problems can be solved by quantum computing technologies. On the enterprise level, quantum computers can reorganize and optimize the production floors that involve thousands of working steps, effectively eliminating any operational waste. Consultancies and analytics providers are able to generate insights more quickly using an unfathomable amount of data. The financial industry is able to leverage quantum computing for new types of risk assessments or portfolio management. In addition, car companies are already experimenting with quantum computing to improve traffic management.

In light of the sheer computing power of quantum computers, the current standards of security protocols come into question. This is especially relevant for massively aggregated data in data centers, which is a likely target for cyber criminality.

In a so-called brute force attack, an attacker tries to guess a password out of all the possible options. With current security standards that is at 3.026×10^15 options. This number of options is so high that it makes a brute-force attack not impossible but infeasible to conduct, for the most part, both in terms of available time, and compute power. Quantum computers can crack most of these codes in a matter of hours, thus threatening cybersecurity on a global scale.

In summary, there are many existing challenges that can be solved by applying quantum technologies. Quantum science and computing might open up entirely new possibilities for use cases that span across industries and markets that are yet to be defined.

Video 7: The Oligopoly of the Public Cloud Market

The cloud storage and computing market is an oligopolistic playing field, at least with regard to public clouds. Using public clouds, as opposed to private clouds, means that you as a customer are not responsible for any of the management of a cloud solution. Your data is stored in the provider’s data center, and the provider is responsible for the management and maintenance of the data center. Private clouds, on the other hand, reside on a company's intranet or hosted data center, where all of their data is protected and managed by the company itself or a third party behind a firewall. So, then why is the public cloud space an oligopoly?

The reasons are both the major investments required to build a data center, and the massive economies of scale that can materialize. These drivers have led to the creation of 5 major international players in the public cloud market, including Amazon at 45%, Microsoft at 17.9%, Alibaba at 9.1%, Google at 5.3%, and Tencent at 2.8%.

IBM, a legacy infrastructure as a service provider, had initially watched other cloud providers outrun it in the race for cloud market share. However, IBM acquired the open cloud leader Red Hat in 2019, which brought IBM 6 billion USD in cloud revenues. Furthermore, acquiring Red Had complements IBM's strategic position to build the infrastructure of the future. This will enable organizations to access their specific cloud solutions through a stable intermediary, being IBM. However, Amazon, Microsoft, and others continue to build their cloud platform business so successfully that they are shaping the digital economy through their platform-driven applications. The dependency on this oligopoly is the cause for concern by governments and companies alike, especially since the public cloud market is projected to grow by 150% between2019 and 2022.

The German government for example, introduced the idea of a European cloud alternative, called Gaia-X, funded by the European Union through a Belgian non-profit foundation. Originally conceived as an answer to the almost total dependency of European companies on U.S. and Chinese data storage companies, Gaia-X aims to establish a European cloud network that abides by European data privacy regulations. Up until now, the European market failed to independently put forth such a solution. There are doubts that a non-profit entity and politically-driven project like Gaia-X is exposed to the right market mechanisms necessary to compete against the innovative strength and R&D budgets of its Chinese and US competitors. Thus, as business executives, we should very carefully assess where we place our bets in the public cloud market.

Video 8: Actors: Private Clouds and Data Center Market

Despite the rise of public clouds, there remains a demand for private or company clouds and managed data centers. In 2019, McKinsey discovered that about 65% of workloads will continue to be hosted in private data centers and managed by internal-infrastructure teams over the next several years.

There are a variety of reasons for this, including better total cost of ownership in certain cases, the need to safeguard sensitive intellectual property, the absence of viable public cloud providers in some countries, the skill-set enterprises have been building up around managing legacy systems, and the perceived need for control over security and regulatory issues. These factors, combined with the growing use of edge computing, mean that we should expect to see a hybrid, multi-cloud infrastructure picture for the coming decade. Companies need to be prepared for this reality, and not focus all of their efforts on transitioning to a public cloud. This is especially important as new solutions like quantum or edge computing appear on the horizon and continuously gain importance.

While trailing behind competitors in the public cloud market, Oracle, for example, is focusing on legacy cloud offerings instead. The offerings are centered around "Infrastructure as a Service," for private or hybrid solutions. This provides more customized, specialized, and client-controlled offerings for financial and employee management solutions.

There are also managed services data centers, third-party managed data centers, and colocation data centers. The managed data centers simply host your data in their centers. The third-party managed data centers don't just host data, but lease server equipment and physical infrastructure to their clients. Colocation data centers provide the physical equipment to multiple tenants in one location, who then manage the rented servers, storage, and firewalls themselves. While Amazon hosts its data on wholly owned data centers, Microsoft, Google, Intel, Apple, and Cisco are colocation customers.

The diversity of options clearly shows that companies are well advised to conceive an emerging strategy for cloud and data center access.

Video 9: Semiconductor Supply Chain Actors (Part A)

Let's dive into the landscape of semiconductors, which form the core ingredient to building the hardware of cloud or data center solutions. The oligopolistic tendencies we see in the cloud market can also be found in the supply chain of semiconductors.

The following supply chain of semiconductors generally applies to all kinds of semiconductors, although the market forces and implications differ greatly depending on the specific kind of semiconductor. The 8 sets of activities include: Chip Design, Design Software, Intellectual Property, Fabrication, Equipment, Assembly, Chemicals and Wafers, and Testing. Behind each of those activities are specific actors.

The first activity set is Chip Design. It’s the complex art of designing increasingly smaller yet faster integrated circuits, and is an oligopoly with a strong USA dominance. The list of companies in the U.S. includes Intel, Nvidia, AMD, Micron, Texas Instruments, and Qualcomm. The other major hubs for integrated circuits are South Korea with Samsung, Taiwan with Spreadtrum and China.

However, the proliferation of artificial intelligence and the emergence of the cognitive era have given rise to AI chips. No broad design standard has yet emerged from the over 100 companies currently designing AI chips, including giants like Alibaba, Alphabet, and Amazon, but also Taiwanese Novatek and RealTek. Chinese Huawei’s subsidiary HiSilicon is another noteworthy actor in this space.

Various well-funded start-ups are venturing into Chip Design, leveraging methodologies of machine learning. This is a super interesting area of research, with many potential new approaches to machine learning and artificial intelligence challenges. Scientists will continue to develop both narrow and broad AI models. Correspondingly, opportunities for specialized chip companies building the appropriate Application Specific Integrated Circuits, or ASIC chips, will not run dry.

The second segment of semiconductors’ supply chain is Design Software, or Electronic Design Automation, also called EDA. To create any of those integrated circuits, companies rely on design software. Globally, only three companies exist that create suitable software, all of which sit in the U.S., including Cadence Design Systems, Synopsys, and Mentor. This consolidation was the result of extremely short innovation cycles, which forced companies to acquire other players to remain on top and recoup innovation progress.

The third segment of the semiconductors supply chain is Intellectual Property, the blueprint of the technological architecture that is running our everyday lives. Almost every computer in the world runs on an x86 architecture, which refers to the architecture of the chip. The only three owners of x86 patents are Intel and AMD in the U.S., and VIA Technologies in Taiwan. Changing the chip architecture would mean reubuilding operating systems and the applications of practically all computers.

ARM is the owner of the equivalent of x86 for computers for most things mobile. They effectively own the intellectual property of CPUs on all mobile devices, from smartphones and tablets to sensors, automotive computing, edge and certain AI chips. ARM, as the IP owner, doesn’t manufacture chips but licenses the IP broadly. This allowed more companies to build their own processors on top of it, and enabled a fierce competitive landscape for mobile chip manufacturing. Manufacturers like Qualcomm and Broadcom are among the biggest producers of mobile semiconductors, reaching revenues of about 24.3 and 20.8 billion USD in 2020.

Video 10: Semiconductor Supply Chain Actors (Part B)

The best processors in the world include Qualcomm's Snapdragon series chips and Apple's A series Bionic chips. Both are extremely efficient, and exhibit an extremely low level of power consumption. Apple took the lead after it scrapped its partnership with Qualcomm, possibly redefining the power structures in the semiconductor industry.

Besides manufacturing chips, Qualcomm is also the owner of a variety of foundational mobile technology patents, for example the technology that connects a mobile device to any mobile network. Now, everyone has to license at least that technology from Qualcomm. However, as Qualcomm bundles most of its tech into a flat fee, it creates strong incentives for companies to rely heavily on Qualcomm's other patents since they are paving the way. Also, Qualcomm doesn't license its patents directly to chipmakers, tying mobile phone vendors and manufacturers directly to its technology. Apple's strategic move to design its own chip and have it manufactured by TSMC set it free of Qualcomm's ecosystem, which licensed more than 300 phone vendors in 2019.

When deciding on a chip strategy, the options are slim but all the more important. ARM owns the architecture of most mobile chips, and Qualcomm owns many essential patents for mobile technology as well as manufactures some of the best mobile chips available. An old rival Intel has lost many mobile chip clients. In addition, with Apple's announcement to put mobile chips in MacBooks and the emergence of ARM powered Chromebooks and powerful Tablets, non-mobile chips are losing traction.

In terms of the manufacturing in the supply chain, there are two approaches. On the one hand, we have integrated device manufacturers, in short IDMs, such as Intel or Samsung. They keep the entire supply chain in house, from design to fabrication and assembly. On the other hand, we have Taiwanese TSMC, the world's biggest foundry or chip producer. Apple, Nvidia, and Broadcom are also customers of TSMC.

The foundries themselves require highly specialized equipment, designed by other vendors. Here, Dutch ASML is leading in the super high tech market for foundries fitting circuits on the tiniest wafers smaller than 7nm. ASML has an effective worldwide monopoly for their equipment, similar to the German Zeiss who produces mirrors that focus light on wafers to cut structures into silicon.

The final assembly is either done in-house at a company like Intel or Samsung, or in an outsourced semiconductor assembly and testing company, called OSAT. OSATs are predominantly U.S. based but have seen increasing Chinese market shares.

Silicon is the base for practically all computer chips. While the resource itself is the second most abundant element in the world right after oxygen, it is only economically sourced in a few countries, led by China, distantly followed by Russia and Norway. Shin-Etsu and Sumco in Japan, Siltronic in Europe, GlobalWafers in Taiwan, and SK Siltron in South Korea are the major players that turn silicone into silicon wafers. Besides silicon, other chemicals are provided by Japanese Shin-Etsu and Sumitomo Chemicals. Alternatively, there are much smaller European actors, such as BASF, Linde, and Merck.

It's easy to see that the supply chain of semiconductors is highly complex, and therefore fragile in light of disruptions such as pandemics or trade conflicts. While those supply chains span the globe, the core hubs are still in the US, Japan, South Korea, Taiwan and to a lesser extent, Europe. Yet, China is catching up fast. No region has the entire production stack in its own territory, since companies often specialize on particular process steps or technologies in pursuit of economic efficiency.

So, what does this mean for you as a business leader? If semiconductors are a core component of your products, it's critical to establish alternative sources of supply as a matter of contingency planning. This is especially true in light of global shocks that could disrupt the supply chains. While the current market structure is shaped by oligopolistic patterns, we see the emergence of new actors. For one, new companies are designing specialized AI chips for various use cases in the cognitive era. Also, China's push towards semiconductor independence has opened the door for new joiners in the "traditional" semiconductor market. This is worth observing since those new actors could become partners. They could offer critically needed alternatives to the current oligopoly, or an opportunity to secure a new competitive edge through chips.

Video 11: Actors: The Quantum Players

There are very few players currently capable of developing or operating general purpose quantum computers. Google, Microsoft, IBM, and Alibaba are all well-known platform companies in the digital economy, and D-Wave is the major player researching general purpose quantum computers. A variety of other organizations are rallying to provide specialized quantum computers.

Airbus, while not traditionally in the computing sector, is researching applications for existing specialized quantum computers, to uncover breakthroughs in both the optimization of wing design and the handling of aerospace data. Accenture, in a partnership with Canadian 1QBit, is seeking quantum capabilities to compare and design drugs for neurological conditions based on molecular comparisons.

China's Alibaba is developing quantum computer hardware for its AI, e-commerce, and data centers. AT&T is looking to develop secure quantum communication. As quantum technologies represent a new technological paradigm which is arguably still in its beginning, governments are directly investing as well.

While general purpose quantum computers are said to materialize only in the next 5-10 years, use case specific breakthroughs are becoming more frequent. For example, DARPA is partnering with 6 U.S. universities and ColdQuanta Inc. to research quantum information use cases at the intersection of quantum and classical computing.

Video 12: Colliding Trends: Globalized and Politized Supply Chains

The industry of CPU manufacturers has gained international and political interests, because of the skill, experience and a strong financial muscle required to design, manufacture, test, and assemble the CPUs. They can be clustered into two major business models, each with distinct globalized supply chains. One business model is integrated device manufacturers, or IDM, and the other is fabless manufacturers, which outsource manufacturing, test and assembly.

Integrated device manufacturers, like Samsung or Intel, have followed the competitive pressure by vertically integrating various activities. Fabless computer chip developer AMD doesn't own fabrication plants or major assembly components. Instead, AMD focuses on the development and design of microchips, an immensely complex process, involving high R&D expenditures. Hence, AMD relies on strong and mutually beneficial relationships with foundries such as the world's biggest foundry TSMC in Taiwan or Samsung in South Korea.

However, both supply chain types are global in nature, but with China and the U.S. controlling a critical mass of the value added. Therefore, these global supply chains are of enormous interest for political intervention.

China and the U.S. both aim to foster more independent supply chains to create resilience against political and economic shocks. China is working to establish Chinese-designed microchips to gain independence from the US-owned x86 architecture. By basing much of its research on ARM's mobile architecture, it's threatening the leading positions of AMD and Intel. Subsequently, and as part of the US-China trade war started in 2019, US based electronic design automation was put on the export ban list.

For the strategic bystander, it's important to note that China is not the only one trying to escape the reign of x86. Apple is increasingly moving its offering onto ARM's architecture, developing more computer-like tablets, but more importantly also starting to sell laptops with their own computer chips based on ARM architecture. Organizations operating in the software development industry should already be anticipating this switch. Not only will it more deeply integrate the hardware and software, but the application architecture of mobile and laptop or desktop devices will merge as well.

While the power play in the semiconductor industry is in full swing, the colliding uncertainties around the politicization and fragility of supply chains and their increasing demand is giving way to quantum computing as the new source of compute power. The U.S. Government already projects that quantum computing will likely be even more impactful on the economy, infrastructure, and security than semiconductors. We are seeing early indications of this happening with long-running U.S. Government investments in quantum information science, transforming this scientific field into a nascent pillar of the American research and development enterprise.

Video 13: Colliding Trends: Energy Consumption of Data Centers

By 2030, 20% of the world's electricity usage will result from Information and Communication Technology for two reasons: the electronic voltage needed to run and store information in data centers, and the power needed to cool down the servers to prevent overheating. Memory Integrated Circuits, the heart of cloud data storage, store information in millions of transistors in the form of voltage or no voltage. Then, they need to be frequently topped up to hold that information. Data centers alone accounted for 1% of the world's energy consumption at more than 200 TWh in 2019. This is both a shockingly high and shockingly low number. Internet traffic has increased 12.1 times since 2010, and data center workloads have increased 7.5 times. Thanks to the consolidation of data centers and vast improvements in efficiency, the energy consumption remained almost flat at around 1% of world total energy consumption.

Microsoft stores data centers underwater to leverage the cooling capability of the ocean. Most public cloud data centers can be found in northern Virginia. It's one of the world's biggest data center marketplaces, excluding colocation centers and most of AWS data centers, which are run and operated directly by Amazon. This extraordinary aggregation of data centers in just one U.S. state pushes the energy consumption into focus.

But what does this all mean? The energy demand created by data centers is creating buying power vis-a-vis their energy providers. In May 2019, AWS, Microsoft, Apple, LinkedIn, Salesforce, and other actors in the space jointly pressured Dominion Energy, the North Virginian energy provider, to intensify investment in renewable energies.

However, even with all that buying power, the high electricity cost of data centers remains. As a result, data center provider Iron Mountain ventured into vertically integrating this energy production to better control the cost, and has run its own renewable energy plants since 2007. Given this intense and intertwined relationship with energy providers, the United States Environmental Protection Agency has partnered with the Top 30 Tech and Telecom enterprises to foster green energy initiatives.

While Amazon signed the letter to Dominion Energy, they have already been following their usual strategy of vertically integrating entire value chains and thus venturing into energy. Investing in solar plants that are powered and operated by third parties allows Amazon to create catchpenny initiatives to reduce its carbon footprint, increasing long-term planning benefits as renewable energy's price is stable compared to oil and gas. Meanwhile, it's gaining independence from energy providers which power their empire. Amazon launched its first solar energy facility in Shandong, China, expecting to produce 128,000 megawatt hours per year.

Video 14: Colliding Trends: Energy Consumption of Blockchain

A noteworthy technology at the intersection of data processing and energy is blockchain. While in itself not requiring massive data centers, there are crass energy costs.

Bitcoin, as any blockchain technology, is built on the idea that one can build a transparent, decentralized ledger that can be fool proof. To achieve that, each transaction is basically a hypercomplex mathematical function that gets added to the record of previous functions as a block. Hence the name blockchain, which is essentially a chain of equations or blocks.

Based on this concept, every Bitcoin transaction triggers a request to decentralized computers to solve the next equation, and thus validate the transaction. As the technology becomes more and more common the chain of transactions, meaning the actual memory size of Bitcoin grows bigger as well, signalling the tremendous growth in application. In 2015, Bitcoin's blockchain was 45GB but only 5 years earlier in 2010, it was only 1GB.

To incentivize people on the network to invest in solving those equations, they are offered a fee, called the block reward. Miners refer to the people investing computing power in the network. They are looking to optimize the complexity of the transactions that are bundled into the next 1MB block added to the blockchain. Due to the increasing complexity and the decreasing reward, miners need to continuously upgrade their equipment, including the application specific integrated circuits, to remain profitable. As a result, a single transaction requires as much energy as an average American household consumes in 18 days. The 2019 energy consumption of the Bitcoin network was equal to all of Switzerland, while the carbon emissions of mining were 20Mt CO2, that is, two thirds of the total carbon emission in 2019 of the UK.

Wherever trends collide, we also see innovations appear. One emerging technology, optical computing, promises increased computing power by 100% compared to Integrated Circuits, while cutting down to merely 10% of the energy consumption required. Optical computing harnesses the speed and efficiency of light to conduct computational processes rather than electricity. Lightmatter, a spinout startup from MIT, is one player to develop this technology. Full blown optical computing might still be a little far out, but it offers an outlook on a promising, more energy efficient paradigm. In the short term, it's a strategic imperative for leaders to consider their energy consumption as a rising fixed cost, as one requires more and more of it to power computing.

Video 15: Strategic Considerations: How to Decide Between a Public or a Private Cloud? (Part A)

Processing the growing amounts of available data requires new approaches to handling computing power requirements. Especially with the most recent shift to remote work at scale, the obvious answer is outsourcing data storage and computing power to cloud centers. However, the risk associated with sending confidential and business critical data around the globe to cloud providers is just one of many concerns both business leaders and chief information officers have.

Additionally, the open cloud driven strongly by IBM's Red Hat is breaking into the market. Open cloud is aiming to establish an industry-wide standard in cloud architecture, which will allow more seamless integration of different providers. More so, an open cloud architecture is an important prerequisite for hybrid cloud architectures.

As usual, there is not a one-size-fits-all solution. When it comes to preparing businesses for the challenges, a holistic approach must be taken, involving considerations of staffing, infrastructure readiness, and use case specific approaches. Let's take a look at what this means. No matter your industry, the following points are crucial when deciding on a public or private cloud.

First and foremost, having skilled staff to manage either onsite or outsourced datastreams is critical. Data infrastructure engineers are already in short supply. Research by Vertiv showed that organizations could lose up to 16% of their infrastructure workforce to retirement by 2025, with little new entrants in the highly specialized field. Self-healing devices that are able to self-diagnose and adapt to dynamic external forces pose an opportunity to soften the talent shortage. Outsourced data infrastructures are an equally promising solution, but run the risk of creating business-critical dependencies on the providers. Only 10% of businesses indicated that they would be able to operate without a data center.

As a leader, the decision to invest in in-house or external data infrastructure must be coupled with an actionable plan to create or maintain the critical human resource required, even when global supply decreases. The capability of maintaining and managing one's own infrastructure presents critical indication as to what data management strategy to follow.

Another point to consider is infrastructure readiness. Forbes found that only 29% of organizations report that their current data centers meet their needs. And only 1% of engineers think their organization's data centers are future ready. Hence, the trend goes towards increasingly relying on outside solutions.

The main advantage of private clouds is that they are tailor-made for the organization, serving the processes and internal customers in a highly rewarding way. However, to reap those benefits, one must clearly map out the organizational needs now and the expectations for the next 5 years, in terms of growth, market developments, and demand. If you are able to create reliable forecasts hand-in-hand with your infrastructure team, you can theoretically generate 2 times the improvements in customer service, 1.5 times in reliability and availability, and 4 times capacity deployment when deciding for a private cloud.

However, if you are operating with high degrees of uncertainty that prevents accurate long-term planning, or have high infrastructure complexity in terms of legacy devices, maintaining a private cloud can result in a continuous game of catching up with requirements. In these instances, a public cloud is able to offer superior service and cost efficiency for your organization.

Video 16: Strategic Considerations: How to Decide Between a Public or a Private Cloud? (Part B)

Businesses today are exposed to various kinds of data, including financial data, production data, and customer data. As public, private or hybrid clouds offer different use cases, challenges and price points, organizations are well-advised to consider specific use case solutions, with an outlook on the organization’s 5-year and 10-year strategies. This allows for the organization to evade premature solutions that result in technological or vendor lock-in. Depending on your industry, the following considerations might apply to you.

The first is around bandwidth. Do you offer software solutions to your customers? On a private cloud, are you able to manage usage spikes to prevent performance issues? At the same time, are you using your available bandwidth wisely when traffic is slow? Public clouds usually enable smart bandwidth maximization to meet high usage requirements, and create savings when times are slow.

Next is latency. In manufacturing or the financial services industries, latency is a traditionally major concern. Private clouds on premise offer stark decreases in latency, where public clouds can struggle. Are you able to perform your operations in an outsourced cloud with differing latency? If the answer is no, private clouds are your preferred solutions.

The final consideration is global integration. Do you operate globally with a physically separated workforce? Creating seamless experiences on the private cloud can become difficult when services and data are hosted in different locations. Are your employees able to operate in a high-latency environment? Then, public clouds can help.

Data security is a main concern when deciding on a private or public cloud. The general sentiment is that if it’s in-house, it’s safer than on a public cloud. However, public cloud providers like Amazon and Microsoft have many advantages when it comes to security. Yes, a public cloud is a great target for hacking attempts, but that is exactly why public clouds usually outperform private ones. Public cloud providers attract the best talent internationally and redefine what others should follow in terms of security features. They are tested thoroughly in development everyday through continuous hacking attempts. Also, through economies of scale, they are able to invest in the latest technology.

In contrast, private clouds often run on outdated equipment. Security testing is done on a much smaller scale, and the talent available is costly and sparse. Yet, imagine a bank in Switzerland that enjoys the reputation of being a trusted swiss institution. Having its own Swiss-based data center instead might actually be best to enforce its brand, and hence its positioning in the market.

So, when making a decision based on security, you must ask yourself the hard questions. Are you able to continuously invest in the head-to-head race on security? Are you able to maintain the people to do it? Are you willing to regularly upgrade your equipment to not fall behind on security standards?

While on the topic of security, I would like to circle back to quantum technology. Quantum computing can be used to crack passwords in order to perform a security attack. Additionally, quantum technology can also be used to defend. Due to its fragile states and the property of quantum entanglement, a quantum internet would allow for virtually non-interceptable communications, as any activity from intruders would irrevocably alter the message. China is already developing a prototype of the first quantum internet.

As a leader, it’s important to take actions now. Symmetric encryption algorithms can already provide enhanced protection against quantum brute force attacks. It’s vitally important to identify security threats early, and keep apprised of industry leaders like IBM, Microsoft’s PICNIC or Google’s NewHope as they develop post-quantum cryptography.

Video 17: Strategic Considerations: Build Advantage into Chips

Arguably, there has never been a better time to venture out and invest in new businesses. At first glance, the processor, data computing, and cloud landscapes seem saturated. However, instead of living in a saturated market, we are at the brink of a paradigm shift, and the cards are still very much being shuffled. We don’t have the next AI chip standard, we don’t have a solution for the high energy costs of the internet globally, and we don’t know how to manage the global yet fragile supply chains that enable the backbone of our era. What we do know is that sooner rather than later, innovators will introduce new approaches to these and other challenges we’re facing.

When it comes to blue sky research, consider IonQ that experimented on a different path for quantum computing that was different from that of Google’s and IBM’s, and developed a quantum computer 200.000x as strong as theirs.

Looking at the artificial intelligence chips industry, we have learned that there are more than 100 players in the industry, all working on slightly different designs for slightly different platforms. Since current AI technology will not grow into a generally applicable AI anytime soon, narrow AI is moving into niche applications, as companies are seeking relief from industry specific challenges.

For business leaders, this brings unique opportunities to build partnerships with chip designers within your industry, and to build the hardware platform for your industry’s cognitive revolution. Irrespective of your industry, you now have the unique opportunity to partner with providers of AI solutions to develop products tailored to your customers. Then, you can work with AI chip designers to integrate upstream into the compute power value chain to design use case specific chips.

Video 18: Strategic Considerations: Become Energy-Ready

The vast aggregation of data centers puts financial, environmental, and hence also political pressure on data center providers. Emerging Edge Computing technology is based on low-powered devices. However, it’s not a viable antidote to the rising energy needs, as its low-powered devices only conduct basic computing, usually before aggregating the data in the data center after all.

Outsourcing allows for more efficient energy use, and to some degree passes on electricity savings. If outsourced data storage is strategically infeasible, or if a hybrid approach to computing power is critical to business success, partnerships with energy management companies or proprietary investments in energy-savings solution providers can present strategic cost saving opportunities.

Some of the big tech companies like Amazon have pledged to achieve carbon neutrality by 2040, with some even going further by wanting to become carbon negative by 2030. Microsoft has the ambitious goal of removing all its historical carbon emissions by 2050.

As companies grow increasingly concerned with the cost capability boundaries and second or third order effects of energy consumption, we can expect to see a lot of innovation in the space of energy-improved computing, decentralized and smart grids, and battery technology. This will all impact the way we store and compute data in many ways.

For business leaders, becoming energy efficient is more than a service to the environment, but something of strategic importance. When everything runs on data, everything requires energy. Identifying energy efficient solutions and building efficient infrastructures early-on can generate critical cost and branding advantages.

Video 19: Strategic Considerations: Create Compute Power Resilience

The 2019 U.S.-China trade war, or the sudden stress on international supply chains due to the COVID-19 pandemic, can all have substantial effects on your organization's ability to source its required computing power. The Integrated Circuits supply chain is highly interconnected, globally dispersed, object of political plays, and therefore vulnerable.

The obvious short-term answer to computing power resilience is therefore investing in redundant failover cloud infrastructure, i.e. ensuring that failing servers and data centers are backed up by other servers and centers that are on “hot standby,” as it were, should the original equipment fail or be under attack. . Medium term this also means that we need to start thinking about alternative sources, alternative forms or approaches to creating resilience for critical capabilities in the future. Resilience for computing and data storing technologies is not a one-time decision but a strategic business principle. It creates brand equity with customers, competitive advantage vis-a-vis less vigilant rivals, operational cost-reduction and regulatory compliance.

Similar to supply chain analysis for precious metals in the battery market for example, Chief Information Officers must also conduct horizon scanning and scenario analysis to discover opportunities for new developments in the industry. That enables contingency plans early-on to create compute power resilience.

Especially with regards your own contingency plans, striking partnerships and alliances with designers and manufacturers creates resilience and strategic optionality for you.

More broadly speaking, I recommend that business executives also invest in broad analysis and due diligence by public policy and government relations departments. It’s critical to continuously scan and diagnose the health of geo-political and economic conditions that can impact your cognitive computing assets. This is not only helpful in the area of compute power, but across all tech domains as they are all reshaping the relationship between technology, society, and the economy.