BANK OF AMERICA INSTITUTE

BANK OF AMERICA

Transformation

Next Gen Tech: Computing

07 June 2024

Key takeaways

- The number of connected devices is expected to reach 350 billion by 2025 and one trillion by 2035, underscoring that we live in a world of exponential data growth. And with the AI revolution only intensifying, we'll see AI training and inference costs continue to rise, and in turn, so will the demand for faster and stronger computing power.
- There's no shortage of computing technologies on the horizon to address these challenges, but here we highlight computing innovations that have the potential to change the way we live and work, including: high-performance, spatial, edge, neuromorphic and quantum computing.
- Bank of America Institute's 'Next Gen Tech' series explores 30 breakthrough technologies across artificial intelligence (AI), computing, robots, communication, healthcare, energy and mobility, that are about to alter our lives. Join us as we discuss what's next on the tech horizon.

This publication is part of Bank of America Institute's 'Next Gen Tech' series – focused on sharing 30 breakthrough technologies that have the potential to transform the world. The series will highlight one of seven categories (artificial intelligence, computing, robots, communication, healthcare, energy and transport), and share advancements within each, so stay tuned for more.

Computing evolution

In our series introduction, Next Gen Tech: Breakthroughs that will transform the world, Bank of America Institute discussed how rapid shifts in innovation are transforming businesses and the world. In fact, we noted that the fastest transformation in human history is ahead of us. First, we discussed innovations in artificial intelligence (AI) and here, we'll share the second of seven categories of breakthrough technologies – computing.

BofA Global Research notes that advances in processing power have led to an evolution in the computer, where traditional processing units and large compute clusters cannot breach the boundaries of computational complexity. We live in a world of exponential data growth and, for the first time, Moore's Law (the observation that the number of transistors on an integrated circuit doubles every two years), falls short of explaining the demand for faster and stronger computing power.

And then came generative AI...

Demand for semiconductors has been accelerated by secular growth in connectivity. According to BofA Global Research, the number of connected devices is expected to reach 350 billion by 2025 and one trillion by 2035. By 2025, we could be interacting with connected devices as often as once every 18 seconds (4,785 times a day), compared to "only" every 2.4 minutes today.[1]

Additionally, AI training and inference costs continue to rise. Model sizes (i.e., number of parameters) for large language models (LLMs) have grown exponentially over the past few years, evolving from 94 million parameter LLMs in 2018, to the commercially available 175 billion parameter GPT-3, and the estimated more than one trillion parameter GPT-4. And according to Nvidia, the processing power needed to train generative AI models is increasing 275x every two years. So, we ask, what's here, what's next, and what will transform computing to address the challenges ahead?

1) Exascale computing & high-performance computing

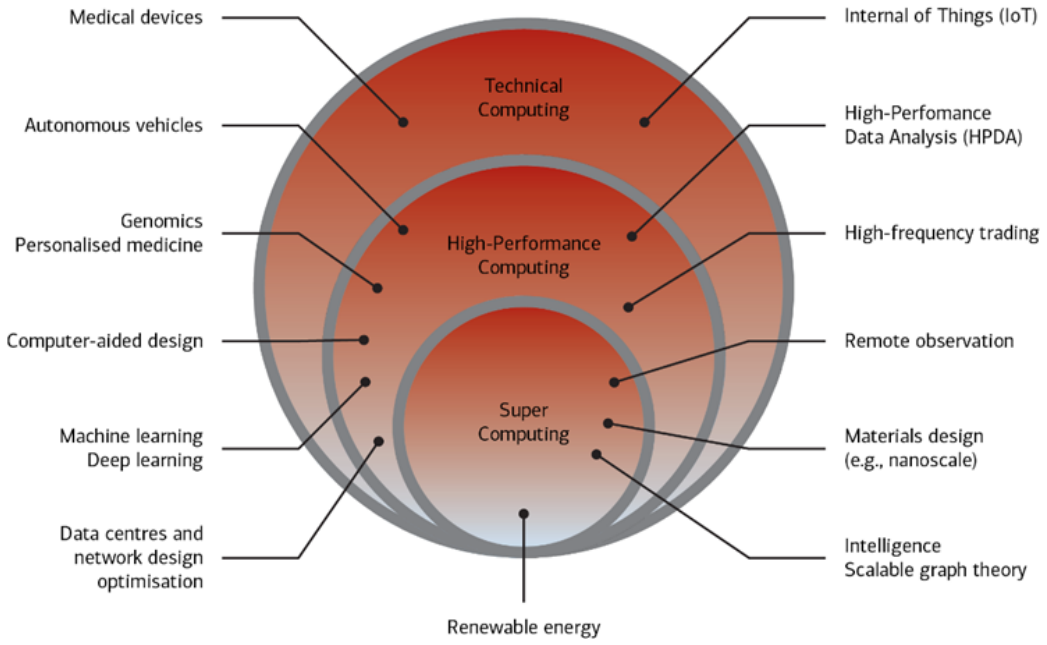
High-performance computing (HPC) is technology that uses clusters of powerful processors, working in parallel computing, to process massive multi-dimensional datasets (big data) and solve complex problems at extremely high speeds. HPC systems are typically more than one million times faster than the fastest desktops, laptop or server systems.

For example, HPC systems can perform quadrillions (i.e., 10^15) of calculations per second, compared with regular laptops or desktops that can perform at most three billion per second (with a three GHz processor). Exhibit 1 shows how HPC, as a methodology, can be used in many domains including sequencing DNA, automating stock trading, and running AI algorithms and

[1] International Data Corporation (IDC), IoT Analytics

simulations – like those enabling self-driving automobiles – that analyze terabytes of data streaming from IoT (internet of things) sensors, radar and GPS (global positioning) systems in real time to make split-second decisions.[2]

Exhibit 1: HPC, as a methodology, has a wide applicability to both established and emerging domains, such as autonomous vehicles, the Internet of Things or precision agriculture
Relevance of HPC to strategic and emerging domains



Source: Irish Centre for High-End Computing (ICHEC)

BANK OF AMERICA INSTITUTE

TOP500 List: Accelerated compute drives performance

TOP500, a project that assembles and maintains a list of the 500 most powerful computer systems in the world, provides bi-annual updates based on recent advancements. The key takeaways from the most recent TOP500 List of supercomputers are:

- Aggregate HPC performance of >7 exaflops (one exaflop is equal to one billion billion (10^18) FLOPS) (vs. 5.2 exaflops in May 2023), representing the highest half-over-half (HoH) increase on an absolute basis.
- Most of the incremental growth in performance came from the addition of four new computers in the top 10.
- 186 systems have been accelerated in total, representing an attach rate of 37%, in line with prior performance results (185 systems in June 2023)
- While accelerated systems are in the minority, they account for over 70% of aggregate compute performance.

From serial to parallel computing

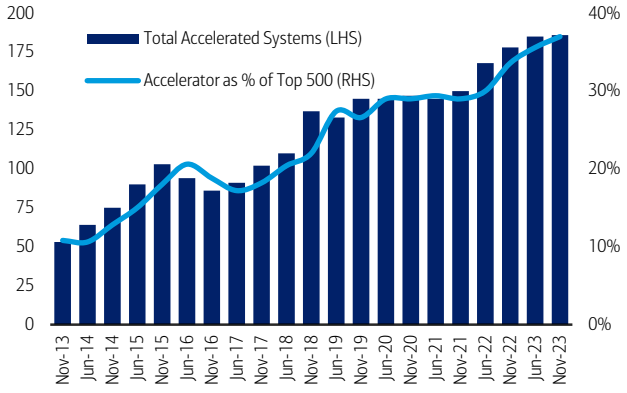
The maturation of Moore's Law and serial computing is shifting more workloads to parallel computing, implemented with separate co-processor/accelerators such as GPUs (graphics processing units), custom chips (ASIC, or application-specific integrated circuits) and programmable chips (FPGA, or field-programmable gate array). As of November 2023, 186 machines on the TOP500 List employed a co-processor, up from 137 systems five years ago (Exhibit 2). Co-processor/accelerator use across the TOP500 was flattish HoH and up ~5% year-over-year (YoY). Total compute performance of the TOP500 supercomputers grew to 7 exaflops (or 7,032 petaflops), which in November of 2023 was up 45% YoY (Exhibit 3).

In terms of aggregate compute capacity (petaflops, PFLOPS), the US accounts for 53% of deployed PFLOPS globally, recently benefitting from the new Frontier system (the world's first true exascale supercomputing system). This is up from ~38% realized in November 2018. Meanwhile, Japan has claimed the #2 spot with 10% of deployed petaflops, while China is #3 (<6%).

The United States share of the TOP500 systems is ~32%, up from 25% in November 2022 and ~22% five years ago (Exhibit 4, Exhibit 5). China previously held the number one spot globally (as recently as November 2022) and accounted for 45% of total systems five years ago (flattish HoH), but today, its share has fallen to ~21% (down ~1,2000bps YoY).

[2] IBM, TechTarget, NetApp

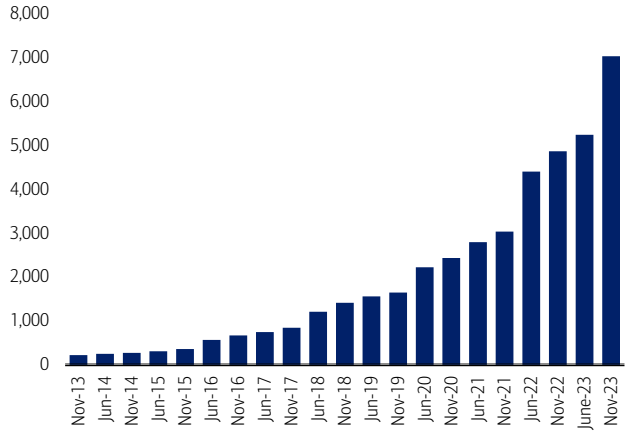
Exhibit 2: 186 of the Top 500 supercomputers have acceleration/co-processing, up from 137 five years ago
The number of Top 500 supercomputing systems with acceleration/co-processing



Source: TOP500.org

BANK OF AMERICA INSTITUTE

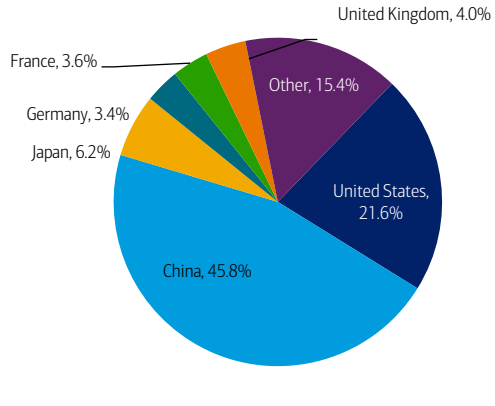
Exhibit 3: Total petaflops are up ~45% year-over year (YoY) in June 2023
Total petaflops for Top 500 over time



Source: TOP500.org

BANK OF AMERICA INSTITUTE

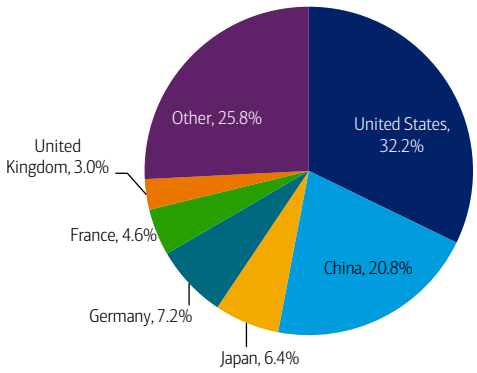
Exhibit 4: The US held a ~22% share of the TOP500 systems in 2018 and lagged China's ~46%...
November 2018 - % of Top 500 systems by region



Source: TOP500.org

BANK OF AMERICA INSTITUTE

Exhibit 5: But the US is now the market leader with 32% of the TOP500 systems
November 2023 - % of Top 500 systems by region



Source: TOP500.org

BANK OF AMERICA INSTITUTE

2) Spatial computing

Desktop workstations and personal computers (PCs) have been around for more than half a century, yet the way most people interact with them hasn't changed much. The keyboards we use today evolved from typewriters, a technology that dates back almost 150 years. Even the graphical user interface (GUI) has been around for a while – the first to gain popularity in the consumer market was on the Macintosh in 1984. Considering that computers are far more powerful today than they were 50 years ago, following Moore's Law, the basic interfaces haven't changed dramatically.

That said, we are reaching an inflection point in human-computer interaction. We are on the cusp of moving away from the traditional keyboard-and-mouse configuration, and towards touch gestures, conversational AI and augmented vision computing interaction. Spatial computing changes human/machine interaction by using AR/VR (augmented reality/virtual reality) to blend the graphic interface for the user to take place in the real, physical world.

Spatial computing revolution comes after the smartphone

Spatial computing has the potential to drive the next wave of disruption, following PCs and smartphones, where technology becomes integrated into our everyday behavior, with real-time data and communications bridging our physical and digital lives.

BofA Global Research recently noted that computing eras occur every 10-15 years, with each major new cycle completely reshaping the computing landscape. Original consumer PCs were introduced and commercialized in the 1980s and 1990s. Then,

mobile computing and the smartphones of the early 2000s allowed for the free flow of information, and enabled mobile messaging, social networking, the sharing economy, and the overall mobile economy. It all leads us to ask: Is the spatial computing revolution next?

3) Edge computing

Edge computing is a rapidly evolving segment that addresses the need to process data closer to where it is generated (the physical location where things and people connect with the digital world).

Edge versus cloud: What's the difference?

Edge computing complements cloud computing by addressing latency, bandwidth, autonomy, and privacy requirements. Most hyperscalers develop their offerings based on a 'cloud-out' principle that pushes the public cloud architecture to the edge. The core, typically the cloud or a centralized data center, is the upstream system that supports the edge with centralized storage, processing, and analytics at scale.

While edge compute is location-specific, many cloud attributes can still be applied, such as continuous integration and deployment, DevOps (a partnership between software development and operations used to enable coordination and collaboration), 'as-a-service' hardware management, and OPEX (operational expenditure) pricing.

Edge compute is still in an adolescent stage of maturity, with a survey by IDC (International Data Corporation) finding that 42% of enterprise respondents struggle with the design and implementation of key components, including infrastructure, connectivity, management, and security. In the long run, however, the combination of edge data aggregation, analytics and cloud access for analytics and model training, will create a new economy built on digitally enabled edge interactions. Use cases for edge computing range from inventory and fraud management to real-time retail promotions, or from contactless checkout to VR-enabled customer experiences.

Where is edge computing located?

Omdia defines the edge as locations with a maximum of 20ms (milliseconds) roundtrip time to the end user. Telecom operators refer to the edge as telco-operated sites, including central offices, regional data centers, and access networks, while cloud edge refers to cloud-provider operated sites. Enterprise edge sites include branch offices, industrial locations, and regional data centers.

Many enterprises are investing in edge locations (from internal IT (information technology) and OT (operational technology) to external, remote sites) to get closer to end users and where the data is generated. The enterprise edge is IT-heavy and usually includes locations that are classified as remote/branch offices without data center facilities. Most enterprise edge locations are controlled via specific management and provisioning capabilities based at the core.

What are the challenges?

The number of nodes needed can vary wildly and requires intense planning, which entails higher costs because of its distributed nature. Furthermore, as more devices are added to the network, data breach risks increase. However, keeping the data distributed so that a data breach only impacts a fraction of the data or applications can address this. Another barrier is the lack of a standard edge compute stack and API (application programming interface). Additionally, edge computing nodes have typically been use-case dependent to keep costs down, however, use cases will continuously change and enterprises will need to change their deployment strategies so that edge computing platforms are special-purpose and extensible.

Content delivery network is a form of edge computing

A content delivery network (CDN) is a distributed infrastructure composed of internet points-of-presence (POPs), which are clusters of locations containing caching servers, or nodes, power/cooling, switches, and optics. Some CDN vendors have light enough POP footprints that they can collocate their servers in locations where other providers can't fit their footprints. CDN vendors "rent" bandwidth from service providers, as well as space within their data centers to house caching servers, which store content at the edge of the network. CDNs can balance traffic, decrease latency, prevent network outages, and provide secure connections and real-time analytics.

Whilst content and media providers typically use CDNs to augment their networks, especially at high-traffic events (e.g., the Super Bowl or a TV show premiere), the digitalization of enterprises is driving the adoption of CDNs, especially those with large amounts of digital property (e.g., news outlets, social media companies, travel companies, financial service providers, and others). Hyperscalers use CDN vendors too, though some have chosen to build their own CDNs using existing network infrastructure – a strategy BofA Global Research believes is cost-prohibitive for most enterprises.

CDNs x edge compute: Extending the digital world to the physical world

Edge compute runs on the distributed topology of CDN providers and is the closest level of compute to end users. It takes place on a continuum between the edge and the central core, with parts of compute running on devices such as smartphones and tablets, while other portions take place at edge servers, small data centers, or CDN locations.

In the context of CDN providers, most edge compute uses existing POPs and CPUs (central processing units) to execute compute functions at the edge. Unlike centralized cloud, edge compute brings specific, lightweight pieces of compute functionality to edge servers, often in containers, which hold the code and dependencies necessary to perform the compute tasks. Edge compute solutions leverage the global footprint of CDN networks to dynamically optimize compute resources and enhance the overall efficiency of the compute process.

AI opportunity from inferencing

Unlike training, which will occur in core compute with hyperscalers, inferencing requires a distributed, scalable, low-latency, low-cost model, which is what the edge compute model provides.

The current divide in edge compute is whether to use CPUs or GPUs for inferencing. While all main vendors support both capabilities, BofA Global Research believes CPUs are the optimal choice to support inference at the edge. GPUs are powerful for compute but are uniquely bad at living anywhere outside of core compute. They consume a lot of power, resulting in costly cooling measures, and make hardware design complicated. Under the GPU model, only six to eight requests can be processed at a time. However, CPUs can subdivide servers by users, making it a more efficient processing system at the edge.

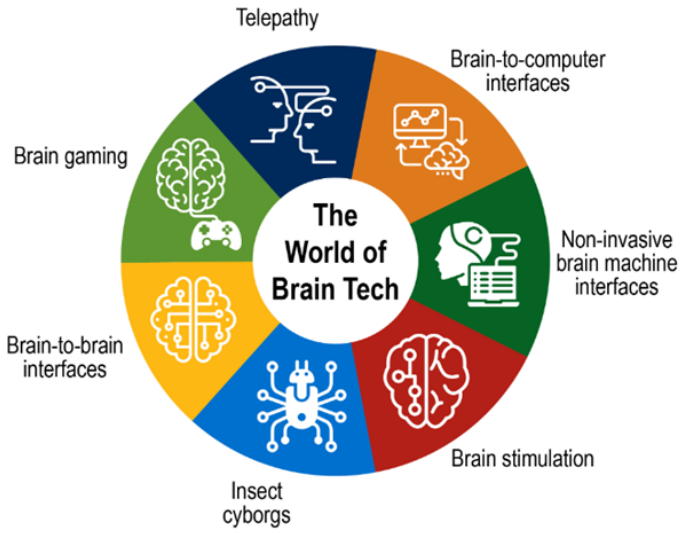
4) Brain-computer interface / neuromorphic computing

Neuromorphic computing or brain-computer interface (BCI), sometimes called brain-machine interface, is where the brain waves of humans and animals directly interact with the external world and vice versa.

Human-machine implants: If you can't beat them, join them

Though AI has the potential to become more intelligent than humans, some start-ups are focusing on creating a human-machine cooperation that will prevent people from being left behind. As this is a nascent technology, different approaches are being trialled with different motives and consequently, brain computer interfaces differ in invasiveness and ability (Exhibit 6).

Exhibit 6: The world of brain tech. Technological approaches vary from brain stimulation techniques to brain gaming.
How is brain technology being developed?



Source: BofA Global Research

BANK OF AMERICA INSTITUTE

How would brain computer interfaces work?

This concept of brain implants is not completely new. One demonstration of a brain implant already in use is cochlear implants that stimulate the auditory nerve to provide some form of sound for hearing-impaired individuals. Academics have already completed studies of implants using a few hundred electrodes to collect brain data in animals. In addition, there are already procedures to stimulate parts of the brain for patients with Parkinson's Disease and clinical depression, and minimally invasive implants are now being trialled in humans. However, regulation of such a contentious technology will no doubt play a key part in the rollout of BCIs.

5) Quantum computing

According to BofA Global Research, quantum computing could be one of the biggest revolutions yet. In short, it leverages sub-atomic particles to store information and uses superpositions for complex calculations. A quantum computer can solve problems near instantaneously that would take a classical computer billions of years.

How it works: Bizarre and mysterious lesson 101

Although the possibility of a quantum computer was proposed as far back as 1982, it was only in 2012 that the first venture was undertaken to explore commercial use. A quantum computer is a machine based on quantum mechanics – the branch of physics that describes behavior at the sub-atomic level. Quantum computers operate on three principles of quantum mechanics, which, technically, are not possible: superposition, entanglement and qubits.

- **Superposition**: If a classical computer can work on two possibilities only (0 or 1, electricity is either on (i.e., 1) or off (i.e., 0)), the superposition phenomenon theoretically translates into endless possibilities and, as a result, endless calculations. Unlike classical physics, superposition posits that a particle can be in more than one state at any given time: two states, more than two or even none. For example, a photon could be up, down, up and down, in no state, or anything in between, all at the same time.

  But here is where it gets even weirder. Superposition has one condition – it requires uncertainty. Once we observe and measure the position (i.e., add certainty to the equation) the superposition states collapse into one. Meaning that in order to stay in a superposition and make endless calculations, we need uncertainty.

- **Entanglement**: This refers to two or more quantum systems being linked in an unseparated bond, even if far apart – meaning this status cannot be described independently, even if the distance between them is vast. Any effect on one particle will instantly impact the other, faster than the speed of light, no matter the distance between them. If superposition offers endless states and therefore calculations, entanglement provides parallel processing of the system and allows it to scale up and work as one and other properties like error correction or control interference.

- **Qubits:** In other words – quantum bits – a quantum computer's basic unit of information. Superposition of a particle can exist in a qubit and multiple qubits can be entangled – meaning one qubit can affect another. In short, these units contain the quantum mechanics properties of superpositions and entanglements. Without qubits, those properties could not be harnessed to create a quantum computer. To complicate matters yet further, qubits can act like waves, or particles and can tunnel through energy barriers – which is impossible according to classical physics – and just "turn up on the other side." However, to achieve these qualities, qubits need to be fully isolated, and if not, they "collapse" and become "normal bits."

The "power couple" that will change the world

We can talk about many applications, but one of the most exciting is the marriage of the two most powerful technologies: AI and quantum computers. The convergence of AI and quantum (AQ) technologies can enable fundamental improvements in the physical world as well as the digital one. While quantum computers will provide endless calculations when available, the increased capability of AI technologies can unlock several transformational use cases in the meantime.

LLMs complemented with simulation, knowledge graphs, computer vision and predictive analytics are 'the new AI toolbox' that companies can deploy. Iterative research tasks that would have taken years can now be achieved in weeks. Life sciences, chemicals, materials, and finance/logistics are all in scope to benefit. And in the long run, artificial general intelligence (AGI) – when AI reaches human cognitive abilities and even self-awareness and singularity – the point when AI surpasses human intelligence – will lead to exponential, radical transformation of technology.

A quantum computer is not suitable for regular tasks like using the internet, office tasks or emails, but for complex big data calculations like blockchain, machine and deep learning, or nuclear simulation. A combination of quantum computers and 6G mobile networks would be a game-changer in every industry. For example:

- **Big data analytics:** The amount of data created is projected to double every two-to-three years, reaching 183ZB (zettabyte) by 2024E from 120ZB in 2022 and 12ZB in 2015.[3] The untapped big data potential is huge given that only c.0.5-1.0% of data generated has ever been analyzed. Today, we are storing, transmitting, and using only 1% of global data,[4] because we lack the computing power needed to process more. Quantum computing could change that and unlock real economic value. Using 24% of global data (unlike 1% today) would double global GDP.[5]

- **Financials:** Quantum computers can master almost every aspect of banking. For example, the data calculation capabilities of quantum computers are better at predicting market trends, portfolio optimization, data analytics in real time, pattern detection, encryption and fraud detection.

[3] Statista, IDC
[4] IDC
[5] IDC

- **Cybersecurity:** Quantum computing can technically challenge all current encryption methods (based on large number factoring), including blockchain, by parallel processing capabilities of up to 1 trillion calculations per second.[6] This also opens the door to new encryption technology, based on quantum computing elements.
- **AI and machine learning:** Quantum computers can speed up machine learning capabilities by using more data faster and solving complex connections between data points. Machine learning and deep learning progress is limited to the pace of underlying data calculation. The faster data is calculated and used, the faster machine and deep learning algorithms will evolve.
- **Infrastructure:** One area where quantum calculation is needed is complex network calculations. Quantum computers can reorganize global telecoms, utilities, transportation and other combined infrastructures that are becoming more integrated and too complex for current computing abilities.
- **Healthcare and genomics:** If global data is expected to increase twofold every two-to-three years, medical knowledge is expected to double every 73 days.[7] Each person will generate enough health data in their lifetime to fill 300 million books. Genomic data will double every 50 days and the amount of global genomic data will surpass YouTube and X by 2025.[8] Quantum computers could be the solution for all big data processing in this field. Big data utilization would not only potentially bring down costs, but also help provide better healthcare through targeted treatment and predictive analytics.[9]
- **Science:** Complex calculations used in space (e.g., big bang simulations), physics (e.g., dark matter, string theory), nuclear simulations, complex materials and 'super-connectivity' can leapfrog to the next level with quantum computing.
- **Cloud:** This could be one of the winners, as the cloud could be the platform where all data creation, sharing and storage take place. Once the commercialization of quantum computers begins, cloud access will be needed, and data generation should jump exponentially. Thus, cloud platforms will be the solution.
- **Autonomous vehicle (AV) fleet management:** One connected AV will generate the same amount of data as 3,000 internet users.[10] However, two cars will generate the same amount as around 8,000-9,000 users as the two will need to communicate with each other, and they will be generating data and so on. The growth of data from AVs alone will be exponential. Thus, on this calculation, a fleet of 1,000 cars will generate more data than the entire global population. A regular computing system would not be able to handle this.

[6] Bernard Marr
[7] American Clinical and Climatological Association, P. Densen
[8] PLoS Biol. "Big Data: Astronomical or Genomical?" 13 July, 2015.
[9] IMS Health
[10] Intel

Contributors

Vanessa Cook

Content Strategist, Bank of America Institute

Sources

Haim Israel

Equity Strategist, BofA Global Research

Felix Tran

Equity Strategist, BofA Global Research

Martyn Briggs

Equity Strategist, BofA Global Research

Lauren-Nicole Kung

Equity Strategist, BofA Global Research

Disclosures

These materials have been prepared by Bank of America Institute and are provided to you for general information purposes only. To the extent these materials reference Bank of America data, such materials are not intended to be reflective or indicative of, and should not be relied upon as, the results of operations, financial conditions or performance of Bank of America. Bank of America Institute is a think tank dedicated to uncovering powerful insights that move business and society forward. Drawing on data and resources from across the bank and the world, the Institute delivers important, original perspectives on the economy, sustainability and global transformation. Unless otherwise specifically stated, any views or opinions expressed herein are solely those of Bank of America Institute and any individual authors listed, and are not the product of the BofA Global Research department or any other department of Bank of America Corporation or its affiliates and/or subsidiaries (collectively Bank of America). The views in these materials may differ from the views and opinions expressed by the BofA Global Research department or other departments or divisions of Bank of America. Information has been obtained from sources believed to be reliable, but Bank of America does not warrant its completeness or accuracy. These materials do not make any claim regarding the sustainability of any product or service. Any discussion of sustainability is limited as set out herein. Views and estimates constitute our judgment as of the date of these materials and are subject to change without notice. The views expressed herein should not be construed as individual investment advice for any particular person and are not intended as recommendations of particular securities, financial instruments, strategies or banking services for a particular person. This material does not constitute an offer or an invitation by or on behalf of Bank of America to any person to buy or sell any security or financial instrument or engage in any banking service. Nothing in these materials constitutes investment, legal, accounting or tax advice.