Ubiquitous Computing Power: the Cornerstone of an Intelligent Society

Position Paper

February 2020
Computing is a way of perceiving the world. From mainframes to PCs, and from PCs to smartphones and wearables, computing has become a de facto extension of human capability.

Our approach to computing is also evolving. Statistical computing will soon become mainstream, and we estimate that five years from now AI applications will account for more than 80% of all computing power used around the world. It will be a new age of intelligent computing.

The technologies behind computing and connectivity are progressing by leaps and bounds every day, and they will soon pave the way for a world where all things sense, all things are connected, and all things are intelligent. Ubiquitous computing power will be the cornerstone of this future world. To make sure that everyone and everything has access to the computing power they need, all industry players need to work together and develop a common understanding around three key concepts.

First, ubiquitous computing power is the cornerstone of an intelligent society. We all know that per capita GDP is the primary measure of a country’s economic productivity. Similarly, we use per capita computing power to mark the different stages of development in smart countries.

Currently, per capita computing power in major countries and regions ranges from 100 to 2,500 GFLOPS. Nevertheless, places with high per capita computing power are still in the very early stages of their intelligent journey; they will enter the next stage of growth – the developing stage – only when their per capita computing power surpasses 10,000 GFLOPS.

Just like the wide adoption of electricity, which laid the foundation for an industrial society, ubiquitous computing power will become the cornerstone of an intelligent society.

Second, we need to build a diversified computing power ecosystem to drive the computing industry forward. In the intelligent world, we will see more diverse application scenarios and data types than ever before, which calls for a much more diversified computing power architecture.

If we hope to succeed in building a prosperous and diversified computing power ecosystem, we will need new innovation in the underlying architecture and greater collaboration between customers, industry partners, and developers. This is the only way to build out the ecosystem more rapidly and provide new momentum for the computing industry as a whole.

For example, the Green Computing Consortium is a joint effort between leading global providers of servers and cloud computing solutions who are working together with Chinese universities and research institutions. This Consortium has brought together a remarkable number of industry resources and experts to help promote the development of a diversified computing power ecosystem.

The European Processor Initiative (EPI) is a similar
project implemented by the EPI Consortium. As of January 2020, it has gathered 27 partners from 10 European countries who work together across a number of domains, including R&D, production, and application scenarios for computing and chips. This has helped to promote robust development in the European computing industry through more concerted cross-sector collaboration.

Third, investing in computing power will promote innovation, boost economic growth, and improve people's lives, raising the overall competitiveness of many nations. Analysis across a number of countries has revealed that investments in computing power boost economic growth in two ways: by directly promoting growth in the local ICT industry, and by contributing to innovation in many others, including manufacturing, transportation, energy, retail, and agriculture. Manufacturing serves as a typical example – every dollar invested to make plants more digital and intelligent can stimulate ten dollars in growth.

Higher computing power in sectors related to people's livelihood, such as education, scientific research, and healthcare, will improve people's welfare and increase their sense of satisfaction and happiness. Therefore, investments in computing power should be placed on a national strategic level, as well as planned and implemented strategically.

In China, we have already begun enjoying the fruits of more readily available computing power. Every day, nearly 900 million people in China – from bustling cities to remote villages – can pay for anything, anywhere with a tightly integrated mobile payment ecosystem. From paying for meals and transportation with QR codes, to shopping online and in brick-and-mortar shops alike, each payment is completed instantly – a feat that would not be possible without widespread connections and computing. This has helped to greatly improve consumer experience and guarantee payment security.

Over the past 30 years, Huawei has committed itself to delivering the world's best connectivity through heavy investment in R&D and ongoing innovation. With the advent of an intelligent world, we have expanded from connectivity alone to connectivity and computing together in order to promote the development of a diversified computing industry. With these efforts, we hope to provide the world with more abundant and more affordable computing power, create long-term value for our customers, and contribute to society as a whole.

In the next 30 years, as computing power becomes effectively omnipresent, AI will become deeply integrated into all industries, driving deep changes in each and every one. It will bring brand-new experiences to every home and every individual, fundamentally changing the way we work and live.

A journey of a thousand miles begins with a single step. To create a bright future with ubiquitous computing power, everyone in the industry needs to go all in. We are ready and willing to work with all of our partners in the computing industry and make "ubiquitous computing power" a reality.
# Table of Contents

**Executive Summary** ................................................................. 02

**Chapter 1. Definition, Evolution, and Future Trends of Computing Power** 05

1.1 Definition and Evolution of Computing Power .......................................................... 05
1.2 Three Key Stages and Future Deployment of Computing Power ................................. 07
1.3 The 4A Requirements for Computing Power by Key Technologies ........................... 14

**Chapter 2. Computing Power Evaluation Model** ...................................................... 25

2.1 Introduction to Computing Power Evaluation Model ................................................. 26
2.2 Results of Computing Power Measurement .................................................................. 35
2.3 Demand for Computing Power Will Grow Exponentially as Smart Society Evolves ........ 38

**Chapter 3. Benefits of Computing Power to Countries and Regions** ................. 47

3.1 Investing in Computing Power Will Generate Sizeable Direct Economic Returns ........ 47
3.2 Investing in Computing Power Will Have Positive Spill-over Effects to Other Industries .... 48
3.3 Investing in Computing Power Will Lift Comprehensive Nation Power by Driving Innovation, Optimize Public Service and Improve People’s Welfare ................................................................. 51

**Chapter 4. Challenges in Computing Power Development and How to Tackle Them** .... 55

4.1 Developing Computing Power Faces Challenges of Power Consumption, Network Limitations and Diversified Ecosystem ................................................................... 55
4.2 Governments can Make Targeted Efforts to Drive Computing Power growth: Making Policy Planning, Building Infrastructure and Providing Services, Extending Fiscal Support, Enhancing Legal Framework and Supervision, etc. .......................................................................................... 65
Executive Summary

From the vintage era of abacus to the present era of smart devices, computing power has increased exponentially and is fitting into more diverse ecosystems. In the present era, along with silicon-based chips, computing power has evolved through three stages, namely, “single-core”, “multi-core” and “networked”. Computing power of single-core chips will peak at 3nm. Moving down from 3nm would be too difficult to manufacture and too expensive to commercialize. Multi-core chips seem to be the solution – but not a perfect one. As more cores stack up, dissipation effects and rising power consumption will make the trade-off of more cores for more computing power uneconomical – the tipping point is at 128 cores. When the capacity of single-core and multi-core chips is exhausted, “networked” computing power will come in to fulfil the demand gap. Meanwhile, to break free from the limitations posed by the existing network technology and bandwidth cost, deploying computing power on the edge will be a must to address the demand-supply mismatch. Eventually, an ideal cloud-edge-device ubiquitous computing power deployment architecture will take form.

The future demand for computing power will be enormous. The undergoing smart society process involves using such key technologies as Artificial intelligence (AI), Internet-of-Things (IoT), blockchain and Augmented Reality (AR)/Virtual Reality (VR) in different scenarios to boost productivity, improve people’s lives and make governments more efficient. Wide application of these technologies is bound to engage more computing units on the cloud, the edge and the device level, driving computing power to grow steadily. Specifically, by 2030:

- AI will penetrate all industries and will carry a computing power of 16,000 EFLOPS\(^1\) (combined computing power of 160 billion Qualcomm Snapdragon 855 NPU\(^2\) embedded in today’s smartphones);
- IoT will connect all appliances in factories and households, and require 8,500 EFLOPS of computing power deployed on the cloud and the edge (combined computing power of 7.9 billion chips used in high-end IoT edge devices);
- Blockchain technology underpinning digital currency and other areas will engage a computing power of more than 5,500 EFLOPS (combined computing power of 1.3 billion units of AntMiner V9);
- AR/VR, when playing to its full potential, will have more than 3,900 EFLOPS of computing power (combined computing power of 2.1 billion units of SONY PS4 consoles).

To meet all the above demand, computing power in the future should be available Anytime, Anywhere, at Any Capacity, and in Any Object.

To better understand computing power and effectively measure that of each country, we have put together a Computing Power Evaluation Model. It measures the computing power of

---

\(^1\) EFLOPS = exaFLOPS. Floating point operations per second (FLOPS, flops or flop/s) is a measure of computer performance. 1 EFLOPS = \(10^9\) GFLOPS = \(10^{18}\) FLOPS

\(^2\) Neural Processing Unit (NPU) is a microprocessor that specializes in the acceleration of machine learning algorithms, typically by operating on predictive models such as artificial neural networks (ANNs) or random forests (RFs).
all typical computing units relying on each of the three levels of computing power-carrying infrastructure, i.e. the cloud, the edge and the device level. Cloud computing power mainly resides in supercomputers and data centres. Their computing power is calculated using the number of servers and average computing power per server, while factoring in dissipation effects and transmission lost. Edge computing power exists in edge servers such as CDNs\textsuperscript{3}, intelligent gateways. This part of computing power is calculated based on the amount of edge servers and their average capacity. Network transmission lost is also considered. Device computing power includes the computing power of all smart devices, i.e. smartphones, laptops, smart wearable devices, etc. By calculating and comparing the computing powers of 27 countries, we conclude that a country's computing power is tightly coupled with its level of economic development, and therefore, is an effective indicator of comprehensive national power.

More importantly, per capita computing power\textsuperscript{4} will be an important indicator marking different stages of development of a smart society. In the early stage, i.e. the current phase, per capita computing power in major countries and regions in the world ranges from 100–2500 GFLOPS\textsuperscript{5}; Smart devices are applied in limited industrial scenarios to facilitate easy tasks like automated production. Going forward to the developing stage where the per capita figure will reach 10,000 GFLOPS, pervasive AI technology capable of proactive analytics will reinvent the ecosystem and business models in such industries as manufacturing and driving, making them vertically integrated and highly intelligent. Into the mature stage, the per capita computing power in major countries and regions will surpass 29,000 GFLOPS; Self-evolving AI will penetrate most industries and enable ubiquitous smart scenarios.

Further analysis of computing powers of different countries reveals that computing power can bring sizeable direct economic benefits, boost growth of peripheral industries, and improve public service efficiency and people’s lives. First, higher computing power directly boosts the growth of hi-tech, high value-added industries like the ICT industry with strong leverage effect. One dollar invested in the R&D, manufacturing, application and talent cultivation related to computing power infrastructure (chips, servers, data centres, intelligent devices and high-speed internet, etc.) will generate 4.7 dollars of return. Second, more investment in computing power leads to more advanced digital and intelligent technologies, which, in turn, will contribute to the output, efficiency, innovation and customer experience of many industries, including manufacturing, transportation, retail, etc. Manufacturing serves a typical example – every dollar invested to make plants more digital and intelligent will create ten folds of return. Investing in computing power will

\textsuperscript{3}Content Delivery Network or Content Distribution Network (CDN) is a geographically distributed network of proxy servers and their data centres. The goal is to provide high availability and high performance by distributing the service spatially relative to end-users.

\textsuperscript{4}Computing power per capita= Total computing power of Country A / Country A's population

\textsuperscript{5}gigaFLOPS
promote innovation, boost government efficiency and improve people's lives, raising a nation's overall competitiveness.

Investment in computing power will pay back. Higher computing power in innovation-intensive sectors will improve a nation's scientific and technological excellence; Higher computing power in public service industry will make government more efficient in serving the public, thus contributing to a peaceful and harmonious society; Higher computing power and more digital infrastructure in sectors concerning people's livelihood like education, scientific research, healthcare, employment and community service will improve people's welfare. Even though the marginal value of computing power investment is bound to diminish as the society becomes more intelligent, the absolute benefits of such investment is still substantial.

While striving to develop computing power and harvest its advantage, there are three challenges and one concern that should be addressed by governments.

As a prerequisite, governments should actively build computing power infrastructure, develop key technologies like cloud computing and AI and make them more accessible.

- The first challenge is surging power consumption. Governments should offer guidelines for deploying energy-conserving and green computing power infrastructure, ensuring sustainable development of the smart society. To this end, governments can: 1) advocate industry association to set energy-conservation goals, 2) directly invest in new energy industry, and 3) formulate a carbon emission exchange mechanism and put a high price tag on limited emission quota, forcing market players to cut emissions.

- The second challenge is demand for higher-quality network. Governments should mobilize its policy tools and resources to improve the network infrastructure and lay a solid foundation for developing ubiquitous computing power. Core policies may include: 1) spread out an extensive monitoring matrix to supervise the status of network deployment and 2) leverage private capital to build network infrastructure through Public–Private Partnership (PPP).

- The third challenge is increasingly diverse ecosystem. Governments should encourage enterprises of different types to thrive and attract more high-calibre talents, who can make full use of available computing power to materialize more intelligent application scenarios. Potential policies may include: 1) set goals to support the development of diverse enterprises who will contribute to a mobile, multi-dimensional architecture of computing power and the growth of computing power ecosystem, 2) encourage industrial alliances to integrate upstream and downstream resources and provide talent support, and 3) broaden access to computing power market and streamline relevant administrative procedures to create an enabling policy environment.

Apart from the above challenges, a crucial concern is security. While building and operating a smart society, governments should be able to monitor and prevent potential security risks. Feasible policies may include: 1) coordinate with other governments to formulate international security standards, 2) establish a public-funded supervising authority for computing power security and allow third-party supervision, and 3) establish an accountability mechanism to regulate market participants.
Chapter 1

Definition, Evolution, and Future Trends of Computing Power

Key Takeaways from Chapter One

1.1 Definition and evolution

- Computing power refers to the capability of delivering designated results by processing data
- Computing power could be realized through devices in various forms, evolved from manual device, mechanic device, electric device, integrated circuit to today’s mobile device

1.2 Three key stages and future deployment

- Single-core processor will reach its limit at 3nm due to performance and cost constraints: quantum tunnelling effect and commercial feasibility
- Multi-core processor will reach its limit at 128 cores: data transfer, operating system, software and power consumption per computing power
- “Networked” computing power calls for more edge devices to form the “cloud-edge-device” ubiquitous architecture: flexible and cost-efficient edge devices offer solutions to network and bandwidth limitations

1.3 Technology-driven demand & 4A features

- Computing power will be the key driver for a smarter society
- Thanks to AI, IoT, blockchain and AR/VR technologies, future computing power will be available Anytime, Anywhere, in Any Capacity and on Any Object,

Source: Desktop research; Roland Berger

1.1 Definition and Evolution of Computing Power

As a trending topic, computing power has induced a significant amount of discussion, yet the definition of it remains obscure. William D. Nordhaus\(^6\), the 2018 Nobel laureate in Economic Sciences, defined computing power as "the amount of information delivered per second by the machine – that is, the quantity of information produced as the machine moves from one internal state to another". Referring to his and other authoritative definitions, we define computing power as "the capability of delivering designated results by processing data".

Evolution of the form of devices enabling computing power

“Computing power is the amount of information delivered per second by the machine – that is, the quantity of information produced as the machine moves from one internal state to another.”

---William D. Nordhaus, 2018 Nobel Laureate

Abacus (manual device)
- Abacus is a simple calculation tool invented by ancient Chinese

Ancient times

1642
- Mechanical calculator (mechanical device)
  - French scientist Blaise Pascal invented the first mechanical calculator based on abacus

1937
- Electronic digital computer (electronic device)
  - Atanasoff Berry invented the first electronic digital computer

1947
- Transistor and integrated circuit
  - Bell Labs invented the transistor in 1947; Jack Kilby and Robert Noyce invented the integrated circuit in 1958

1993
- Mobile phone (mobile device)
  - IBM launched Simon, the first smartphone with NEC V30HL 16MHz 16-bit Processor

Since 1993, the form of devices enabling computing power are further diversified

Cloud data center
Edge computing device
Smartphone
Smart watch

The form of devices enabling computing power have become more diverse.

Computing power used to be realized through manual devices such as abacus and mechanical calculator (Figure 1.2). Abacus is the earliest known calculating device and was once widely used to perform common arithmetic operations. But it is inefficient for relying solely on human brain. In 1642, French scientist Blaise Pascal invented the first mechanical calculator based on abacus principles and largely released manual input. Later, electronics and integrated circuits came into being and became new carriers of computing power. In 1937, John v. Atanasoff and Clifford Berry designed the “Atanasoff-Berry computer”, the first automatic electronic digital computer and significantly boosted calculation efficiency. As Bell Labs invented transistor (1947) and Jack Kilby and Robert Noyce invented integrated circuit (1958), carriers of computing power rapidly shrunk in size and became exponentially more competent, paving the way toward integrated circuit computers. In 1993, IBM launched the world’s first smartphone Simon, marking the beginning of wireless portable and mobile computing power. Since then, the device forms enabling computing power have been drastically diversified, from servers underpinning internet infrastructure to smartphones, PCs and wearable smart devices. Computing power is now everywhere and closely knitted in to the daily life.
1.2 Three Key Stages and Future Deployment of Computing Power

Evolution of computing power architecture

<table>
<thead>
<tr>
<th></th>
<th>Single-core</th>
<th>Multi-core</th>
<th>Networked</th>
</tr>
</thead>
<tbody>
<tr>
<td>Key Constraints</td>
<td><strong>P</strong>=Performance; <strong>C</strong>=Cost; <strong>Po</strong>=Power</td>
<td><strong>P</strong>=Performance; mismatch between processor, memory, and software architecture; <strong>Po</strong>=Power</td>
<td>Standalone devices inadequate</td>
</tr>
<tr>
<td>Technological and commercial feasibility</td>
<td>· Performance: quantum tunnelling effect</td>
<td>· Performance: mismatch between processor, memory, and software architecture</td>
<td>· “Networked” computing power</td>
</tr>
<tr>
<td>· Cost: chip beyond 3nm unaffordable for mass market</td>
<td>· Power: sharp increase in power consumption as number of cores goes up</td>
<td>· Cost: supply and demand mismatch due to bandwidth cost</td>
<td>· More edge device</td>
</tr>
<tr>
<td>Silicon-based processor reaches limit at 3nm</td>
<td>· Multi-core processor reaches limit at 128 cores</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1.3 Three key constraints drove computing power to evolve from “single-core” to “multi-core” and “networked” stages. Since integrated circuit was invented in the 1960s, the development of computing power has been constrained by the performance, cost and power consumption of silicon chips. Performance refers to the chip’s computing and data processing capacity, and is impacted by physical effects, manufacturing techniques, packaging, etc. Cost means the economic investment into the chip for its computing power, including design cost, manufacturing cost, etc. Power consumption can be measured for the chip itself or for each unit of computing power.

To break free from these three constraints, silicon chips, along with their computing power, have evolved through 3 stages: “single-core”, “multi-core” and “networked”. Computing power of a single-core chip will peak at 3nm. Moving down from 3nm would be too difficult to manufacture and too expensive to commercialize. Multi-core chips seemed to be the solution – but not a perfect one. More cores don’t mean proportionately more computing power due to the dissipation effects caused by the mismatch between processor, storage media, operating system and software. Energy consumption per additional unit of computing power will also stop to optimise when the number of cores reaches a certain level. The tipping point is at 128 cores. When the capacity of single-core and multi-core chips is exhausted, it requires “networked” computing power to fulfil the demand gap. Meanwhile, to address the tidal effects in the demand for computing power caused by the existing network limitations, flexible and low-cost edge devices may be deployed to solve the demand-supply conflict. Therefore, in the future, computing power will be ubiquitous in a “cloud-edge-device” deployment architecture, ready to meet any demand of the increasingly smart society. (as shown in Figure 1.3).
1.2.1 Computing power of single-core chips will reach its limit at 3nm\(^7\) due to performance and cost constraints

Performance constraint. Physical limits of silicon-based materials such as quantum tunnelling effect make it difficult to keep boosting the performance of single-core chips by simply making them smaller. Improving chip performance through advancing nano-process technology is an arduous challenge. As the size of the transistor decreases, its gate length also decreases. When the gate length is less than 3nm, quantum tunnelling effect (when the grid narrows to a certain extent, the source and drain will be very close. Even with no voltage applied, the source and drain are interconnected) will cause the logic circuit to fail. By then, improving the performance of single-core chips by advancing manufacturing process will be very difficult.

Cost of mobile phone SoC

> **Smartphone SoC (Qualcomm Snapdragon 855) cost (US$)**

<table>
<thead>
<tr>
<th>Grid Length</th>
<th>Cost (US$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>16nm</td>
<td>33</td>
</tr>
<tr>
<td>10nm</td>
<td>43</td>
</tr>
<tr>
<td>7nm</td>
<td>64</td>
</tr>
<tr>
<td>5nm</td>
<td>108</td>
</tr>
<tr>
<td>3nm</td>
<td>263</td>
</tr>
</tbody>
</table>

Source: IBS, Roland Berger

![Figure 1.4 Cost of 3nm flagship SoC will increase significantly.](image)

Cost constraint. Making chips at 3nm is prohibitively expensive. Enormous initial investment and unaffordable retail price mean difficulty in commercializing. According to our analysis, at 10 million units of shipment, cost of 3nm flagship SoC (e.g. Qualcomm Snapdragon 855) is US$200 higher than the 7nm variant. The mass market won’t be able to afford. The increased cost is mainly used to fund the R&D, design and manufacturing of more advanced nano-process. For example, photomask making, a key procedure in semiconductor manufacturing costs less than USD 1 million at 28nm. At 7nm, the cost is at least USD 10 million. The surge in chip cost will translate into a daunting retail price and put off most users. The small number of high-end users are not sufficient to support chip manufacturers to make

\(^7\)When describing silicon-based chips, 12nm and 7nm refers to the grid length of CPU or GPU chip. The shorter the grid length, the more transistors are integrated on one silicon chip. In order to achieve stronger single-core chip performance, chip manufacturers must continuously improve the nano-process to include more transistors in one chip.
meaningful investment. Such dilemma will eventually impede tech advancement. In 2018, UMC and Grofand, two of world’s top 5 foundries, announced their withdrawal from the race beyond 7nm.

1.2.2 **Networked computing power will be necessary when the 128-core limit is hit due to performance constraints (storage, system and software) and rising unit power consumption**

Multi-core processor has two or more processing units (i.e. cores) and supports parallel computing techniques. It is naturally suitable for tasks requiring high computing power, such as AI (image recognition, natural language processing etc.), 3D modelling, game rendering and video editing. However, to a certain point, the computing power of multi-core processor will not increase as more cores are crammed in.

**Multi-core processor computing power growth forecasts**

![Graph showing the growth of computing power with the number of cores.](image)

Source: Intel, Roland Berger

Performance constraints. We monitored the correlation between multi-core processors’ computing power and number of cores (under the current Von Neumann architecture) and found that growth multiple of computing power declines dramatically as number of cores goes up. Specifically, from 4-core to 8-core, computing power grows by 1.74x; but from 128-core to 256-core, the growth is merely 1.16x (Figure 1.4), lower than the 1.2x of threshold for being economical. Behind such computing power dissipation are constraints posed by data transmission bottleneck, operating system limitation and nature of serial tasks.

- Data transmission bottleneck: When CPU bottleneck happens, the computing power from increased number of cores won't be fully utilized. CPU bottleneck happens when data transmission speed (determined by the bus linking CPU and memory) between the chip and the memory does not
match the CPU’s computing speed, thus limiting CPU’s output. Theoretically, more cores mean more computing power. But when the memory bus is shared by more cores and responds to more concurrent requests, a mismatch between the processing capacity of processor and the data transmission capability of memory will happen.

- Operating system limitation: Mainstream operating systems such as Windows and Linux are designed to work with 8-16 cores. They are not sufficient to optimize the scheduling, arbitration, and load balancing when the processor sports more than 32 cores, rendering the computing power of the additional cores idle.

- Nature of serial tasks: Multi-core processor is designed to handle parallel tasks, not serial tasks. In serial tasks, like a typical RTS video game, the input of any latter task depends on the output of the previous one. Many serial tasks in real practice cannot be decomposed into parallel subtasks and therefore cannot harness the computing power of multi-core processors.

Power consumption constraints. “Networked” computing power is required to resolve the mismatch between growing demand for computing power and rising chip power consumption. With current chip-manufacturing technology, power consumption per unit computing power (or, power per performance) will increase along with the number of cores. For instance, the 6-cores i5-8600k and the 8-cores i7-9700k are both 14nm 3.6GHz Intel chips. But their powers per performance are 1.11 W/GFLOPS and 1.23 W/GFLOPS, respectively, a difference of 11%. However, without battery and cooling technology breakthroughs, the power capacity per device is not expected to grow at the same rate as more cores are stacked on processors. Computing power distributed in different locations could be “networked” via internet and better fulfill the increasing demand while power capacity of devices stagnates.

In practice, distributed computing can group together the numerous computers in the world and pool their computing power to accomplish higher-level computation by spreading the task between the machines.

For instance, distributed computing helped scientists in the European Organization for Nuclear Research (Conseil Européen pour la Recherche Nucléaire, CERN), the world’s largest particle physics laboratory, to prove the existence of the Higgs boson8 using the Large Hadron Collider9. CERN connected worldwide supercomputers to develop the "networked" computing pool. The pool fully utilized DCN10 (Data

---

8 The Higgs boson is an elementary particle in the Standard Model of particle physics, produced by the quantum excitation of the Higgs field, one of the fields in particle physics theory.

9 The Large Hadron Collider (LHC) is the world’s largest and most powerful particle accelerator. It first started up on 10 September 2008 and remains the latest addition to CERN’s accelerator complex. The LHC consists of a 27-kilometer ring of superconducting magnets with a number of accelerating structures to boost the energy of the particles along the way.

10 DCN (Data communication networks) transmit digital data from one computer to another computer using a variety of wired and wireless communication channels. One such network is the Internet.
Communication Network) to build the internal network of computing nodes and DCI\(^\text{11}\) (Data Centre Interconnect) to promote the interconnection between computing nodes, which are important to each device's computing power output. Data distribution follows a pyramid structure. Data was initially distributed from a zero-level site (Tier-0) at CERN to 12 primary sites worldwide (Tier-1) through a high-speed (up to 100 Gbps) network. Then data were sent to nearly 200 secondary sites (Tier-2) from primary sites. With the distributed network, the pool calculated 100 PB\(^\text{12}\) 5-year experimental data collected by scientists (equivalent to data carried on 6 million Blue-ray discs).

In addition, SETI@home\(^\text{13}\), launched on May 17, 1999, has been harnessing the idle drives of personal computers to search for extraterrestrial intelligence. The project broke up the massive data collected by radio telescopes into data packages and uploaded them into the distributed networking system. Then the system distributed those packages to those PCs with SETI@home for data analysis. Within just 5 years, the project accumulated nearly 2 million years of CPU running time, processed more than 1.3 billion data units, and performed nearly \(5 \times 10^{21}\) floating-point operations.

1.2.3 "Cloud-edge-device" ubiquitous deployment can solve the computing power demand mismatch caused by performance and cost constraints

There will still be computing demand left unfulfilled even with supplementary "networked" computing power. Many scenarios (such as Smart Security Network, CDN acceleration) require low latency and high bandwidth at low transmission cost. Edge computing is the solution for these scenarios. Edge computing means deploying computing power closer to the terminal device to achieve faster reaction, lower latency and lower transmission cost. Tom Lantzsch, Senior Vice President of Intel, said, "It is necessary to migrate data from cloud to edge as our current bandwidth and transmission speed cannot handle our increasing amount of data"\(^\text{14}\). Utilizing edge computing device could bring down the cost of computation, storage, and network by more than 30%\(^\text{15}\).

In terms of performance, when network bandwidth limits cloud computing output and network latency causes computing power mismatch, it is time for edge computing to come in. Therefore, it is an absolute need to deploy a multi-level “cloud-edge-device” system.

---

\(^{11}\) Data Centre Interconnect (DCI) technology connects two or more data centres together over short, medium or long distances using high-speed packet-optical connectivity.

\(^{12}\) Pettabyte

\(^{13}\) SETI@home ("SETI at home") is an Internet-based public volunteer computing project employing the BOINC software platform created by the Berkeley SETI Research Centre and is hosted by the Space Sciences Laboratory, at the University of California, Berkeley. Its purpose is to analyse radio signals, searching for signs of extraterrestrial intelligence, and as such is one of many activities undertaken as part of the worldwide SETI effort.

\(^{14}\) eefocus.com, With edge computing market expanding, computer power is far from limit, 2018

\(^{15}\) China Institute of Electronic Technology Standardization, White Paper on Edge Cloud Computing Technology and Standardization, 2018
Bandwidth and latency jointly determine the transmission quality of network channels, and in turn, influence the full output of network computing power infrastructure (Figure 1.6). Utilizing “networked” computing power involves two-way data transmission between terminal device and cloud, through the channels between network nodes like gateways, base stations and data centres. Under the current network technology where bandwidth is restricted by network types and transmission methods, and latency caused by distance and interference, cloud computing power output never reaches its full capacity (Figure 1.7). Edge computing device is great supplement. Demand for computing power will keep increasing in an evolving smart society where new technologies, smart applications and smart scenarios abound. However, due to network limitations and cost constraint, mismatch may occur between the “networked” computing power supply and actual demand. Deploying more flexible and cost-effective edge computing devices is a great way to facilitate the long-distance data transmission between cloud and terminal devices. For example, in the era of 5G, deploying multi-access edge computing (MEC)\(^{19}\) can ensure low latency and high bandwidth as required in certain scenarios such as smart factory, smart

\(^{16}\)For the cloud, computing power is generated by chips like GPU or NPU and its rational use can be achieved by complicated data processing on virtual platform scheduling servers.

\(^{17}\)For the edge, computing power is generated by chips like CPU or FPGA and the stability and low latency can be guaranteed by real-time data filtering and response on edge service platforms.

\(^{18}\)For the device, computing power is generated by chips like CPU, GPU or DSP. Better user experience can be achieved by resource management of software and hardware on the operation systems.

\(^{19}\)Multi-access edge computing (MEC), formerly mobile edge computing, is an ETSI-defined network architecture concept that enables cloud computing capabilities and an IT service environment at the edge of the cellular network or any network.
port and smart transportation. Edge computing can also help reduce low-value data flow (e.g. repetitive data) or hyper-scale flow (e.g. video streaming). In the smart port scenario, crane control requires ultra-low latency (within 10-20 milliseconds) and ultra-high reliability (above 99.999%). Such requirement is beyond the capability of the current 4G technology.

In terms of cost, deploying edge computing power closer to terminal devices will reduce network bandwidth cost. In 2018, Alibaba Cloud and Huya.com\(^2\) jointly created edge node service, cutting latency to less than 5ms and bandwidth cost by more than 30%.

**Overview of computing power architecture**

To sum up, facing technological and cost constraints, it is a must to supplementing “networked” computing power with edge computing. In the future, to satisfy the ever increasing demand from a smarter society, a “cloud-edge-device” deployment architecture will take shape where large scale and complex computing tasks take place on the cloud, simple tasks are computed on the edge, and cognitive interaction happen on the terminal devices. (as shown in FIGURE 1.7).

---

\(^2\)Huya.com is a video livestreaming platform covers game, entertainment, and sports topics.
1.3 The 4A\textsuperscript{21} Requirements for Computing Power by Key Technologies

Four key technologies in a smart society

In the future, computing power demand will be drastically driven up by the underpinning technologies of a smart society, including AI, IoT, blockchain, AR/VR, etc. (Figure 1.8) These technologies, as they become more mature, will help realize a large number of smart scenarios. Specifically, AI will turn autonomous driving, smart office, smart security, and smart self-services into reality; IoT will make smart firefighting, smart factory, smart shopping malls, and smart home possible; Blockchain will be used for digital certificate and information encryption; AR/VR will improve experience in smart malls, games, and smart classrooms. When these scenarios are realized, industries will leap forward with more innovation, higher digital level and more vigorous market; Governments will run more efficiently and effectively, and better serve its citizens; people will enjoy better social security and living environment; The society will achieve sustainable development and become more intelligent with collective contribution from all.

\textsuperscript{21} “4A” refers to “Anytime, Anywhere, Any Capacity, Any Object”
Requirements of key technologies for computing power

<table>
<thead>
<tr>
<th>Anytime</th>
<th>Fast response: low latency and uninterrupted support</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI:</td>
<td>Autonomous driving</td>
</tr>
<tr>
<td>IoT:</td>
<td>Smart factory</td>
</tr>
<tr>
<td>Blockchain:</td>
<td>Bitcoin</td>
</tr>
<tr>
<td>AR/VR:</td>
<td>VR games</td>
</tr>
<tr>
<td></td>
<td>· Higher level of autonomy requires better network</td>
</tr>
<tr>
<td></td>
<td>· Level 4 need</td>
</tr>
<tr>
<td></td>
<td>&gt;100Mbps bandwidth &amp; 5-10ms latency</td>
</tr>
<tr>
<td></td>
<td>· Smart factory requires continuous computing</td>
</tr>
<tr>
<td></td>
<td>power support</td>
</tr>
<tr>
<td></td>
<td>&gt;1Gbps bandwidth &amp; 5-10ms latency</td>
</tr>
<tr>
<td></td>
<td>· Blockchain requires high bandwidth and latency</td>
</tr>
<tr>
<td></td>
<td>· &gt;180Mbps of bandwidth and &lt;20ms latency to</td>
</tr>
<tr>
<td></td>
<td>guarantee smooth user experience</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Anywhere</th>
<th>Wide coverage: Free of geographical limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI:</td>
<td>Autonomous driving</td>
</tr>
<tr>
<td>IoT:</td>
<td>Smart factory</td>
</tr>
<tr>
<td>Blockchain:</td>
<td>Bitcoin</td>
</tr>
<tr>
<td>AR/VR:</td>
<td>VR games</td>
</tr>
<tr>
<td></td>
<td>· In 2040, all new cars will have</td>
</tr>
<tr>
<td></td>
<td>autonomous driving function</td>
</tr>
<tr>
<td></td>
<td>· Home, factory, power grid, etc.</td>
</tr>
<tr>
<td></td>
<td>· Finance, government, retail, payment,</td>
</tr>
<tr>
<td></td>
<td>logistics, etc.</td>
</tr>
<tr>
<td></td>
<td>· Games, studios, smart shopping malls,</td>
</tr>
<tr>
<td></td>
<td>etc.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Any Capacity</th>
<th>Flexible: Adaptive to peak and off-peak demand</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI:</td>
<td>Autonomous vehicle</td>
</tr>
<tr>
<td>IoT:</td>
<td>Smart factory</td>
</tr>
<tr>
<td>Blockchain:</td>
<td>Bitcoin</td>
</tr>
<tr>
<td>AR/VR:</td>
<td>VR games</td>
</tr>
<tr>
<td></td>
<td>· In 2030, computing power demand will</td>
</tr>
<tr>
<td></td>
<td>increase by 390x</td>
</tr>
<tr>
<td></td>
<td>· In 2030, computing power demand will</td>
</tr>
<tr>
<td></td>
<td>increase by 110x</td>
</tr>
<tr>
<td></td>
<td>· IoT devices will reach 21.5 billion units</td>
</tr>
<tr>
<td></td>
<td>in 2025</td>
</tr>
<tr>
<td></td>
<td>· In 2030, computing power demand will</td>
</tr>
<tr>
<td></td>
<td>increase 2,000x</td>
</tr>
<tr>
<td></td>
<td>· In 2030, computing power demand will</td>
</tr>
<tr>
<td></td>
<td>increase by 300x</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Any Object</th>
<th>Diverse forms: computing power realized through various devices</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI:</td>
<td>Autonomous vehicle</td>
</tr>
<tr>
<td>IoT:</td>
<td>Smart factory</td>
</tr>
<tr>
<td>Blockchain:</td>
<td>Bitcoin</td>
</tr>
<tr>
<td>AR/VR:</td>
<td>VR games</td>
</tr>
<tr>
<td></td>
<td>· Autonomous vehicle</td>
</tr>
<tr>
<td></td>
<td>· Edge computing device</td>
</tr>
<tr>
<td></td>
<td>· Cloud computing center</td>
</tr>
<tr>
<td></td>
<td>· IoT device</td>
</tr>
<tr>
<td></td>
<td>· Edge computing device</td>
</tr>
<tr>
<td></td>
<td>· Cloud computing center</td>
</tr>
<tr>
<td></td>
<td>· Cloud computing centre</td>
</tr>
<tr>
<td></td>
<td>· Cloud computing centre</td>
</tr>
<tr>
<td></td>
<td>· Cloud computing centre</td>
</tr>
<tr>
<td></td>
<td>· Cloud computing centre</td>
</tr>
<tr>
<td></td>
<td>· AR/VR device</td>
</tr>
<tr>
<td></td>
<td>· Phone, computer, etc.</td>
</tr>
<tr>
<td></td>
<td>· Cloud computing center</td>
</tr>
</tbody>
</table>

Source: IDC, Broadband Development Alliance; Roland Berger

In the future, computing power will be 4A (Anytime, Anywhere, Any Capacity and Any Object), which is driven by the technologies such as AI, IoT, blockchain and AR/VR (Figure 1.9):

- Anytime: In the future, computing power needs to guarantee low latency to allow real-time data processing, and ensure uninterrupted support. In the autonomous driving scenario, the self-driving vehicle should react in real-time to any emergency on the road. Its computing power should monitor the traffic condition and vehicle condition throughout the trip. Level 4 and Level 5 autonomous driving will require an above 100Mbps bandwidth and tolerate 5-10ms of latency.

- Anywhere: In the future, computing power will be everywhere as smart scenarios will be so common, including facial recognition, smart security, smart traffic, mobile payment, smart factory, smart agriculture, etc.

- Any Capacity: The future computing power should be able to satisfy demand of any scale by any terminal devices through adaptive deployment based on the actual dynamic demand. For instance, the computing power in autonomous driving should vary and adapt to peak and off-peak traffic periods.

- Any Object: In the future ubiquitous computing power deployment architecture, computing power will exist on any object: on the cloud (data centres of different scales), the edge (smart road piles and other edge computing units) and the device (human brain, autonomous vehicles, IoT devices, PCs and AR/VR devices).
1.3.1 Artificial Intelligence’s demand for computing power

Al’s demand for computing power

<table>
<thead>
<tr>
<th>AI has great application potential</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autonomous driving: replace driver with AI</td>
</tr>
<tr>
<td>AI Healthcare: imaging, diagnostics, predictive analytics</td>
</tr>
<tr>
<td>IDC predicts that the market value of artificial intelligence will grow to US$208.1 billion in 2025 from 2018’s US$24.9 billion</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>AI development needs computing power for support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demand for computing power is growing rapidly in each scenario</td>
</tr>
<tr>
<td>- Higher driving autonomy calls for higher computing power</td>
</tr>
<tr>
<td>Anytime, Any Object: real-time response in any environment</td>
</tr>
<tr>
<td>- Extensive coverage</td>
</tr>
<tr>
<td>- Low latency</td>
</tr>
<tr>
<td>Anywhere, Any Capacity: penetration and coverage increase</td>
</tr>
<tr>
<td>- Level 3-5 autonomous vehicles more popular</td>
</tr>
<tr>
<td>- Higher market penetration</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>AI growth demands more computing power</th>
</tr>
</thead>
<tbody>
<tr>
<td>2018: 41 EFLOPS</td>
</tr>
<tr>
<td>2020: 134 EFLOPS</td>
</tr>
<tr>
<td>2025E: 1,709 EFLOPS</td>
</tr>
<tr>
<td>2030E: 16,206 EFLOPS</td>
</tr>
<tr>
<td>~390x</td>
</tr>
</tbody>
</table>

Source: IDC, Roland Berger

Figure 1.10 Artificial intelligence’s demand for computing power

AI will empower all industries. It is estimated that, by 2025, AI will be a market worth US$208.1 billion and will be at full play in many fields such as autonomous driving, smart finance, smart healthcare, smart retail, and entertainment. As computing power of chips used in AI-powered smart scenarios goes up and AI penetrates more industries, both demand for computing power and demand for cloud computing centre, edge devices and terminal NPUs will rocket. By 2030, the demand for computing power in AI-powered fields will to reach 16,000 EFLOPS, equivalent to the combined computing power provided by 160 billion Qualcomm Snapdragon 855 chips with built-in artificial intelligence. We use the autonomous driving scenario to illustrate the computing power demand driven by AI.

Computing power demand of individual autonomous vehicle will increase significantly.

Ubiquitous Computing Power: the Cornerstone of an Intelligent Society
Position Paper

**Autonomous Vehicle's demand for computing power**

- **Very low requirement**
  - Low amount of data generated per second
  - No need for real-time support
  - Rely on human judgment

- **Low requirement**
  - Data from 1-3 720p camera and radar
  - Limited machine involvement
  - Acceleration and deceleration control

- **Medium requirement**
  - Data from about 8 HD cameras and 10+ radars
  - Partial machine driving
  - Human involvements is limited to system request

- **High requirement**
  - Data from 8 HD cameras, 20+ radars
  - Complex software framework & safety redundancy
  - Highly autonomous driving
  - Full autonomous driving under selected conditions

- **Further requirement**
  - Data similar to L4
  - Recognition, deduction and decision needs to be real-time
  - Complete autonomous driving

1) Xavier SoC is the core of the DRIVE AutoPilot system (Designed for L2 Autonomous Vehicle), provided by NVIDIA

Source: Roland Berger

Figure 1.11 Autonomous vehicle's demand for computing power grows 5,000 times from Level 2 to Level 5.

From Level 2 to Level 5, autonomous vehicles require 5,000 times more computing power to cope with the huge amount of additional data (Figure 1.11).

Level 2 autonomous vehicles only need so much computing power as to deal with the data from 1-3 720P cameras and 1-2 radars, and not in real-time.

Tesla's latest model is a Level 3 autonomous vehicle with a Full Self-Driving Computer system onboard. Thanks to the neural processing unit (NPU) built in the self-driving chip and Tesla's proprietary Autopilot 3.0 system, the vehicle has enough computing power to handle simultaneously the data generated by 8 HD cameras and 10+ radars, with which it analyzes and responds to the road condition and vehicle condition in real-time throughout the journey.

Huawei Mobile Data Centre (MDC) system can support Level 4 autonomous driving. Having the latest Ascend CPU chip, AI chip, ISP chip and SSD control chip built in, the system ensures low latency: less than 200ms end-to-end latency, less than 1ms ROS internal latency and less than 10ms kernel scheduler latency. The MDC system also can deal with the huge data flood from 16 cameras, 16 ultrasonic radars, 8 lidars and 6 millimeter-wave radars. According to Nvidia, Level 4 autonomous vehicles require 50 times more computing power than Level 3 vehicles to deal with more complex driving conditions, more sophisticated correction mechanisms, and software architecture with more layers.

At Level 5, autonomous vehicles will need to deal with harsh and complicated road conditions. Therefore, with a similar amount of data with Level 4 to process, chips onboard Level 5 vehicles need 10 times more computing power to analyze faster and make decisions accordingly in real-time.
Anytime, Any Object.

To achieve high level autonomous driving, the multi-level and multi-source driving data need to be processed at any place and with millisecond-level latency. This is where edge servers come in. If such crucial driving data as road condition and pedestrian density were processed by cloud servers located hundreds of miles away, the potential latency might cause serious consequences to a fast-moving autonomous vehicle (e.g. traffic congestion, traffic collision, etc.). Level 3 autonomous driving can tolerate a latency of 10-20ms, but Level 4/5 can do less than 10ms. Based on the current autonomous driving roadmap featuring individual intelligent vehicles and relying on internet of vehicles, such demand for low latency and high reliability will not be satisfied even in a 5G context. The solution lies in edge computing power deployment (deploying RSS\textsuperscript{23} for roadside sensing and edge computing platform for data collection, routing and distribution).

Anywhere, Any Capacity.

Computing power in the autonomous driving field will surge as high level autonomous vehicles roll out rapidly. By 2020, Level 1 and Level 2 autonomous vehicles will have a penetration rate of 40%, and Level 3, 5%. Level 3 and Level 4 autonomous vehicles will be mass-produced in 2020 and 2025, respectively. By 2040, all new cars will be capable of autonomous driving at different levels, and more than 30% of them will be have Level 4 or Level 5 autonomy. By 2040, there will be about 200 million autonomous vehicles on the road. Their 300-500 TFLOPS\textsuperscript{24} NPU chips’ computing power will dramatically drive up the overall demand for computing power.

AI-powered autonomous vehicles can eliminate manmade traffic accidents and make driving safer. By responding in real-time to all driving data, including surrounding vehicles, road condition, traffic lights, etc., and predicting traffic conditions, they can also reduce congestion on the road. There are peak traffic periods (i.e. rush hours in the morning/evening and holidays) and off-peak periods (i.e. late night and normal working hours). For example, Shanghai has 80,000\textsuperscript{25} vehicles running on the road during peak periods, 1.6 times\textsuperscript{26} that of off-peak periods. Cloud and edge computing power should be able to adapt to the dynamic traffic conditions.

\textsuperscript{23}Road side server is used for data gathering and edge computing

\textsuperscript{24}TeraFLOPS

\textsuperscript{25}Roland Berger estimation

\textsuperscript{26}Roland Berger estimation
1.3.2 IoT’s demand for computing power

IoT’s demand for computing power

The IoT will extend to smart factories, smart homes and many more scenarios. ZION estimated that the IoT market will worth US$158 billion\(^\text{27}\) by 2025. The popularization of IoT devices, and the increase of cloud computing and edge computing units will all demand for more computing power. By 2030, such demand is estimated to reach 8,500 EFLOPS, equivalent to the computing power provided by 7.9 billion AMD EPYC 7401 chips used in today’s high-end edge devices. We use smart factory scenario to illustrate IoT’s demand for computing power.

Anywhere, Any Object.

21.5 billion IoT devices by 2025 will call for rapid growth of computing power. Among them, 13.7 billion will be IIoT\(^\text{28}\) devices like the ones in a smart factory (infrared sensors, photoelectric sensors, ultrasonic sensors, etc.), the rest will be distributed in smart homes, smart grids, smart cities, etc. All of them need to source computing power from IoT chips like Huawei Boudica 150\(^\text{29}\).


\(^{28}\) Industrial Internet of Things

\(^{29}\) Huawei’s new highly integrated V150 NB-IoT SoC which can reduce development cost and shorten development cycle for IoT device
Smart factory’s demand for computing power

Any Capacity.
In a smart factory, high performance cloud and edge servers (Figure 1.13) are indispensible to process the enormous amount of data. Every single day, the hundreds of thousands of IoT devices in a smart factory will produce hundreds of gigabytes of data, including but not limited to machine names, material IDs, program names, standard parameters, production statistics, pass/defect rates, machine status, indicator lights and alarm codes. Strong computing power able to process the large magnitude of data benefits a smart factory mainly in two ways: first, real-time data portraits real-time operation status, and enables factory managers to detect and resolve any emerging issues; second, data across all links of production line, from raw material sourcing, warehousing, manufacturing, assembly, installation to shipment will be efficiently analyzed on the cloud, making automated, predictable and smart production possible.

Anytime.
Smart factory has high requirements for network latency and bandwidth to ensure real-time data uploading. Transmission latency in any link may end up shutting down the entire product line, impacting production efficiency and putting operators’ lives at risk. Both the magnitude of data (generated by tens of thousands of devices) and the frequency of data transmission in the smart factory call for higher bandwidth: local area network bandwidth in a smart factory needs to reach at least Gbps level, and the maximum latency should be within 5-10ms.
1.3.3 Blockchain’s demand for computing power

Blockchain’s demand for computing power

Blockchain has great application potential

- Digital currency
- Supply chain finance
- Information sharing
- Copyright protection
- Cross-border payment
- Logistics chain

IDC predicts that the blockchain-related market will grow to $38.5 billion by 2025

Blockchain development needs computing power

- **Any Capacity, Any Object**
  - Proof of Work (PoW) security mechanism
  - More computing power is needed to prevent the “51% attack”
- **Anytime, Anywhere**
  - Blockchain will be used in increasingly more scenarios and drive up the demand for computing power

Blockchain growth demands more computing power

In 2030, Blockchain-related demand for computing power will equal to the combined computing power of 1.3 billion units of Antminer V9.

Source: Roland Berger

Figure 1.14 Blockchain’s demand for computing power

Blockchain will empower many industries, particularly digital currency, supply chain finance, information sharing, copyright protection, cross-border payment, and logistics. By 2025\(^{30}\), blockchain will be a market worth US$38.5 billion. In numerous scenarios of these industries, blockchain technology will be employed to enhance data security. The resulted computing power demand will mainly be supported by cloud servers. By 2030, blockchain-related computing power demand is expected to reach ~5,500 EFLOPS, equivalent to the amount of computing power provided by 1.3 billion Antminer V9. We use the bitcoin network as an example to illustrate blockchain’s demand for computing power.

Any Capacity and Any Object.

In the future, the pool of computing power (i.e. the computing power sustaining the bitcoin system) should keep expanding to ensure the security of the blockchain system. Blockchain solves the problem of double-spending\(^{31}\), because at its core, blockchain is an open, decentralized ledger where transaction data, once recorded, cannot be easily faked or changed. Transactions of digital assets, i.e. the so-called

\(^{30}\) IDC, Worldwide Semiannual Blockchain Spending Guide, 2019

\(^{31}\) Double-spending is a potential flaw in a digital currency scheme when the same single digital token is spent more than once
"cryptocurrency" like the bitcoin, can be recorded in blockchain. In the bitcoin network, participants compete to produce new blocks to the chain with Proof-of-Work (PoW). PoW solves the problem of determining representation in majority decision-making. Under the competitive bookkeeping mechanism, the majority decision is represented by the one who has the greatest PoW effort. In an extreme case, a participant on the blockchain controlling 51% mining power will be able to intentionally exclude or modify the ordering of transaction, or reverse transactions, leading to a double-spending problem. Prevention of such security risk requires computing power supported by cloud computing centre to keep rising.

Anytime, Anywhere.

In the future, blockchain, along with cryptography, the mechanism that maintains the security of the blockchain network, will create a new basis of trust in a smart society. Their application will go beyond cryptocurrency, and expand to many scenarios in industries like general finance, government, retail, payment and logistics.

Here are six examples:

1) Information sharing. Participants of a blockchain-based information platform will always be updated in the same way, just as each node on a blockchain is always synchronized with others and contains a full copy of the transaction history of the blockchain. (e.g. Tencent’s finding-missing-people blockchain initiative).

2) Copyright protection. Blockchain technology can be harnessed to protect copyrights and enhance the credibility of legal evidence (e.g. electronic ID card, timestamp protection, and Guangzhou Arbitration Chain project).

3) Logistics chain. Blockchain will ensure full traceability of imported goods (e.g. tracing the imported goods purchased from e-commerce platforms like Alibaba, JD, and Tencent).

4) Supply chain finance. Replace traditional paper documentation (warehouse receipts, contracts, bills, etc.) with blockchain-powered system to enhance data reliability.

5) Cross-border payment. Set up cross-border blockchain ledgers to reduce payment procedures, shorten settlement cycle and cut settlement cost (e.g. Ripple, Circle, etc.).

6) Asset digitalization. Use blockchain to digitalize securitized assets to facilitate asset separation and transaction (e.g. Tencent’s WeGold application).

Realizing the above scenarios with blockchain technology require massive computing power from cloud computing centres.
1.3.4 AR/VR’s demand for computing power

AR/VR’s demand for computing power

AR/VR technology will fully empower the gaming industry and the commercial sector. By 2025, it is estimated that the AR/VR market will worth US$59.6 billion\(^{32}\). Computing power deployed to cloud computing centres and edge computing units will need to keep up to support the wide application of more computing power-intensive VR/AR devices in the future. By 2030, AR/VR applications will need a computing power of around 3,900 EFLOPS, equivalent to the amount of computing power provided by 2.1 billion SONY’s top PS4 consoles. We use VR games as an example to illustrate AR/VR application’s demand for computing power.

Individual device demands higher computing power.

In the future gaming industry, heavy rendering tasks, advanced special effects, and low latency requirements will press the computing power of GPU (the device), the cloud and the edge to grow quickly.

\(^{32}\) Marketsandmarkets, Virtual Reality Market by Offering, Technology, Device Type, Application and Geography - Global Forecast to 2024, 2019
In terms of rendering, with 4K and 8K display technology readily available, the render target resolution of future VR devices will increase by 4 to 16 times (from 1920*1080 to 3840*2160 or 7680*4320). Not only that VR devices will need more computing power to support higher-resolution display, computing power on the cloud and the edge shall also keep up. In terms of special effects, advanced special effects algorithms like dual-channel rendering, lens correction, complex interaction and physical effects will work properly only on VR devices with sufficient computing power. In terms of latency, given that VR headsets display images on screens placed very close to the eyes, the images should be of high quality parameters: latency below 20ms, refresh rate above 75Hz, and gyroscope refresh rate above 1K. For instance, to achieve the basic 1080P rendering resolution, NVIDIA’s recommended GeForce GTX 970 GPU\textsuperscript{33} for its GTX GeForce VR Ready Program has a computing power of 2.44 TFLOPS. According to NVIDIA, if VR devices were to meet all the above requirements, performance of GPU should exceed 100 TFLOPS, i.e. 40 times as powerful as its current level.

Anywhere, Any Capacity, Any Object.

Computing power will increase as VR/AR devices are widely used. High definition VR headsets like Oculus and Hololens and high-quality applications will enable VR/AR devices to rapidly penetrate multiple industries, such as business, healthcare, education and entertainment. IDC predicted that global shipments of VR devices will exceed 7 million units in 2019, and is expected to hit 36.7 million by 2023 with a CAGR of above 51.3%. In the future, as 5G networks becomes the mainstream, on-demand VR content streaming will be as common as today’s video streaming. This requires computing power deployed on the cloud and the edge to support video encoding, stitching and rendering.

Anytime.

VR depends on high bandwidth and low latency to run efficiently. Specifically, to achieve a 3D immersive panorama experience, VR/AR devices need a minimum of 180Mbps of bandwidth to power higher retinal resolution, wider viewing angle and more accurate position tracking (3D audio tracking, 3D hand gesture tracking, etc.). In addition, the motion-to-photon latency should be lower than 20ms to avoid dizziness.

\textsuperscript{33} NVIDIA’s GTX Geforce VR Ready project
## 2.1 Introduction to the Computing Power Evaluation Model

- The Model calculates the computing power of typical carriers at each of the cloud, the edge and the device level, factors in the dissipation effects and impact of network limitations, and adjust the results with applicable coefficient.

## 2.2 Findings of the Computing Power Evaluation Model

- There is a strong positive correlation between a country’s per capita computing power and its economic power. Therefore, it serves as a proxy of a country’s digital competitiveness and forms the foundation for future digital economy growth.

## 2.3 Demand for Computing Power Will Grow Exponentially as Smart Society Evolves

- At the present stage, i.e. the early stage of smart society, per capita computing power of major countries/regions is below 3,000 GFLOPS. Threshold for the development stage is 10,000 GFLOPS and 29,000 GFLOPS for the mature stage.
- As the smart society evolves, the structure of the ubiquitous computing power deployment architecture will also change: more computing power will concentrate on the cloud and the edge.
2.1 Introduction to Computing Power Evaluation Model

Computing Power Evaluation Model:

Since computing power is physically carried by the three-level cloud-edge-device architecture, the Computing Power Evaluation Model 1) calculates the computing power of typical carriers at each level, 2) factors in the dissipation effects and impact of network technology limitations, and 3) adjusts the results with applicable coefficient (Figure 2.2).

Dissipation effects refer to the fact that the overall computing power output of a multi-core processor is less than the sum of single-core performances. The main reason is that some computing power will be lost when shared resources such as L3 cache, memory and QPI bus are spread thinner between more cores. Network technology limitations means that while the cloud centre is more than capable of processing massive amount of data, limited bandwidth and high latency can become the bottleneck in data transmission. Per capita computing power can be obtained by dividing the overall computing power of a country by its population.

2.1.1 Cloud computing power is mainly carried by supercomputers and computing centre servers.
Computing Power Capacity of a Country

<table>
<thead>
<tr>
<th>Supercomputer</th>
<th>Computing power of supercomputer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Aggregate computing power of TOP500 supercomputers in the country 1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Computing power of computing center</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU chip penetration rate × number of servers × CPU computing power of typical server</td>
</tr>
<tr>
<td>GPU chip penetration rate × number of servers × computing power of typical GPU chip</td>
</tr>
<tr>
<td>FPGA chip penetration rate × number of servers × computing power of typical FPGA chip</td>
</tr>
<tr>
<td>ASIC chip penetration rate × number of servers × computing power of typical ASIC chip</td>
</tr>
</tbody>
</table>

Adjustment Coefficient

| Computing power effective utilization rate 90% (lost to dissipation effects is assumed at 10%) |
| Network efficiency 80% (lost to network bandwidth and latency limitations is assumed at 20%) |

Weighted average method is used to calculate computing power of server chips.
Four major types of chips used in computing centers are used in the calculation, namely CPU, GPU, FPGA and ASIC.

1 Supercomputers included in the global TOP500 Supercomputer List
2 Source: Gartner, Microway, Nvidia, Itsco, Desktop research, Roland Berger

Figure 2.3 Cloud computing power measurement

1) Computing power of supercomputers: Aggregated computing power of TOP500 supercomputers in the country * computing power dissipation coefficient * network efficiency.

Total computing power of TOP500 supercomputers is used as a proxy of a country’s overall supercomputer computing power; computing power dissipation coefficient is assumed at 10%; network efficiency is assumed at 80% (computing power output lost 20%).

2) Computing power of computing centres: (CPU chip penetration rate × number of X86 servers in the country × typical CPU computing power of X86 servers + CPU chip penetration rate × number of ARM servers in the country × typical CPU computing power of ARM servers + CPU chip penetration rate × number of other RISC architecture servers in the country × typical CPU computing power of other RISC architecture servers + CPU chip penetration rate × number of z system servers in the country × typical CPU computing power of z system servers + CPU chip penetration rate × number of other architecture servers in the country × typical CPU computing power of other architecture servers + GPU chip penetration rate × numbers of server in the country × typical GPU computing power + FPGA chip penetration rate × number of servers in the country × typical FPGA computing power + ASIC chip penetration rate × number of servers in the country × typical ASIC chip penetration rate * CPU chip penetration rate * GPU chip penetration rate * FPGA chip penetration rate * ASIC chip penetration rate)
penetration rate × number of servers in the country × typical ASIC computing power) × computing power dissipation coefficient × network efficiency.

There are 440,000 computing centres worldwide (including server rooms of small and medium enterprises), with about 55 million installed servers\(^{36}\). We use historical shipment data of x86 servers from IDC\(^{37}\) to estimate existing number of x86 servers in each country. Furthermore, we leverage the market share of each architecture to estimate the volume of other servers. Current computing centre servers mainly use four types of chips, namely CPU, GPU, FPGA and ASIC. We estimate market penetration rate of each type based on their respective market shares and average prices. Typical computing power of the four types of chips are used to estimate the overall computing power of the computing centre servers. Dissipation rate and network efficiency are also assumed at 10% and 80%, respectively. Average computing power of each type of chips is estimated by considering the computing powers of chips in different price ranges and their penetration rates. For example:

1) Computing power of CPU chip of computing centres is calculated by first selecting chips of different architecture. Among the four chip types, the CPU chip have a market penetration rate of 100%\(^{38}\) as it is installed in all computing centre servers. In addition, we categorize X86 chips into different price ranges before calculating their computing power (see Figure 2.3.). Specifically, we select the typical CPU chip types in each architecture (Intel Xeon E5, IBM Z14, Huawei Kunpeng 920 etc.) based on the worldwide server inventory\(^{40}\), measure the computing power of each selected model, and estimate the overall computing power of CPU chips in computing centre.

\(^{36}\) Zhongtai Securities, 2018  
\(^{37}\) IDC, Worldwide Server Market Revenue Declined 11.6% Year Over Year in the Second Quarter of 2019, According to IDC, 2019  
\(^{38}\) Each Cloud data center server includes at least one CPU. 25% of server includes a GPU and so on.  
\(^{39}\) AMR, Artificial Intelligence Chip Market by Chip Type, Application, Technology, and Industry vertical - Global Opportunity Analysis and Industry Forecast, 2018-2025, 2019  
\(^{40}\) IDC, Worldwide Server Market Revenue Increased 12.6% Year Over Year in the Fourth Quarter of 2018, According to IDC, 2019
2) GPU, FPGA, and ASIC chips in the computing centre are emerging in recent years as they are more suitable for the large-scale neural-network parallel computing tasks in AI. Therefore, they are not yet installed in many computing centre servers. The penetration rates of GPU, FPGA, and ASIC chips are 41.7.8%, 20.4%, and 1.6% respectively. At the same time, performance of different chips of the same types is similar, so there is no need to categorize them in different price groups. In the calculation, we selected Nvidia Tesla P4 GPU, Intel Stratix 10 GX/SX FPGA and HiSilicon Ascend 910 as samples.

2.1.2 Edge computing power is mainly carried by 5G+MEC, CDNs, router servers, intelligent monitoring edge servers, and IIoT edge servers.
To estimate the edge computing power, it is first necessary to understand the location and deployment of the edge computing devices in the network (Figure 2.5). Edge computing is applied in various intelligent application scenarios, such as smart home, smart healthcare, smart factory and security surveillance, smart energy, games and entertainment (Figure 2.5). Here are three examples:

1) Industrial Internet of Things (IIoT): Edge computing and cloud computing work together. Smart devices installed and connected in the edge computing environment can process critical data and respond in real-time with nearly zero latency.

2) Smart home: Smart gateway and smart device can act as edge computing nodes to process massive heterogeneous data and then upload all processed data to the cloud, enabling users to control all smart home terminal devices.

3) Smart traffic: Cloud-edge collaboration enables all-dimensional V2V and V2I (Vehicle to Vehicle, Vehicle to Infrastructure) information exchange to improve traffic safety and enhance efficiency. Current technologies such as MEC and edge cloud rely on linkage between network side and field side. Therefore, calculation of edge computing power should be done by looking at different types of edge computing solutions (Figure 2.4). On the network side, edge servers are mainly loaded through CDN server nodes and base stations. On the field side, computing power output is mainly carried out through smart gateways and smart devices (considering that smart devices interact directly with human users, their computing power is calculated in the terminal device side).
Ubiquitous Computing Power: the Cornerstone of an Intelligent Society
Position Paper

Figure 2.6 Edge computing power calculation

Edge computing solutions mainly include 5G+MEC, CDN, smart gateway, intelligent monitoring, and IIoT. Their computing power is calculated in the following logic: (Figure 2.6)

1) Edge computing power of 5G+MEC: The deployment of 5G would significantly stimulate and enrich the application of edge computing, however, as the development is still at an early stage at the moment, we would calculation the computing power as zero for now.

2) Edge computing power of CDN: Number of CDN nodes × average number of servers per node × average server computing power × network efficiency

3) Edge computing power of smart router: Number of routers × proportion of smart routers × average router computing power × network efficiency

4) Edge computing power of intelligent monitoring: Number of monitor nodes × average number of servers per node × average server computing power × network efficiency

5) Edge computing power of IIoT Edge Server: Number of IIoT production line × average number of servers per line × average server computing power × network efficiency

Chips selected in calculating computing power on the edge:
- We selected AMD EPYC 7401, Dlink OTT Gateway / Wireless Router, Hisilicon Hikey 970, Intel Xeon Phi 5120D as representatives chips used in CDN, smart router, intelligent monitoring node, IIoT edge server respectively

Source: Desktop research, Roland Berger

Network efficiency 95% (lost to network bandwidth and latency limitations is assumed at 5%)

Adjustment Coefficient

Computing Power Capacity

5G+MEC

CDN

Smart router

Intelligent Monitoring Node

IIoT Edge Server

Assume as zero since only begin to deploy

Computing power of CDN

Number of CDN nodes × Number of servers per node × Computing power per server

Computing power of smart router

Number of routers × Smart router penetration rate × Computing power per router

Computing power of Intelligent monitoring node

Number of monitoring nodes × Number of server per node × Computing power per server

Computing power of IIoT Edge Server

Number of IIoT production line × Number of server per line × Computing power per server

Figure 2.6 Edge computing power calculation
Computing power of CDNs: There are 1.2 million servers deployed on CDN nodes worldwide. Research shows that penetration rate of edge solution in CDN is about 100% and each CDN node has 110 servers. We estimate the number of CDN servers of each country based on geographical breakdown of CDN nodes by the world's top 25 CDN providers, as well as proportion of digital industry in each country. We selected AMD EPYC 7401 as the sample of CDN server chip and assume the network loss rate at 5% (network efficiency 95%).

Computing power of smart routers: There are 9.3 million routers worldwide. With development of AI, big data, cloud computing and IoT, more smart routers start to adopt edge computing solutions. Smart routers account for about 15% of the total routers. We selected Dlink OTT Gateway as sample and assume the network loss rate at 5% (network efficiency 95%).

Computing power of intelligent monitoring node: There are 300.5 million surveillance cameras worldwide, while on average 400 cameras would be control by one monitor node with 4 edge servers. We estimate the number of intelligent monitoring servers of each country based on the proportion of digital industry in each country. We selected Hisilicon Hikey 970 as sample and assume the network loss rate at 5% (network efficiency 95%).

Computing power of IIoT: Per CAICT, global IIoT market has reached 3.27 billion USD, and, for instance, while it is estimated that on average develop one production line with IIoT technology would cost 1.5 million dollars and each line would have one edge server. We estimate the number of IIoT edge servers of each country based on the proportion of digital industry in each country. We selected Intel Xeon Phi 5120D as sample and assume the network loss rate at 5% (network efficiency 95%).

---

42 GlobalDots, Content Delivery Network Explained, 2019
43 CDN expert seminar
44 GlobalDots, Content Delivery Network Companies, 2019
45 Euromonitor
46 IDC, IDC Trackers Show Moderate Growth in Worldwide Ethernet Switch and Router Markets in Q2 2019, 2019
47 TPLINK expert seminar
48 Euromonitor
49 Euromonitor
50 CAICT, IIoT Platform White Paper, 2019
51 Euromonitor
2.1.3 **Computing power on the device level resides in such smart devices as mobile phones, tablets, laptops, PCs, smart speakers, wearable devices, smart TVs, smart set-top box, smart in-car unit, drones, game consoles, mining machines, and smart surveillance camera. The overall calculation logic is as shown in Figure 2.7:**

Computing power of smart devices: Device ownership in each country $\times$ market share of each chip manufacturer $\times$ (high-end chip computing power $\times$ high-end chip market share + mid-end chip computing power $\times$ mid-end chip market share + low-end chip computing power $\times$ low-end chip market share)

---

**Figure 2.7 Calculation of device computing power**

Since terminal device interacts directly with the end user directly, it is not affected by network limitation. So we assume no adjustments are needed for terminal device computing power capacity.

**Calculation of device chip computing power (take smartphone SoC as an example)**
- Use weighted average method based on market share breakdown of major SoC suppliers, average computing power of high-/medium-/low-end SoC offered by each SoC.

Source: Roland Berger
When calculating the computing power for different types of smart devices, we rely on the follow information to work out the weighted average results: number of smart devices in each country\(^2\), market share of different supplier of smart device chips, high/medium/low-end chips from each selected supplier and computing power of typical chips.

Take smartphone as an example, the mainstream smartphone SoC manufacturers in the global market include Qualcomm, Apple, Mediatek, Samsung and Huawei. To calculate the computing power of Samsung SoC, we select Exynos 9825, Exynos 7885 and Exynos 7270 as samples for Samsung’s high-end, mid-range and low-end SoC, respectively. Based on inputs from experts on mobile phone SoC chips\(^3\), we assume that the market share high-end, mid-end and low-end SoC to be of 20%, 20% and 40%, respectively. Crunching the above statistics leads us to the overall computing power of Samsung mobile SoC. With the results for each individual supplier, and their respective market share, the weighted average computing power of smartphone will be clear. Other smart devices such as wearable devices, drones and mining machines, are estimated by the same method. In terms of adjustment coefficients, since the smart devices directly interact with end users, its power output is subject to only minimal constraints. Therefore, we assume computing power lost rate to be zero. (Figure 2.8)

\(^2\) Euromonitor

\(^3\) Expert interviews with chip manufactures
2.2 Results of Computing Power Measurement

Ranking of countries by total computing power [Unit: $10^9$ GFLOPS]

For this study, we chose a variety of countries in the world to calculate their computing power level. From total computing power perspective, USA and China are significantly ahead of other countries, ranking top 2. (Figure 2.9)

Ranking of countries by per capita computing power [Unit: GFLOPS]

Comments

- Strong positive correlation between computing power and economic power
- Computing power is highly relevant to the size of innovation-intensive economy (digital, IoT, etc.)

Countries with high per capita computing power (more than 1,000 GFLOPS) are mainly developed countries such as USA, Germany, the UK and Japan.

Countries with medium per capita computing power (460~1,000 GFLOPS) include Spain, Chile, Italy, and China.

Countries with low per capita computing power (below 460 GFLOPS) include Brazil, South Africa, etc.

Source: Desktop research, Roland Berger

Figure 2.9 Ranking of countries by total computing power

Figure 2.10 Ranking of countries by per capita computing power
Categorized by their per capita computing power, the 27 countries measured in our model belong to three categories. (Figure 2.10)

- Countries with high per capita computing power (more than 1,000 GFLOPS) are mainly developed countries such as USA, German, the UK and Japan. This is consistent with the TOP500 supercomputer list. According to the latest list published in June 2019, the US accounts for 37.1% of total computing power of the world’s TOP500 supercomputers, ranking the first among all countries. Its “Summit” and “Sierra” supercomputers top the list with 148.6 PFLOPS and 94.6 PFLOPS, respectively.

- Countries with medium per capita computing power (460~1,000 GFLOPS) include Spain, Chile, Italy, and China. Among the four, China stands out in terms of supercomputing power. In 2019, China houses 228 of the world’s TOP500 supercomputers, carrying 23.2% of the total supercomputing power.

- Countries with low per capita computing power (below 460 GFLOPS) include Russia, Brazil, Columbia, and South Africa. These countries have less than 10 of the world’s TOP500 supercomputers.

Countries evaluated in the model fall into three levels of per capital computing power: the high, the medium and the low. Countries with high per capita computing power also lead in economic development and cutting-edge industries such as digital industries and IoT platforms. For example, the United States has high computing power; it is also a leader in terms of economic power and digital economy. In 2018, the US per capita GDP and per capita income were USD 62,641 and USD 63,390, respectively; the share of digital economy per thousand of population was USD 5,685; Its IoT market was worth USD 27bn. In contrast, the Philippines, while falling into the low per capita computing power group, it also lags in terms of total economy size, scale of digital economy and IoT industry. In 2018, its per capita GDP was $3,103 and per capita income $10,720; the share of digital economy per thousand of population was c. USD1; and its IoT market was worth only USD 190mm.

---

54 Euromonitor
55 World Bank, GDP per capita (current LCU) – the United States, 2018
56 Statista, Internet of Things in the U.S. - Statistics & Facts, 2019
57 Euromonitor
58 Frost & Sullivan
Per capital computing power versus World Digital Competitiveness

Moreover, there is a highly positive correlation between a country’s per capita computing power and its ranking in the IMD World Digital Competitiveness Ranking. The IMD\(^59\) (International Institute for Management Development) is a world-renowned research institution in Lausanne, Switzerland. Since 2017, it started to publish the IMD World Digital Competitiveness Ranking to measure the capacity and readiness of 63 economies to adopt and explore digital technologies as a key driver for economic transformation in business, government and wider society. The Ranking is based on 50 criteria. The 50 criteria are grouped in nine sub-factors that, in turn, are classified in three factors: Knowledge, Technology, and Future-readiness. Comparing the IMD World Digital Competitiveness Ranking and the per capita computing power ranking shows a highly positive correlation: Countries with high digital competitiveness such as the US, Singapore and the UK also have a high per capita computing power; countries with low digital competitiveness such as Thailand and Indonesia also have below per capita computing power (Figure 2.11).

Clearly, per capita computing power is an effective proxy for a country’s economic development and comprehensive national power. Computing power, a strong indicator of a country’s digital competitiveness, is also the foundation for developing digital economy and digital society.

\(^{59}\) IMD is a Swiss scientific research institution with worldwide authority. It established the international digital competitiveness evaluation system in 1980. Since then, it has published the World Digital Competitiveness Report annually to assess the level of digitalization of 63 economies in the world. Intelligence level refers to the ability of an economy to develop and utilize digital technology for governmental, commercial and social transformation. According to IMD, the measurement of intelligence level includes 3 primary indicators, 9 secondary indicators, and 50 tertiary indicators. The primary indicators include Knowledge, Technology, and Future-readiness.
2.3 Demand for Computing Power Will Grow Exponentially as Smart Society Evolves

Advanced technologies promote smart society development and demands high chip performance and network capacity. We predicted the thresholds of computing power for different stages of the future smart society using the Computing Power Evaluation Model. The conclusion is that per capita computing power will surpass 10,000 GFLOPS in developing stage and 29,000 GFLOPS in mature stage.

2.3.1 Computing power demand in different stages of a smart society

**Level of digital applications penetration in various industries**

<table>
<thead>
<tr>
<th>Stage Definition</th>
<th>Product concept smart verification</th>
<th>Product concept smart decision</th>
<th>Product idea active generation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source: Accenture, Boston Consulting, International Robotics Federation, McKinsey, Signify Research, Synergy Research, Desktop research, Roland Berger</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Product concept smart verification**
- Smart computing technology can only be used for short and incomplete implementation of R&D ideas in the manufacturing industry

**Product concept smart decision**
- Smart system can initiate a new prototype and provide decision suggestions for the development of product concepts based on algorithm

**Product idea active generation**
- Smart R&D system can actively generate R&D plans for the reference of the enterprises

Figure 2.12 Different stages of a smart society\(^{60}\) and their computing power demand to support smart scenarios\(^{61}\)

\(^{60}\) According to world-renown research institution, Gartner, the industry will gradually step into the developing stage, and then the mature stage from the early stage with the improvement of smart penetration. The technologies, methods, infrastructures and ecosystems in early stage start to develop and have a high growth rate. The uncertainty of technologies is also high and the penetration usually exceeds 1%. The technologies in developing stage get verified and the values can be predicted in most scenarios. Profitability grows well and industry’s penetration reaches 1%-5%. The technologies in the mature stage get verified; the industry value proposition get well understood, and technologies get large-scale commercial applications, where the penetration usually exceeds 20%. It takes more than 10 years to evolve from early stage to mature stage. Considering most countries are stepping into the developing/mature stage, their smart penetrations are higher than that (5% and 20%) defined by Gartner: the average penetration in early stage is 8% and 30% in mature stage. Major countries will step into the developing stage in 2035 and into the mature stage in 2045.

\(^{61}\) Smart penetration in early/developing/mature stage can be estimated by selecting typical scenarios in major industries, for example Robo-adviser in financial industry, Industry Robot in manufacturing, Auto-driving in transportation and logistics industry, and smart report reading in the medical industry, etc.
Computing power level marks different stages of smart society and significantly improves along with the smart society’s evolution. The development of computing power is influenced by three factors: industry coverage rate, industry penetration rate and natural growth of computing power performance.

- Industry coverage rate: Coverage of smart scenarios, i.e. scenarios enabled by advanced technologies like AI, IoT, big data, blockchain and AR/VR across different industries (for example, coverage of smart driving in transportation industry) is calculated as “number of industries using smart applications in the industry / total number of industries”.

- Industry penetration rate: Penetration of smart devices in an industry (e.g. penetration of smart driving devices in the driving device market) is calculated as “shipment of smart devices used in smart scenarios/ total shipment of the industry”.

- Natural growth of computing power performance: Moore’s Law still holds though it has slowed down, meaning that computing power per unit of cost will naturally grow as technology advances. Such natural growth is factored in our measurement. Without major technology breakthrough, Moore’s Law will end in 10 years and the natural growth of computing power performance shall not be considered anymore after 2029.

---

62 The penetration of smart devices in a particular industry is calculated as “smart industry production value/ total production value”.

63 It can be predicted that the chip calculation performance per unit cost shall decrease with the smart society development, considering the existence of Moore’s Law and the probability to its slowing down in the following 10 years (2019 to 2029). Take Intel i7 processor for desktop computers as an example, it is found that the natural growth factor decreased along time and the growth multiples and years are negatively related by comparing the computing power per unit costs of the top chips of the year from 2009 to 2019. Without major technology breakthrough after 2029, Moore’s Law won’t be considered anymore, that is, the growth multiple is always 1.
Levels of computing power for Three Stages of the Intelligent Society

<table>
<thead>
<tr>
<th>Computing power level at each stages [GFLOPS per capita]</th>
<th>Description of each stage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early Stage</td>
<td>Developing Stage: around 2035 with major countries/regions’ smart penetration is about 8%. At this stage, computing power is about 10,000 GFLOPS per capita</td>
</tr>
<tr>
<td>Developing Stage: around 2045 with major countries/regions’ smart penetration is about 30%. At this stage, computing power is about 29,000 GFLOPS per capita</td>
<td></td>
</tr>
<tr>
<td>Mature Stage: around 2045 with major countries/regions’ smart penetration is about 30%. At this stage, computing power is about 29,000 GFLOPS per capita</td>
<td></td>
</tr>
</tbody>
</table>

1) From purely penetration rate perspective, 5% and 20% are Gartner’s points of separating 3 stages. However, after taking into account the differences of different countries’ economic development, we used a 5 year delay and therefore used 8% and 30% as the milestones of two stages for the most countries/regions.

2) Moore’s Law has already began to slow down, without major technological breakthrough, it is expected to begin fail around 2029.

Based on the penetration rate of different development stages of the industry, smart society can be divided into three stages: the early stage, the developing stage and mature stage (Figure 2.13).

1) Early Stage: According to computing power model, all countries in the world are now in the early stage with per capita computing power lower than 3,000 GFLOPS. Particularly, the per capita computing power in the US is 2,522 GFLOPS and 2,132 GFLOPS in Singapore (Figure 2.12). Smart scenario coverage in all industries is low (below 8% in some industries) and features dotted distribution. At this stage, application of advanced technologies based on smart computing like AI just started. However, the overall digital competitiveness remains low due to limited investment in smart technologies, complicated industry chain and unbalanced mechanization level of different links of industry chain. In this era of 4G, the peak transmission rate of the mobile network bandwidth reaches 150 Mbps; the network latency is about 50ms; 100,000 devices can be well connected to 4G per square kilometer. Deep into the early stage when 5G becomes the mainstream, network bandwidth will jump to 10 Gbps; latency will drop below 10ms, fully satisfying the demanding latency requirement of VR

---

64 Mainly the mobile networks with higher differences in network intergenerational evolution.

65 The broadband bandwidth is about 500 Mbps in the early period of the early stage.

66 There is no obvious difference between broadband bandwidth and mobile network bandwidth. Therefore, no distinction is made...
interaction (10-30ms) and smart home (20-40ms)\textsuperscript{67}; 1 million 5G devices can be well connected to 5G per square kilometer (including NB-IoT scenarios, largely all less power-intensive IoT devices can be covered).

2) Developing Stage: Per capita computing power of 10,000 GFLOPS is the milestone to this stage. Coverage of smart scenarios will reach 8~30% in most industries like smart manufacturing, smart driving and smart government. At this stage, chip performance will not be limited by its integration level. AI and other advanced technologies based on smart computing will prevail across industries. Smart decision-making, AI-enabled smart cloud service and smart robot will be commonplace. In the developing stage, technology development will see major breakthroughs. 6G will appear and gradually mature. Network bandwidth will increase up to 100Gbps to 1Tbps\textsuperscript{68}; latency will plummet below 0.1ms, fully enabling the end-to-end transmission of vehicle control signals\textsuperscript{69} and immersive VR/AR experience (latency lower than 8ms)\textsuperscript{70}; up to 100 million 6G devices can be well connected per square kilometre. Meanwhile, progress in network technology will help realize new smart scenarios (3D real-time rendering, smart R&D test, and high-level autonomous driving, etc.), which will shorten production cycle, improve productivity, lower labour cost, make production safer and more environment-friendly, and eventually fully upgrade industrial ecosystem and business model.

3) Mature stage: Per capita computing power of 29,000 GFLOPS is the milestone to this stage. Smart scenarios will cover all industries with an above 30% penetration rate. Penetration rate in certain industry like transportation will even surpass 90%. Network technology more advanced than 6G will be commercialized; bandwidth will be well above 1Tbps; latency will be negligible; and there will be no major limitation to device density per square kilometre. In the mature stage, quantum computing chip will be used; advanced technologies like brain-machine interface and holographic projection will mature; smart computing will fully empower all industries, covering the entire value chain from R&D, production, manufacturing to after-sales service.

\textsuperscript{67} China Information and Communication Research Institute, Gigabit Broadband Network Business Application Scenario White Paper, 2019
\textsuperscript{68} University of Oulu, Key drivers and research challenges for 6G ubiquitous wireless intelligence, 2019
\textsuperscript{69} GIV, White Paper on the Commercial Application Scenarios of Gigabit Broadband Network, 2019
\textsuperscript{70} China Institute of information and Communication, White Paper on the Commercial Application Scenarios of Gigabit Broadband Network, 2019
As the smart society evolves, the structure of the ubiquitous computing power deployment architecture will also change: more computing power will concentrate on the cloud and the edge. Based on our prediction, in 2045 when the smart society leaps to the mature stage, the weight of computing power distributed on the cloud, the edge and the device will be 55%, 35% and 10%, respectively. At present, data process happens largely on the device level, but it will gradually move to the cloud and the edge. This trend is evidenced by, among others, the increasing popularity of smart scenarios. For example, on Google’s cloud game platform Stadia, games are hosted on remote servers and video can be streamed through cloud to connected devices, including smartphones, tablets, computers and TVs, reducing the computing power load on the device level. The trend is justified from both the demand side and the supply side:

1) On the demand side: Users will demand smart devices to be more convenient, more integrated and lighter. This means the computing power carried on the device level will decrease.
2) On the supply side: Future network will be able to carry more computing power as it breaks free from the current limitations and support larger magnitude of data transmission with higher efficiency.

71 IDC, Data World, From Edge to Core, 2018
72 IDC, Cloud IT Infrastructure Revenues Continue to Expand Despite Slow Down in Spending in 2019, 2019
2.3.2 Computing power demand and typical smart scenarios at different stages of smart society

Examples of technical level signs and scenes at different stages of a smart society

<table>
<thead>
<tr>
<th>Stage</th>
<th>Technological Indicators and Smart Scenarios</th>
<th>Smart Manufacturing</th>
<th>Autonomous Driving</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early Stage</td>
<td>Chip performance: · Neural network chips available</td>
<td>Smart concept verification</td>
<td>Conditional Automation</td>
</tr>
<tr>
<td></td>
<td>Other technological indicators: · Key technologies: AI, blockchain, etc. are not mature yet</td>
<td>Automated production</td>
<td>Parking route planning</td>
</tr>
<tr>
<td></td>
<td>· Bandwidth: fixed network bandwidth ~500 Mbps (~20min1), mobile network bandwidth ~150Mbps (~60min1); in the later part of this stage, bandwidth will increase to 10Gbps (~60s1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>· Latency and connected device density: decrease from 50ms to below &lt;10ms, up to 1-10 million devices per square kilometer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Developing Stage</td>
<td>Chip performance · 3nm chip might be mass-produced</td>
<td>Smart decision-making</td>
<td>High automation</td>
</tr>
<tr>
<td></td>
<td>Other technological indicators: · Key technologies: smart decision-making, robotics and speech recognition are mature</td>
<td>Self-initiated automatic production</td>
<td>Autonomous parking lot searching</td>
</tr>
<tr>
<td></td>
<td>· Bandwidth: up to 1Tbps (~6s1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>· Latency and connected device density: about 0.1ms, up to 100 million devices per square kilometer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mature Stage</td>
<td>Chip performance · chip may realize quantum computing</td>
<td>Autonomous concept generation</td>
<td>Full automation</td>
</tr>
<tr>
<td></td>
<td>Other technological indicators: · Key technologies: brain-computer interface, quantum computing, holographic projection, etc. are mature</td>
<td>Predictive optimization</td>
<td>Parking concept eliminated</td>
</tr>
<tr>
<td></td>
<td>· Bandwidth: way above 1Tbps 2 (close to instant completion1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>· Latency and connected device density: negligible; unlimited</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Time required to download a 2 hours long, 4K UHD Blu-ray movie (~60G);
- Assuming no significant difference between broadband and mobile networks

Source: Desktop research, Roland Berger

Figure 2.15 Technology development of different stages will lead to higher per capita computing power

Through some performance patterns of different industries in different stages of smart society, the important role of computing power is illustrated (Figure 2.15).

- Overview of smart manufacturing technologies and scenarios in different stages of smart society
### Evolvement of smart manufacturing

<table>
<thead>
<tr>
<th>R&amp;D</th>
<th>Production</th>
</tr>
</thead>
</table>
| **Early Stage** | - Smart concept verification  
  - Smart computing can only provide partial support by verify the feasibility of product concept | - Automatic production with human input  
  - Human collaboration is required to achieve automated production |
| **Developing Stage** | - Smart decision-making  
  - Smart system can generate prototype options and contribute to the decision-making on product concept | - Automatic production without human input  
  - Smart systems can execute production orders without human input |
| **Mature Stage** | - Autonomous concept creation  
  - Smart R&D system can autonomously generate product concepts through sophisticated algorithm | - Production plan - predictive optimization  
  - Smart systems can use sophisticated algorithms to predict macroeconomic development, laws of science development, social and political events, probability of epidemics, development trend of both the industry and the company, and make optimization accordingly |

Source: Desktop research, Roland Berger

Figure 2.16 Smart scenarios in manufacturing industry will upgrade as smart society evolves.

1) At the early stage, technology limitations hamper the digitalization in manufacturing and only stereotyped and low-level, modular smart production can be realized (Figure 2.16). Specifically, the AI-powered neural network chips are capable of algorithm training with big data input. Therefore, it is possible to realize dynamic equipment adjustment, streamline the links on production line and assembly line, and eventually bring up the overall productivity. As emerging technologies such as AI cloud services mature, smart factories can upload large amount of data to cloud data centres for storage and analysis, and as a result optimise smart production procedures and further boost productivity. With above 1Gbps bandwidth and below 10ms latency, real-time data synchronization will be possible, and can be used to track in real-time the semi-finished and finished customized goods and optimize the mixed production process accordingly. The flexibility achieved in this way will address the uncertainties in a diversified and sophisticated production line.

2) At the developing stage, advance in technologies will enable more smart manufacturing scenarios such as rapid R&D, independent innovation, and human-machine collaboration, leading manufacturing into an era of mid-level intelligence. The breakthrough in chip integration will significantly improve chip performance. The neural network integrated in the chip will be very powerful in extracting information and will lead to simpler design, faster computing and as a result, faster R&D. From the technical perspective, when a smart system is more capable of smart decision-making, it will be able to develop new prototypes and facilitate the decision-making on product concepts. Network with higher capacity ...
will enable real-time industrial operation. Smart systems can work without human input and process production orders with high efficiency and high quality.

3) At the mature stage, smart systems will be able to create product concepts autonomously and optimize demand prediction, bringing the manufacturing industry to a fully intelligent era. Using quantum computing chips, which will then be widely available, smart systems can use sophisticated algorithms to predict macroeconomic development, laws of science development, social and political events, probability of epidemics, etc. Smart systems can predict the development trend of both the industry and the company, and make optimization accordingly. By then, smart robot technology will be largely mature; latency will be negligible; cloud-enabled robots can move data computing and storage on the cloud to cut the hardware cost and power consumption of the robot itself.

- Autonomous driving levels in different stages of smart society

### Transportation Industry

<table>
<thead>
<tr>
<th>Autonomous Driving</th>
<th>Parking</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Early Stage</strong></td>
<td></td>
</tr>
<tr>
<td>Conditional Automation (Level 3)</td>
<td>Parking route planning</td>
</tr>
<tr>
<td>- Under certain road conditions, vehicles can assist in blind-spot detection, lane-keeping, active cruise and braking.</td>
<td>- CPU can work out the optimal parking route and strategy by processing such data as vehicle location, destination and environment parameters</td>
</tr>
<tr>
<td>- Human override is required in certain conditions.</td>
<td></td>
</tr>
<tr>
<td><strong>Developing Stage</strong></td>
<td></td>
</tr>
<tr>
<td>High Automation (Level 4)</td>
<td>Parking slot searching</td>
</tr>
<tr>
<td>- Vehicles are fully capable of driving itself without need for human input or override</td>
<td>- Smart systems on-board will use deep learning to precisely pinpoint optimal parking slot by automatically detecting vehicle position and surrounding boundaries</td>
</tr>
<tr>
<td><strong>Mature Stage</strong></td>
<td>Shared commute (No parking)</td>
</tr>
<tr>
<td>Full Automation</td>
<td>- People may use implanted chip in the brain to find and use idle vehicles and share them with the others in destination, saving the trouble of parking and eliminating the concept of parking</td>
</tr>
<tr>
<td>- All driving operations are completed by the autonomous driving system under any condition</td>
<td></td>
</tr>
<tr>
<td>- No need of human drivers</td>
<td></td>
</tr>
</tbody>
</table>

Source: Desktop research, Roland Berger

![Figure 2.17 Driving autonomy will upgrade along with the smart society](image)

1) At the early stage, only entry-level smart functions like early warning and 3D image simulation can be realized in the autonomous driving scenario, (Figure 2.17). Neural networks in the chip can only manage visual control. Level 3 driving autonomy is mature enough to provide primary assistance, including blind-spot detection, lane-keeping, active cruise, and braking assistance. However, human override is required in certain conditions. As the mobile network bandwidth reaches 10Gbps, the CPU can work out the optimal parking route and strategy by processing such data as vehicle location, destination and environment parameters.

2) At the developing stage, higher rate of autonomous driving will be possible, enabling optimal parking route planning, etc. Smart driving will enter the era of mid-level intelligence. Breakthroughs in chip integration and performance will significantly improve computing power of the chips. In the automatic parking scenario, smart systems on-board will use deep learning to precisely pinpoint optimal parking.
slot by automatically detecting vehicle position and surrounding boundaries. Moreover, with the improvement of smart decision-making and deep neural network learning, the autonomous driving system is fully capable of driving itself without need for human input or override. The network bandwidth will approach 1Tbps and improve the communication efficiency through the Internet-of-Vehicle.

3) At the mature stage, smart driving will enter the era of high-level intelligence, where vehicles can be controlled from the cloud. In the cloud control scenario, quantum computing chips in the vehicle can collect and process environment data, identify emergencies and respond in real-time. By then, brain-machine interface technology will be fully developed. People may use implanted chip in the brain to find and use idle vehicles and share them with the others in destination, saving the trouble of parking (eliminating the concept of parking). Network transmission latency will be negligible. After the driving rights are passed to the cloud, the cloud can control the vehicle with real-time vehicle and environment data, avoiding any accidents caused by latency.
Chapter 3
Benefits of Computing Power to Countries and Regions

Key Takeaways from Chapter Three

3.1 Investing in computing power will generate sizeable economic returns

- Investing in the cloud-edge-device computing power infrastructure, a high value-added high-tech industry, will bring direct economic benefits to countries.
- USD 1 invested in public cloud services will bring a return of USD 4.7

3.2 Investing in computing power will have positive spill-over effects to other industries

- More investment in computing power leads to more advanced digital and intelligent technologies, which, in turn, will contribute to the output, efficiency, innovation and user experience of many industries, including manufacturing, transportation, healthcare, retail, etc.
- USD 1 invested in smart factories of USD 10

3.3 Investing in computing power will lift comprehensive nation power

- Investing in computing power will improve a nation’s scientific and technological excellence, make government more efficient in serving the public and improve people’s welfare

Source: Roland Berger

3.1 Investing in Computing Power Will Generate Sizeable Direct Economic Returns

Investing in computing power infrastructure, a high value-added high-tech industry, will bring direct economic benefits to countries\(^{73}\). Revenue of sales, leasing and service from products on each level of the cloud-edge-device architecture, including data centres, servers, CDNs, base stations and smart devices, can all add up to the value of the industry. For example, on the cloud level, cloud services like IaaS (Infrastructure as a Service), PaaS (Platform as a Service) and SaaS (Software as a Service) will be the main revenue sources. Organizations and individuals can pay to use the cloud services as per their need. Actual statistics show that, investing in the cloud-edge-device computing power infrastructure including chips, servers, smart devices, data centres will directly contribute to a country’s GDP.

\(^{73}\)Investment in computing power means: 1) investment made to such end products as servers, data centres, edge servers, intelligent devices and high-speed internet, or/and 2) investment made to fund R&D, production, talents training, infrastructure deployment in computing power related areas.
Ubiquitous Computing Power: the Cornerstone of an Intelligent Society
Position Paper

Result of direct return calculation and methodology

<table>
<thead>
<tr>
<th>Methodology</th>
<th>ROI Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample scenario: public cloud service eg. Dell PowerEdge R940 Server</td>
<td>Investment</td>
</tr>
<tr>
<td>$1</td>
<td>$4.7</td>
</tr>
</tbody>
</table>

Comments

- Investing in computing power, a high value-added sector, will create considerable value to the ICT industry

Figure 3.2 Measurement of direct economic returns of computing power investment

Take public cloud server as an example (Figure 3.2). Return on investment (ROI) can be calculated with the hardware cost of server suppliers and their leasing revenue. DELL PowerEdge R940 Server, at an ex-factory price of USD 25,100, 5 years of life and 75% of utilization rate, will generate a leasing revenue of USD118,586, an ROI of 4.7x.

3.2 Investing in Computing Power Will Have Positive Spill-over Effects to Other Industries

More investment in computing power leads to more advanced digital and intelligent technologies, which, in turn, will contribute to the output, efficiency, innovation and user experience of many industries, including manufacturing, transportation, healthcare, retail, etc. For example, progress of emerging technologies based on intelligent computing (such as smart robots) is expected to add USD170 billion of value to U.S. manufacturing industry. The value creation is reflected on both the macro and the micro level. On the macro level, investing in computing power can improve the size of the whole smart industry. For instance, investing in smart driving system will expand the scope of the transportation and logistics industry. On the micro level, investing in computing power can enhance the value creation of smart scenarios. For example, more digital hardware devices mean higher revenue for cashier-less Amazon Go convenience stores.

74 AWS, AWS Storage Gateway, 2019
75 United States Department of Labor
## Returns on investment in computing power

<table>
<thead>
<tr>
<th>Industry</th>
<th>Investment</th>
<th>Return</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manufacturing (Smart factory)</td>
<td>Cost of factory digitalization Alibaba official website (^1)</td>
<td>($10) Annual turnover increase Alibaba official website (^1)</td>
</tr>
<tr>
<td>Transportation (Autonomous driving)</td>
<td>Hardware cost of autonomous vehicle LMC, IHS</td>
<td>($5) Expanded size of transportation industry Qianzhan Industrial Research Institute</td>
</tr>
<tr>
<td>Retail (cashier-free Store)</td>
<td>Hardware cost of Amazon Go official website</td>
<td>($5) Revenue of Amazon Go Media Advisory report</td>
</tr>
<tr>
<td>Power (Smart Grid)</td>
<td>Smart Grid investment Electric Power Research Institute</td>
<td>($4.4) Cumulative economic benefits Electric Power Research Institute</td>
</tr>
<tr>
<td>Agriculture (Spraying Drone)</td>
<td>Market size of agricultural drone PwC</td>
<td>($2) Total pesticides saved Agropages, DII</td>
</tr>
</tbody>
</table>

\(^1\) Data source

Source: Desktop research, Roland Berger

We did quantitative analysis on the contribution of computing power investment to manufacturing, transportation/logistics, retail, energy and agriculture, and qualitative analysis on its contribution to the entertainment industry. (Figure 3.3)

1) Manufacturing industry. Take the Hangzhou Dianshi Factory as an example. It is the first digital factory transformed by Alibaba. CNY 50,000 of digitalization investment was made to purchase computing power hardware, including servers, sensors, robots, etc. (software, manpower and network deployment costs were not included). After the transformation, the factory’s annual turnover increased by 6% on top of the previous CNY 1.6 million, i.e. CNY 96,000. Multiply the number by 5 (years of hardware life) is the total return, showing an ROI of about 10\(^x\). \(^6\)

2) Transportation/logistics industry. Take autonomous driving vehicle as an example. In 2018, sales volume of passenger vehicles in China was 23.7 million; penetration rate of vehicles below Level 3 autonomy was 19%; autonomous driving hardware (computing platform, cameras, lidars, etc.) for one vehicle costs $23,735. At a USD/CNY exchange rate of 7.1, the investment in computing power-supporting hardware in the autonomous driving industry is CNY 758.5 billion. Based on China’s CNY 593.6 billion-worth transportation market, an efficiency boost from autonomous driving of 60% and 10-year life 

\(^6\) Alibaba, Digitization of Tao Factory: 2,000 pieces of clothes are produced in 5 minutes, 2018
of hardware invested according to our estimation, it can be calculated that investing in autonomous
driving hardware will add CNY 3,561.6 billion to China’s transportation industry, a ROI of about 5x.77

3) Retail industry. Take the cashier-free Amazon Go convenience store as an example. Based on USD 1.47
million of hardware investment per store covering RFID, computers, cameras, etc., (i.e. computing
power investment), USD 1.5 million of annual turnover per store, and five years of life of hardware
invested, the return on computing power investment amounts to USD 7.5 million, an ROI of 5x.78

4) Energy industry. Take the smart grid as an example. According to the research by the Electric Power
Research Institute, to build a fully functioned smart grid would require $388~$476 billion USD of
investment, which would generate 1.294~2.028 trillion USD of aggregate return, representing an ROI
of about 4.4x.79 In its the collaboration with Huawei80, the Shenzhen Power Supply Bureau of China
Southern Power Grid applied Huawei’s Artificial Intelligence Internet of things (AIoT) technology, and
deployed video monitoring terminals to inspect the power transmission system. This practice reduced
the inspection cycle from 20 days to 2 hours, an efficiency increase of 80 times. In the past, with an
efficiency of 24km/h, inspecting the more than 3,900km of 110kV above overhead transmission/
distribution grid took 20 days. In comparison, AIoT-based video inspection has an efficiency of 195km/
hour with high precision and capability to detect potential risks invisible to the human eyes. The two
key measures behind efficiency improvement are:

(a) Construct the “cloud-edge-device” AI collaboration architecture. On the cloud level, deploy cloud
computing centre and cloud resource pool as AI platforms; On the edge level, deploy Huawei Atlas 500
AI edge station or light cloud server to upgrade AI capability; On the device level, deploy Huawei Atlas
200 AI accelerator module on smart cameras.

(b) Formulate targeted inspection plan. For instance, conduct automatic video analysis and drone video
analysis to overhead grid; Deploy light cloud AI and 1-to-N smart cameras in transformer substation,
distribution room and transmission tunnel. To address the data collection difficulty and poor data
quality in the power industry, Huawei’s IoT data collection technology is utilized to unify data
 specification and improved channel capacity. As a result, 30 times more data was collected on the
device level, contributing to the stability and security of power grid operation.

77 GGII, MarkLines, HIS, NIO Capital, Infineon, Mitsubishi UFJ Morgan Stanley Securities
78 Amazon, Capital Markets of Royal Bank of Canada, Best Reviews, iiMedia Research
79 Electric Power Research Institute
80 Shenzhen Power Supply Bureau of China Southern Power Grid
5) Agriculture industry. Take the agricultural spraying drone as an example. Investing USD 400 million (i.e. the computing power investment) to purchase agricultural spraying drones in China will reduce 50% of pesticide consumption. China’s pesticide market is worth USD 6.5 billion. If agricultural spraying drones can serve 5% of China’s arable land and 5 years of useful life, the investment return will amount to USD 800mm, an ROI of 2x.\(^{81}\)

6) Entertainment industry. Pixar invested in high-performance computers, bringing the quality and definition of 3D films as well as audience experience to a whole new level. To meet the demand from audience for better and more vivid special effects, film makers need more computing power to upgrade their graphic rendering algorithms (to improve hair, texture and light rendering, etc.). As an early mover, Pixar has invested heavily to build its own high-performance computing centre. In 2013, Pixar housed 2,000 servers with a total of over 24,000 cores, and is estimated to be a top 25 supercomputer. As a result, the time required to render animations and special effects films significantly decreased and image quality increased. The large demand for computing power depends on the data amount and resolution requirement per frame. For example, it takes 50 hours of an ordinary workstation to render one frame of a top-level IMAX 3D film. With this speed, a 90-minute movie running at 24 frames per second would take 740 years to render. But Pixar’s high-performance computing centre can get it done in a week or so. It is thanks to the high-performance computing centres that the resolution of films has rapidly increased from 1080P to 4K IMAX 3D and made the popular films like Avatar, Planet of the Apes and Avengers as amazing as they are.

Although investment in smart computing technologies like AI will significant drive economic growth, it is also subject to "marginal effect", which is estimated to diminish when the future smart society evolves to developing stage and mature stage. Take industrial robot as an example. Researchers from the London School of Economics and Uppsala University in Sweden studied data from 17 countries, and found that larger increases in robot density will translate into increasingly small gains in productivity, suggesting that there are some congestion effects (or diminishing marginal gains) from increased use of robots. The same thing may happen in the intelligent computing industry, meaning that as the density of smart devices increases, the marginal effect per additional unit computing power on economic growth will diminish.

3.3 Investing in Computing Power Will Lift Comprehensive Nation Power by Driving Innovation, Optimise Public Service and Improve People’s Welfare

Investment in computing power will pay back. Higher computing power in innovation-intensive sectors will improve a nation’s scientific and technological excellence; Higher computing power in public service industry will make government more efficient in serving the public, thus contributing to a peaceful and harmonious society; Higher computing power and more digital infrastructure in sectors concerning people’s livelihood like education, scientific research, healthcare, employment and community service will improve people’s welfare. (Figure 3.4).

\(^{81}\) CCID consulting, China Business News, Phillips McDougall, Pacific Securities
Ubiquitous Computing Power: the Cornerstone of an Intelligent Society
Position Paper

Figure 3.4 Investing in computing power will life comprehensive nation power by driving innovation, optimize public service and improve people’s welfare

Investing in computing power could boost national comprehensive power by advancing four different types of innovations, namely, efficiency-driven innovation, customer-orientated innovation, engineering/technical innovation and scientific innovation.

Efficiency-driven innovation refers to the innovation in manufacturing sector (e.g. bulk chemicals, textile, electrical equipment and engineering machinery) that will help cut cost, raise efficiency and improve quality. Computing power investment in corporate-level computing centres and devices (e.g. IoT devices in smart factories) could contribute to efficiency-driven innovation by offering, among others, big data solutions to track cost patterns and reduce operation cost.

Customer-orientated innovation leads to progress in products, services or business models that will help address customers’ concerns. It is most needed in industries like internet software and services, home appliances and household products. Computing power investment in cloud computing centre and smart devices (e.g. smartphone and digital freight tag) can contribute to this type of innovation. With stronger computing power, the B2C industries represented by the FMCG industry will drive their product R&D by customer demand.

Engineering/technical innovation refers to new product development in aircraft manufacturing, car manufacturing, telecommunication device industry, etc. Investing in computing power will contribute to this type of innovation as equipment manufacturers may use big data and AI to optimise their production, equipment maintenance and quality control.

Scientific research innovation is crucial to commercialize fundamental R&D achievements. It is most needed in such fields as brand-name drugs, biotechnology, semiconductor design and specialty

<table>
<thead>
<tr>
<th>Use Case</th>
<th>Investment</th>
<th>Return</th>
</tr>
</thead>
<tbody>
<tr>
<td>China’s National University of Defense Technology invested in the Tianhe-1 supercomputer project to promote medical innovation</td>
<td>Invested CNY 600 million to build the Tianhe-1 supercomputer (No. 1 of the world’s TOP500 supercomputers in 2010)</td>
<td>· Facilitate medical research in new drug screening, drug toxicity screening and drug redevelopment · Serving 8,000+ R&amp;D computing tasks per day and promoting national comprehensive power</td>
</tr>
<tr>
<td>· The Estonian government invested in developing e-government to better serve its citizens</td>
<td>Cloud computing center · Internet network · Terminal PC</td>
<td>· 46.7% of Estonian citizens vote online · 98% citizens have electronic ID cards and 99% of government services can be processed online, saving time for citizens and significantly reducing compliance cost</td>
</tr>
<tr>
<td>The EU and the CERN1 set up the Helix Nebula computing center to serve scientific research</td>
<td>Invested about EUR 5.3 million to build hybrid cloud platform</td>
<td>· Helped Pan-cancer initiative to compare cancer samples and raised the success rate of cancer early screening, made early treatment and preventive intervention possible, extended people’s life expectancy</td>
</tr>
</tbody>
</table>

Source: Desktop research, Roland Berger

Innovation (R&D) | Governance (Public Services) | Welfare (Medical Services)

- · China’s National University of Defense Technology invested in the Tianhe-1 supercomputer project to promote medical innovation
- · The Estonian government invested in developing e-government to better serve its citizens
- · The EU and the CERN1 set up the Helix Nebula computing center to serve scientific research

- · Invested CNY 600 million to build the Tianhe-1 supercomputer (No. 1 of the world’s TOP500 supercomputers in 2010)
- · Cloud computing center · Internet network · Terminal PC
- · Invested about EUR 5.3 million to build hybrid cloud platform

- · Facilitate medical research in new drug screening, drug toxicity screening and drug redevelopment · Serving 8,000+ R&D computing tasks per day and promoting national comprehensive power
- · 46.7% of Estonian citizens vote online · 98% citizens have electronic ID cards and 99% of government services can be processed online, saving time for citizens and significantly reducing compliance cost
- · Helped Pan-cancer initiative to compare cancer samples and raised the success rate of cancer early screening, made early treatment and preventive intervention possible, extended people’s life expectancy

Source: Desktop research, Roland Berger

Figure 3.4 Investing in computing power will life comprehensive nation power by driving innovation, optimize public service and improve people’s welfare
Ubiquitous Computing Power: the Cornerstone of an Intelligent Society
Position Paper

chemicals. Investing in computing power, particularly in cloud infrastructure like supercomputers and hyperscale computing centres, will contribute to this type of innovation by enabling R&D institutions and enterprises to discover new drug substance and new materials, and accelerate the commercialization process.

Among the four types of innovation, we now take the scientific research innovation in the medical field as an example. China’s Tianhe-1 supercomputer is built to tackle the key issues in the R&D of traditional medicines, including high screening cost, high toxicity test cost and extremely high failure rate. Since 2000, China has been consistently focusing on researching, investing and constructing supercomputers. As an important milestone of supercomputer development history in China, in November 2010, the Tianhe-1 supercomputer jointly developed by the National University of Defence Technology and the Tianjin government with CNY 600 million of investment, was launched with peak computing speed of 4,700 trillion calculations per second, ranking the first among the world’s TOP500 supercomputers in 2010.

Tianhe-1 and a large number of other supercomputing centres have provided a big data processing platform for biomedical R&D and promoted scientific research innovation in the medical field. Harnessing the power of the platform, many scientific institutions and enterprises in China have made numerous achievements in developing medicines for Epilepsy, AIDS, cancer and other diseases. Supercomputers can help medicine development in three fundamental ways:

New medicine screening: Screening millions of compounds to determine whether a medicine can bind with disease-related target protein;

Medicine toxicity screening: Testing and analysing whether the medicine could potentially bind with any of the nearly 1,000 human proteins and generate toxicity, and eventually preventing the toxic medicine from entering the most expensive clinical trial process;

Medicine redevelopment: Matching existing medicines to different target proteins to shorten the R&D process.

Statistics show that Tianhe-1 performs more than 8,000 R&D calculation tasks every day in biomedical, petroleum resource development, equipment manufacturing and other industries. It has supported more than 1,500 national major technological projects and R&D plans, becoming an irreplaceable platform for national scientific and technological innovations.

Computing power investment in public services will greatly improve the government efficiency, promote social equality and create a better society. Take e-governance as an example. Through building cloud computing centres and national database, the Estonian government has made most public services accessible online, minimizing the time cost for citizens and the compliance costs for enterprises. Since Estonia’s independence from the former Soviet Union, it has been focusing on restructuring government public service network and improving governmental service efficiency. With these objectives in mind, the Estonian government invested in computing power infrastructure, including cloud computing centre, internet network and terminal computer. Now Estonia’s e-government initiative is well under way. Important information of its citizens, including birth records, register of electors, housing deeds and bank
receipts, can all stored in electric forms. The Estonian digital tax system allows taxpayers to complete the tax filing process in just two minutes, significantly reducing compliance costs. In this tax system, the large cloud computing centre conducts data processing, analysing and storing; a large number of terminal computers work as a tax declaration information input nodes, which are supported by highly user-friendly interactive software. This has effectively lowered the technical threshold of electronic tax declaration. Available statistics show that 46.7% of Estonian citizens vote online, 98% of them have electronic ID cards and 99% of government services can be processed online. In total, the digitalization initiative can help the country save operation costs equivalent to 2% of national GDP.

Investing in computing power investment will improve people’s livelihood and increase their sense of satisfaction and happiness. In the scientific research sector, the EU is committed to bringing research institutions, data providers, public information infrastructure, commercialized cloud service providers together to commercialize scientific cloud service. As one of EU’s major cloud initiatives, Helix Nebula - The Science Cloud (HNSciCloud) launched a EUR 5.3 million tender for the establishment of a European hybrid cloud platform. The purpose of the platform is to support the deployment of high-performance computing and big-data capabilities for scientific research. HNSciCloud has played an important role in many scientific domains. Below are three of its use cases:

First, it helped the 3DIX project on ESRD (End Stage Renal Disease) to produce 3D Imaging with X-rays. The objective is to make 3D images (volumes) of nanoscale objects to study their characteristics on a nanoscale. These studies can then be applied to a wide variety of objects and scientific fields such as chemical studies, life sciences, structure of materials etc.

Second, it has helped HADDOCK (High Ambiguity Driven protein-protein DOCKing) build a platform where more than 10,000 users have securely placed the identified or precited protein information coding into AIRs to drive the docking process.

Third, it has assisted the PanCancer initiative in comparing 12 tumour types in 5000+ samples on a monthly basis. PanCancer currently represents the most comprehensive computational study dealing with cancer genomics, with roughly 1 PB of data to be processed. Helix Nebula can efficiently process the large amount of data within a short time window (one month). These practices raised the success rate of cancer early screening, made early treatment and preventive intervention possible, significantly reduced the death rate of cancer patients, extended the life expectancy of residents, and effectively improves people's sense of happiness.
Chapter 4
Challenges in Computing Power Development and How to Tackle Them

Key Takeaways from Chapter Four

4.1 Three challenges and two concerns

- Increasing power consumption challenges the sustainable development of the future smart society
- Uneven network development hampers computing power deployment
- Deficient computing power ecosystem frustrates the realization of diversified smart scenarios
- Insufficient infrastructure and security are two major concerns to address

4.2 Governments can make policy planning, build infrastructure and provide services, extend fiscal support, and enhance legal framework and supervision to make tackle challenges and address concerns

- To support key technologies cloud computing and AI: actively deploy computing power infrastructure construction
- To cut power consumption: 1) introduce industry-wide energy-saving; 2) directly invest in new energy; 3) establish a carbon trading mechanism
- To improve network quality: 1) form a monitoring network and 2) leverage private fund through public-private partnership (PPP) model
- To diversity the computing power ecosystem: 1) set goals to achieve higher level of diversity and promote infrastructure construction and promotion of smart scenarios, 2) encourage stakeholders to form industrial alliance, and 3) broaden market access
- To address security concern: 1) develop international security standard; 2) allow third-party to facilitate supervision monitoring; 3) develop an adequate accountability mechanism

4.1 Developing Computing Power Faces Challenges of Power Consumption, Network Limitations and Diversified Ecosystem

Stronger computing power is indispensable to the evolvement of all industries and the future smart society. However, its development will not be smooth sailing – side effects will start to emerge at a certain point, like the surge of power consumption per unit of computing power. Such problems need collaboration of countries across the world to tackle together. In addition, computing power will have to develop in a less than optimal environment featuring such drawbacks as network limitations and weak ecosystem. Countries need to be aware of these issues and address them proactively. In this paper, we have identified three major challenges in the development of computing power:
- Increasing power consumption challenges the sustainable development of the future smart society;
- Uneven network development hampers computing power deployment;
- Deficient computing power ecosystem frustrates the realization of diversified smart scenarios.

While tackling the above challenges, insufficient infrastructure and security are also two major concerns for the ICT industry to address.

### 4.1.1 Increasing power consumption challenges the sustainable development of the future smart society

Power consumption is one of the three major factors impacting the development of computing power. While the power consumption per unit of computing power is decreasing year by year and gradually approaching its floor, the overall magnitude of computing power is increasing rapidly due to wide application. This is a conundrum: on the one hand, wide application of computing power is necessary for smart society to evolve; on the other, the evolvement will be accompanied by a sharp increase in carbon emission, adding uncertainties to the global climate change. Solving this conflict should be high on the agenda of all countries.

#### Power consumption per unit of computing power

![Power consumption per unit of computing power graph](image)

- **Intel Pentium III 800**
- **Intel Pentium III 1000**
- **Intel Pentium D820**
- **Intel Core 2 Duo E4500**
- **Intel Core 2 Quad Q9400**
- **Intel Core i7-2600K**
- **Intel Core i7-4770K**
- **Intel Core i7-8700K**
- **AMD Ryzen 7 2700X**
- **AMD Ryzen 7 3800X**

*Source: Intel, AMD, Roland Berger*

Figure 4.2 Power consumption per unit of computing power declines significantly as nanometer process upgrades

Overall, the power consumption per unit of computing power, or power per performance of chips, significantly declined in the past 10 years. As it becomes more difficult to upgrade the silicon manufacturing process, the room for further decline is limited. As indicated in Figure 4.2, as the chip process keeps updating, the power per performance of chips (W/GFLOPS) has declined exponentially decline from 90nm (Figure 4.2). As the power per performance of the latest 7nm chip approaches 0 W/GFLOPS, the room for further decline is extremely limited. Theoretically, the chip will see its power...
per performance decline when it becomes smaller. But this rule is being challenged by the quantum tunnelling effect. Quantum tunnelling happens when the transistor gate oxide on silicon-based chips becomes too thin to contain the electrons. Once the critical threshold is hit, it will be impossible for the transistors to have an off state. Therefore, without new materials or technological breakthroughs, it is difficult to further cut the power consumption per unit of computing power.

Adding to the difficulty is that fact that the widespread smart application is inevitable, which is bound to increase total power consumption even further. Take China as an example. The total power consumption of data centres in China has increased by more than 12% for eight consecutive years driven by the development of AI, IoT, blockchain, etc. In 2018, the total power consumption of China's data centres was 160.889 GWh, accounting for 2.35% of the total power consumption of the whole nation and significantly higher than that of Shanghai (156.7 GWh). In the future, the computing power of the data centre will continue to grow along with the rapid development of the ICT industry. In 2023, it will increase by 66% on top of the 2019 level with a CAGR of 10.64%, leading the total power consumption to go further up. Therefore, to ensure the sustainable development of the future smart society, optimising the PUE (power usage effectiveness) of power-intensive data centres is a core challenge to the development of computing power.

**Average PUE of different types of data centers in Europe and China**

<table>
<thead>
<tr>
<th>Data center type</th>
<th>Europe</th>
<th>China</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hyperscale</td>
<td>1.20</td>
<td>1.50</td>
</tr>
<tr>
<td>Large</td>
<td>1.25</td>
<td>1.69</td>
</tr>
<tr>
<td>Small and medium</td>
<td>1.74</td>
<td>2.25</td>
</tr>
<tr>
<td>Average</td>
<td>1.70</td>
<td>2.20</td>
</tr>
</tbody>
</table>

**Comments**
- PUE of large and hyperscale data centers are better than small and medium ones
- Large and hyperscale data centers are better at infrastructure construction and operation management
- Increase the number of large and hyperscale data centers will significantly improve average PUE, therefore reducing energy consumption
- Take China as an example, every 0.1 average PUE reduction can save 7.3 GWh per year, which is enough to power the entire Shanghai city for 17 days
- Bringing down the average PUE value of all data centres in China to the level of large data centres, 37.3 GWh of power can be spared to sustain Shanghai for nearly 3 months, equivalent to reducing carbon dioxide emission by about 30 million tons and planting about 80,000 hectares of forests.

One key approach to optimize the overall PUE of data centres is to build more hyperscale (>10,000 standard racks containing at least 100,000 servers) and large data centres (>3,000 standard racks containing at least 30,000 servers). By analysing the PUE statistics of data centres in China and Europe, we find that larger data centres have lower PUE (Figure 4.3). In China, the PUE of hyperscale data centres (1.5) is 31% better than the total average (2.2). In Europe, the PUE of the hyperscale data centres is 31% better than the total average (2.2).

---

82 Greenpeace & North China Electric Power University, China Data Centre Energy Consumption and Application Potential of Renewable Energy Report, 2019

centre (1.2) is 29% better than the total average (1.7). The reason is that hyperscale/large data centres usually involve large investment and high energy consumption, therefore, they are more inclined to improve power efficiency by optimizing their infrastructure and management system. According to our calculation, bringing down the PUE value of China’s data centres by 0.1 will save 7.3 GWh, enough to sustain Shanghai for 17 days. Bringing down the average PUE value of all data centres in China to the level of large data centres, 37.3 GWh of power can be spared to sustain Shanghai for nearly 3 months. Such power conservation is equivalent to reducing carbon dioxide emission by about 30 million tons and planting about 80,000 hectares of forests.

In terms of infrastructure construction, hyperscale/large data centres are more likely to use power-efficient servers, and to invest more in energy-saving power solutions and cooling devices to optimize their PUE. Take Baidu Yangquan Cloud Computing Centre in Yangquan as an example. The Centre uses advanced rack scale design to accommodate 160,000 servers. The rack scale solution enables the power system, the cooling system and the server to reside in the same rack as modules. The actual operation and energy consumption of each rack server can be adjusted separately, saving 20% of energy every year. In addition, by combining high voltage direct current and direct municipal power supply, the data centre has achieved an energy efficiency of 96%, which is way beyond the reach of traditional UPS power supply, and an impressive PUE value of 1.23.

In terms of data centre operation management software, large data centres are more likely to use AI-based intelligent management systems to dynamically adjust power supply and cooling systems. For example, Huawei Langfang data centre adopts iCooling data centre energy efficiency solution. The AI-based system analyses historical operation data to detect key factors affecting energy consumption (power supply, ventilation, cooling, etc.) to build a prediction model. The model will facilitate dynamically adjustment to the system, and lower PUE and improve energy efficiency.

4.1.2 Uneven network development hampers computing power deployment

As mentioned above, network latency and bandwidth limits will dramatically impact the performance computing power across the cloud, the edge and the device levels. In all countries, there are areas not covered by network infrastructure. And in the areas covered, the quality of connection (bandwidth and latency) varies.
First, network coverage varies significantly across different regions in the world. Study shows that North America has the highest coverage of network infrastructure with only 1% of the population not yet covered. However, even in the best-connected region, 24% of the population have no internet access. In stark contrast, large percentage of people in the developing countries in South Asia and Africa have no internet access. In Africa, the percentage is as high as 30%. (Figure 4.4). The unbalanced network development among countries will surely be a roadblock toward the future smart society. Governments across the globe should make effort to tackle this challenge and lay a good foundation to realize full output of available computing power.

Figure 4.4 Uneven network coverage across different regions in the world

Source: GSMA, Roland Berger

---

GSMA, The State of Mobile Internet Connectivity Report 2018, 2018
Figure 4.5 Network quality varies vastly among different countries and regions

Second, bandwidth and latency vary vastly among the countries. Network speed values tested by a Speedtest, a professional testing software, are used for comparison. (Figure 4.5) Benefited from long-term investment in network infrastructure, Singapore and South Korea lead the world with a fixed bandwidth of nearly 200Mbps and an average mobile bandwidth of 50Mbps, more than 50 folds higher than the countries ranking the bottom, i.e. Venezuela and Algeria. Venezuela has the slowest network bandwidth of 3.61Mbps, which can only support basic web browsing. Average levels of network latency among countries are also vastly different. There is more than five times of difference in network latency between Singapore (the best performer) and Ghana (the worst performer). In Ghana, with its mobile network latency as long as 150ms, even the emergency call function of a car couldn’t work properly.

Bandwidth and latency have great impact on computing power output. Take the power grid as an example. Because the power consumption area is usually remote from power stations, it is necessary to use UHV network for power transmission. Therefore, the capacity of the network will influence all links of power transmission. Currently, numerous power application scenarios (power grid protection, drone patrol inspection, differential protection, smart meter reading, etc.) are constrained by network bandwidth limitations. (Figure 4.6)
Importance of network of high computing power in power industry

5G with low latency, large bandwidth and capable of satisfying a large number of device access is crucial for realizing the above scenarios

Source: Desktop research, Huawei, Roland Berger

In the case of drone inspection of power grid, current 4G network bandwidth is not sufficient for ultra-HD image transmission. As a result, the inspection and maintenance staff won’t be able to detect the minor defects in the grid and promptly eliminate potential risks. Specifically, most inspection drones are connected to the Internet. Their communication with the ground mainly serves three key purposes: image transmission, data transmission and remote control. Among them, image transmission is most demanding in terms of network quality. Although the drone employs the most advanced 4G LTE cellular communication technology (approx.100Mbps) and is covered by network base stations, it only allows 720P (not 4K ultra-HD) video transmission between the drone and base station or remote controller. The 720P resolution is not high enough to help technicians to detect minor defects and resolve potential risks timely, limiting the application of the drone inspection. Therefore, high-speed network like 5G is crucial in this scenario. When 5G is available, drones will be able to transmit HD video and technicians won’t have to climb up the power towers for inspection. By then, not only the inspection efficiency will be improved, operation risks will also be reduced.
4.1.3 Deficient computing power ecosystem frustrates the realization of diversified smart scenarios.

At present, the computing power ecosystem is still far from mature. Although X86 server chips can meet the requirements of common computing scenarios, diversified applications in future smart society will be demand more options of chip architectures. For example, big data search and cloud gaming scenarios demand high concurrency and device-cloud collaboration. However, the existing X86 architectures cannot satisfy the need of all scenarios. Therefore, it is necessary to diversify the computing power ecosystem by developing more chip architectures such as RISC, represented by ARM, and ensure end-to-end adaption. There are three major challenges with regard to the diversity of computing power ecosystem:

a) diverse application scenarios require for diversified chip architectures; b) different chip architectures require for different end-to-end adaption support; c) talent to work on end-to-end adaption is in short supply. Failing to tackle any of the three challenges will make fully realizing future smart scenarios difficult and hinder the evolution of smart society.

RISC architecture and support features for the future intelligent society, take ARM as an example

<table>
<thead>
<tr>
<th>Features of Architecture (ARM as an example)</th>
<th>High Concurrency</th>
<th>Device-cloud Compatibility</th>
<th>Low Power Consumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Search Engine</td>
<td>· The concurrency efficiency of ARM chip with the same clock rate is nearly 70% higher than that of X86 architecture</td>
<td>· Compatible with most mobile and tablet games (Android &amp; iOS)</td>
<td>· For the same performance, power consumption is 20% lower than traditional chips</td>
</tr>
<tr>
<td>Game on Cloud</td>
<td>· Big data search technology is particularly demanding for high computing power and concurrent performance</td>
<td>· When migrating games server to the cloud, ARM server shows better compatibility with lower performance loss</td>
<td></td>
</tr>
<tr>
<td>Cloud Computing Center</td>
<td>· When migrating games server to the cloud, ARM server shows better compatibility with lower performance loss</td>
<td>· The power consumption of cloud computing center accounts for a significant proportion of total cost, which requires lower power consumption solutions</td>
<td></td>
</tr>
</tbody>
</table>

Source: Desktop research, Roland Berger

Figure 4.7 Chip architectures represented by ARM will play a key role in supporting future smart society.

a) Diversified chip architectures are needed to satisfy the various concurrency and compatibility requirements by future smart scenarios. Although X86 servers can meet the requirements of general computing scenarios, RISC servers perform better in certain high technical scenarios (see Figure 4.7). Firstly, in terms of concurrency, take search engine as an example, big data-based search technology (such as MapReduce) is more demanding for high concurrency performance. This is because the technology requires large amounts of data to be sliced and multiple tasks generated concurrently to cope with the parallel search tasks initiated at the same time. The RISC chip has more physical cores. Each of core can operate independently. Therefore, the RISC chip can perform better than traditional
chips by increasing concurrent processing efficiency. When comparing E5-2649 v4 (X86, 10 cores) and Hi1616 (ARM, 32 cores), both sharing the same clock rate (2.4GHz) and design power (90W), it is found that the concurrency efficiency of the dual-path ARM server is 70% higher than the X86 server. Besides that, in big data storage and analytics, RISC architecture also performs better by 25% than X86. Secondly, in terms of compatibility, when migrating games server to the cloud, RISC server shows better compatibility. Because most mobile and tablet gaming applications are based on ARM, a typical RISC architecture. Deploying these applications to ARM servers only requires VM/Docker layer processing, while deploying them to X86 servers need to go through instruction translator, Android emulator and VM/Docker to achieve compatibility. In actual practice, using RISC servers can reduce the performance loss caused by instruction translation by 40%. Finally, in terms of energy consumption control, take cloud computing as an example. The value of cloud computing is that it leverages hyperscale data centres with large-scale server arrays to form a virtual computing platform so as to provide users with fast and flexible computing services. To support sustainable development with green computing power, it is important for data centres to control their power consumption. Actual statistics show that, RISC chips can deliver the same performance with 20% less power consumption than X86 chips. More importantly, the annual power consumption cost of servers is about 40% of its purchase cost. Therefore, deploying RISC servers will not only meet the need of green computing power, but also lower operation cost. Therefore, it is necessary to deploy RISC servers in the cloud computing scenarios (Figure 4.8).

b) Chip architectures require end-to-end adaptation to fully utilize computing power of chips. There are three main levels of end-to-end chip architecture adaptation. First, on the chip level, it requires market player in IP micro-core design, layout stitching, and wafer production to collaborate. Second, on the software level, on one hand, to realize the call of computing resources it’s required to dock the operating system with the compiler; on the other hand, the development kit of the applications is required to identify underlying commands in order to achieve programming development. Third, on the hardware level, we need to realize the full utilization of computing power through matching of different component bus interface bandwidth and chip computing power. It can be seen that whether it is chip, software or hardware, they all need to cooperate with underlying computing architecture to ensure the real application of computing power. None of these three links can be ignored. Nowadays, the lack of end-to-end adaptation, especially the incompatibility of system development, has dragged down the speed of RISC server’s development. To ensure a development and test environment consistent with the server side, and to reduce incompatibilities, most application developers are using X86 computer to develop server-end applications. As mentioned by Xiaoming Pan, global vice president at AMD, an architecture should have more than 25% market share in the server market to achieve sufficient supply chain profits so as to enable sustainable innovation and development.

85 ARM, 2016
86 Huanqiu.com, Pan Xiaoming, AMD Global Vice President: Breaking the Monopoly of x86 Servers, 2015
for the RISC server chip architecture, it needs to have more than 20% market share of the server chip market to have enough system developers, application developers, and cloud computing suppliers to adapt to RISC servers. Only in this way can the ultra-high performance power ratio of ARM architecture server be taken full advantage of to meet the computing power requirements of different scenarios.

c) To push forward the adaptation of various chip architectures from the levels of software, equipment and scenario, many high-end talents are needed to act as “lubricant”, who will help prosper the computing power ecosystem by thoroughly introducing chip computing power to the scenarios. The utilization of computing power needs high-end talents engaged in processes from underlying chip design, to equipment production, then to software development. Developers of operating systems, developers of application software, designers of smart devices, architects and engineers of software and hardware all play crucial roles in the construction of the entire computing power ecosystem. In order to realize further improvement of computing power and large-scale application of smart application scenarios, large-scale various high-end talents are needed.

4.1.4 Network infrastructure construction and security issues hinder computing power development

For instance, in cloud computing construction, we have mentioned above that there is a strong correlation between the national computing power and their economic performance. In addition, the investment in computing power has a strong and positive external effect to all industries. Therefore, if a mismatch between the computing power infrastructure development and the level of economic performance exists, it will influence the industrial development speed, and ultimately affect the national competitiveness. According to McKinsey research data, the cloud access rate of American enterprises is 80% in 2018. However, the rate in China is only 40%, which has greatly influenced the digital operation efficiency of Chinese enterprises. Besides, Fujitsu’s data centre outage is a typical case of insufficient investment in the construction of cloud computing facilities. Its transformer in Silicon Valley data centre is malfunctioned, which led to a power supply damage for the data centre. As a result, all SaaS and public cloud service were temporarily off-line and it took five days for the clients to get back online, which have harmed numerous customers’ daily operations. To ensure computing power security, since the transmission of information in the process of cloud computing power output will go through the four levels of network, software, equipment and chip, its information security needs to have a comprehensive security guarantee mechanism at every level. In the absence of any layer of protection, it is hard for the countries to guarantee the information security of their governments, enterprises and residents, which could lead to a severe information leakage and cause public security issues. For example, the Home Depot in the U.S. has experienced credit card holder information leakage issues. Due to the lack of network and software protection, 56 million credit card holders’ information were leaked, which posed a rigorous challenge to public property security in the U.S.
4.2 Governments can Make Targeted Efforts to Drive Computing Power growth: Making Policy Planning, Building Infrastructure and Providing Services, Extending Fiscal Support, Enhancing Legal Framework and Supervision, etc.

Policy initiatives needed to address the challenges

<table>
<thead>
<tr>
<th>Initiatives</th>
<th>3 Core Challenges</th>
<th>1 Core Issue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planning</td>
<td>✓ Power Consumption</td>
<td>✓ Security</td>
</tr>
<tr>
<td></td>
<td>✓ Network Quality</td>
<td></td>
</tr>
<tr>
<td></td>
<td>✓ Diversified ecosystem</td>
<td></td>
</tr>
<tr>
<td>Resources offering</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fiscal Support</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supervision</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: Desktop Research, Roland Berger

FIGURE 4.8 Summary of policy recommendations to governments

To lay a solid foundation for the development of computing power, government should embrace the challenges of rising power consumption, demand for higher network quality and building a diversified ecosystem, while taking the security issue in mind. Governments should take measures to boost computing power to fulfil the demand from all application scenarios in the future smart society. For example, governments can make policy planning, build infrastructure and provide services, extend fiscal support, enhance legal framework and supervision, etc. (Figure 4.8).

4.2.1 Governments should improve infrastructure to pave way for key technologies like cloud computing and AI to grow and prevail (Figure 4.9)
Ubiquitous Computing Power: the Cornerstone of an Intelligent Society
Position Paper

Many governments across the world, including the EU members, China and Japan, have taken initiatives to build fundamental computing power infrastructure, share data and develop applications.

In terms of infrastructure building: In 2016, the EU officially launched the Europe Cloud Initiative, a blueprint for cloud based services and world class big data infrastructure. The EU commission plans to progressively put in place the initiative through a series of actions, including:

- By 2016: Creating a European Open Science Cloud for European researchers and their global scientific collaborators by integrating and consolidating e-infrastructure platforms;
- By 2018: Launching a flagship-type initiative to accelerate the nascent development of quantum technology, which is the basis for the next generation of supercomputers.
- By 2020: Comprehensively developing and deploying large-scale high-performance computing, data storage and high-speed bandwidth infrastructures, including establishing a European big data process and storage centre, upgrading the backbone network for R&D and innovation, and building a new generation of supercomputer which could rank among the top three in the world.
- Guided by the initiative\(^7\), the EU has selected 8 hosting sites for supercomputing centres plans and acquire 8 supercomputers: 3 precursors to exascale machines capable of executing more than 150 PFLOPS\(^8\) or 150 million billion calculations per second, and 5 petascale machines capable of executing at least 4 PFLOPS, or 4 million billion operations per second. In a major step towards making the EU to a top supercomputing region globally, the EU has formulated a plan for both supercomputing

\(^7\) The European Commission, Digital Single Market: Europe announces eight sites to host world-class supercomputers, 2019

\(^8\) PFLOPS = petaFLOPS. 1 PFLOPS = 10\(^{15}\) FLOPS
infrastructure and talent development. In terms of infrastructure, by 2020, the EU’s objective is to acquire and deploy two supercomputers capable of executing around 10 PFLOPS and another 3-4 supercomputers capable of executing around 1 PFLOPS. In terms of talent cultivation, the EU has launched the Marie Skłodowska-Curie Actions (MSCA)\(^{89}\), a funding scheme supporting the mobility of researchers among EU members to create a pool of European researchers as well as scientific and innovation staff. The scheme will help lay a talent foundation for computing power growth in the EU. Apart from the above-mentioned moves, the EU also has introduced a series of policies to develop cloud computing and AI infrastructure. (Figure 4.10).

### Policies on artificial intelligence and cloud computing in China and EU

<table>
<thead>
<tr>
<th>China</th>
<th>Cloud Computing</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Set goals for core and fundamental industries, manufacturing technology and equipment and support system; require bandwidth and latency in over 90% areas of the country to meet the application demand of AI industry by 2020</td>
<td>- By 2020, the cloud computing infrastructure will be continuously optimized, and the PUE of newly built large cloud computing data center will be less than 1.5</td>
</tr>
<tr>
<td>- 2019.8 Guidance On the Construction of National New Generation Artificial Intelligence Innovation and Development Pilot Zone</td>
<td>- Exabyte (EB) level cloud storage system and other key technologies make breakthrough</td>
</tr>
<tr>
<td>- Strengthen the development of network, big data, and computer infrastructure</td>
<td>- 2017.1 Three-year action plan for cloud computing development (2017-2019)</td>
</tr>
<tr>
<td>- 20 AI innovation and development pilot zones will be established by 2023</td>
<td>- Put forward the development plan for the enhancement of cloud computing technology, industrial development and application promotion</td>
</tr>
<tr>
<td></td>
<td>- Aim to make breakthrough in key technologies and optimize infrastructure by 2019</td>
</tr>
<tr>
<td></td>
<td>- PUE value of the new data center should be lower than 1.4</td>
</tr>
<tr>
<td></td>
<td>- 2018.4 Declaration of Cooperation on Artificial Intelligence (AI)</td>
</tr>
<tr>
<td>- Increase investment and upgrade artificial intelligence infrastructure, and propose the establishment of a world-class European transnational institute for AI research, with a plan to increase investment by at least EUR 20 billion by the end of 2020</td>
<td>- 2012.9 Unlocking the potential of cloud computing in Europe</td>
</tr>
<tr>
<td>- 2018.6 Digital Europe 2021-2017</td>
<td>- By 2020, planning to increase direct investment in cloud computing technology research and development and infrastructure by EUR 45 billion, which is expected to generate a cumulative GDP of EUR 975 billion</td>
</tr>
<tr>
<td>- EUR 2.5 billion will be provided to fund AI development in data and algorithm storage, AI testing and national cooperation</td>
<td>- 2016.4 European Cloud Initiative</td>
</tr>
<tr>
<td></td>
<td>- With a total investment of EUR 6.7 billion, EU plans to strengthen the network platform for scientific research infrastructure by 2017, fully deploy and build high-performance computing centers and big data storage centers by 2020, and enhance technological innovation and supercomputing capabilities</td>
</tr>
</tbody>
</table>

**Figure 4.10 Policies of China and the EU to promote AI and cloud computing infrastructure**

Source: Desktop research, Roland Berger

---

\(^{89}\) European Commission, the Marie Skłodowska-Curie Actions, Horizon 2020, 2019
In terms of data sharing and application development, the EU government has launched the European Cloud Strategy in early 2012. This action aims to build a shared economy based on data and knowledge, allow free access to all resource from interdisciplinary researchers across the EU. Eventually this will unlock the research potential of interdisciplinary academia, improved the understanding of data-concentrated research and data sciences. The specific contents are as follows:

1) Creating a European Open Science Cloud that will offer Europe’s 1.7 million researchers a virtual environment to store, share and re-use their data across disciplines and borders by following the FAIR principle (Findable, Accessible, Interoperable and Reusable). The initial version of “European Open Science Cloud” online portal was officially launch in November 23, 2018 and became a key data source for the EU researcher;

2) Opening up by default all scientific data produced by future projects under the Horizon 2020 research and innovation programme, to ensure that the scientific community can re-use the enormous amount of data they generate.

Both the EU and China have made plans with clear targets, established management authorities, mobilized resources and created legal framework to promote the popularization and development of key technologies like cloud computing and AI.

In terms of making plans and setting targets, the State Council of China released the “New Generation Artificial Intelligence Development Plan” in 2017. The plan outlined a three-step strategy and laid out milestones to reach by 2020, 2025 and 2030. These include making major breakthroughs for fundamental AI theories, boosting the value of core AI industries to reach CNY150 billion and that of AI-related industries to reach CNY1 trillion; establishing a new legal framework that sets ethical principles for AI. Guided by the targets, China has achieved visible progress in AI: China ranks the first among all countries in terms of published papers in AI; many Chinese universities such as the Chinese Academy of Sciences, Tsinghua University and Peking University have set up their institute of AI; the value of core AI industries is approaching CNY 60 billion, well on track to reaching the target value.

90 The State Council, the Development Planning for a New Generation of Artificial Intelligence, 2017
Key measures taken by countries to promote the development of AI

Governments in China, the EU and Singapore have also set up management authorities and mobilized resources to achieve their goals. (Figure 4.11). For example, Singapore has formed "Smart Nation and Digital Government Group (SNDGG)", an umbrella institution that housed the various units and agencies involved in the Smart Nation initiative. SNDGG drives the digital transformation of government by supervising and guiding digital standards and computing power development. It also aims to build long-term capabilities for the public sector, and promotes adoption and participation from the public and industry, to take a collective approach in building a Smart Nation. The EU has also made efforts in these fronts.

**4.2.2 Governments should promote "green computing power"

“Green computing power” will help governments strike a balance between developing a smart society and addressing climate concerns. (Figure 4.12) Green and carbon neutral computing power can help other sectors to cut emission through computing power-based solutions, contributing to the energy efficiency of the whole society. According the GeSI SMARTer 2020 report,

---

91 International Telecommunication Union’s Global Enabling Sustainability Initiative (GeSI), SMARTer 2020 Report, 2018
account for 2.8% of global carbon emissions. But ICT-enabled solution will help cut emission by 7.8 billion tons, 15% of the global total in the year and 5 times of the sector’s emission.

Policy recommendations to promote "green computing power"

<table>
<thead>
<tr>
<th>Initiatives</th>
<th>Specific initiatives</th>
<th>Representative Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Policy planning</td>
<td>☑ Introduce the industry alliance to set energy saving targets, and establish the general consciousness of energy saving and emission reduction</td>
<td>· European Commission published the European Green Deal, aiming to reduce greenhouse gas emissions in order to make Europe climate-neutral by 2050, which addressed the EU would need a digital sector that “puts sustainability at its heart”.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>· Under the GSMA climate action roadmap for the mobile industry, Vodafone and other 50+ member companies committed to zero greenhouse gas emissions by 2050</td>
</tr>
<tr>
<td>Resources offering</td>
<td>☑ Set up a new energy investment plan to boost the scale of new energy with direct government investment</td>
<td>· Indonesian government invested about $36 billion of government funds to add 10.8 million kilowatts of renewable energy power generation in five years</td>
</tr>
<tr>
<td>Enhanced Supervision</td>
<td>☑ Establish a carbon emission trading system for energy, make it more expensive to exchange for limited emission quotas, and guide enterprises to take the initiative in emission reduction</td>
<td>· Germany has helped cut its carbon emissions by nearly 30% from 1990 levels through a cap-and-trade scheme</td>
</tr>
</tbody>
</table>

*Source: Chinese White Paper on Green Data Center, Desktop Research, Roland Berger*

Figure 4.12 Summary of policy recommendations to develop "green computing power"

In terms of making policy planning: To develop green computing power, governments can also mobilize leading companies in the field of cloud server and data centre to set industry-wide emission control target. In this regard, European Commission launched the European Green Deal, outlining a long list of policy initiatives that involves a variety of industries, aiming at putting Europe on track to reach net-zero global warming emissions by 2050. The deal specifically addressed digital sector, as the EU is would consider implementing actual policy to improve energy efficiency and circular economy performance "from broadband networks to data centres and ICT devices". Furthermore, Global System for Mobile Communications Association (GSMA)\(^\text{92}\) is among the first movers. It has raised an initiative to develop a mobile industry climate action roadmap in line with the Paris Agreement with an aim to achieve net-zero greenhouse gas emission by 2050 in line with the Paris Agreement. More than 50 GSMA operator members (including Vodafone, Telefonica, etc.) have joined the initiative and started to disclose their annual energy consumption, and emission control targets and tactics.

In terms of build infrastructure and providing public service: Governments can invest to develop renewable energies (wind, tidal, and geothermal energy, etc.) and encourage computing power infrastructure operators to rely more on renewable energies and cut their carbon footprint. Many countries already launched their plans. For instance, Indonesia introduced a five-year renewable energy investment plan in 2014. The government aimed to invest about US$ 36 billion in hydropower and geothermal power to raise its renewable energy generation capacity by 10.8 gigawatts, bringing the total

\(^{92}\) Global System for Mobile Communications Association
installed renewable energy capacity to 21.5 gigawatts by 2019. The benefits of the investment plan have started to show. In 2017, renewable energy accounted for 12.62% in Indonesia’s energy mix, well above its original target93.

In terms of enhancing legal framework and supervision: Governments can establish a sound carbon trading mechanism under which emission quota is sold at a high price and in limited quality. The government can also set the emission ceiling for each industry and each region so as to control the total emission level of the nation. For instance, the German Emissions Trading Authority at the German Environment Agency is dedicated to administering carbon trading activities, including allocating quotas, supervising auctions and issuing annual emission reports, etc. Thanks to its sound emission trading system, Germany has witnessed its emission reduced to 905 million tons, down by 27.7% from the 1990 level. China also launched carbon emission trading pilots Beijing, Shenzhen and Shanghai since 2012. These initiatives demand companies to step up their efforts in cutting emissions. Computing power-related companies can achieve their “green transformation” by using emerging technologies like AI to monitor and adjust their operations toward a greener way. As an example, Huawei cloud data centre in north China (the Langfang cloud data centre) is using Huawei’s iCooling thermal management solution to achieve intelligent cooling. The solution uses AI to collect actual IT loads and environment variables, incorporates deep learning to draw the appropriate correlations between various cooling equipment with the collected data, applies automatic inference using a deep neural network (DNN) model and determines the system parameters for the optimal PUE. It is in such an intelligent way that the equipment in the data centre (chillers, cooling towers and water pumps, etc.) are controlled and adjusted. After deploying the iCooling solution, the data centre’s annual PUE decreases by 0.116 to 1.304, significantly saving the data centre’s energy consumption.

4.2.3 Government can make policy planning and extending fiscal support to improve network quality

93 Chinese Economic (www.ce.cn), Indonesia Is Promoting the Development of Its Renewable Energy, 2018
To expand and improve network infrastructure, the government can, based on its goals, put in place a monitoring system that follow up on its progress both in terms of geographical coverage and in terms of internet connection speed. Germany, for example, issued the "Digital Strategy 2025" in 2016 and set the target to implement the strategy in four steps. To follow up on its progress, the German government regularly monitor the two key aspects in expanding its digital infrastructure: geographical coverage and bandwidth connection speed. In terms of geographical coverage, the government plans to roll out the digital infrastructure first from central residential areas and then extend to peripheral industrial zones; In terms of bandwidth connection speed, the goal is to provide at least 50Mbps broadband internet connection nationwide and develop 5G network. With effective monitoring, the digital infrastructure expansion is well on track: by the end of 2017, 75.5% of German households have access to internet connection with a download speed at least 50 Mbps.

Spain is another important. In 2009, Spain's National Authority for Markets and Competition launched the Wholesale Offer of Access to Registers and Ducts of Telefónica (MARCo) to expand the coverage of fibre-optic network, laying a good foundation to develop ultra-fast network and realize smart application scenarios. As the MARCo obliges all operators to share network access infrastructures, many international telecom operators started to enter the Spanish market and compete with Telefónica. The heated competition drove down deployment cost of new generation networks, particularly fibre-optic network, by 75%. By 2014, Spain had the highest fibre-to-the-home (FTTH) coverage in the EU. To further encourage operators to deploy fibre-optic networks to rural areas, where the deployment cost is relatively high, the Spanish government updated MARCo in 2016. The updated MARCo provides that new fibre-optic infrastructure constructed after 2016 are privately owned by operators and are not subject

---

**Policy recommendations for the development of network infrastructure**

<table>
<thead>
<tr>
<th>Specific initiatives</th>
<th>Representative cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Form a matrix monitoring system with geographical coverage and network speed layout to continuously promote the network construction process</td>
<td>· Germany carried out matrix monitoring on a regular basis to promote the orderly development of network construction, with “four-step” development goal</td>
</tr>
<tr>
<td>Adjust use right of network infrastructure to promote competition to push for better network coverage</td>
<td>· The Spanish government has promoted the construction of fibre-optic networks by adjusting the use right of infrastructure through the MARCO1</td>
</tr>
<tr>
<td>Adopt Public-Private Partnership model (PPP), use a small amount of government financial resources to attract social capital in network infrastructure</td>
<td>· The European Commission and the European Investment Bank invests 240 million euros, and then drive the social capital such as Cassa Depositi e Prestiti (CDP) to jointly invest and promote the bandwidth network infrastructure in the underdeveloped regions of the EU</td>
</tr>
</tbody>
</table>

Source: Desktop research, Roland Berger

---

1) MARCO is a Spanish governmental policy which obliging all telecom operators to provide access to vertical infrastructure within buildings
to mandatory sharing requirement. The updated policy effectively stimulated the telecom infrastructure construction in the sparsely deployed areas. By 2018, Spain's FTTH coverage reached 63%, 2.6 times of EU average.

Another policy measure is to apply the public-private partnership (PPP) model, where the government can leverage large amount of private fund by a small amount of fiscal input. In the "Digital Single Market 2025" issued in 2015, the EU set such strategic objectives: access to 1Gbps for all schools, transport hubs and main providers of public services and digitally intensive enterprises; and access to download speeds of at least 100 Mbps to be upgraded to 1Gbps for all European households. To deliver on the objectives, the EU Commission and Europe Investment Bank jointly established the Connecting Europe Broadband Fund and invested EUR 240 million to this the first investment platform under the European Fund for Strategic Investments. Later, private financing started to roll in from European banks like the German KfW\textsuperscript{94} and the Italian Cassa Depositi e Prestiti to support broadband infrastructure expansion in less developed areas in the single market. Another example is the OpenNet\textsuperscript{95}, a joint venture initiated by the Info-communications Development Authority of Singapore. It is responsible for building and operating an all-optical fibre network in Singapore. Previous operators like the Singapore Telecommunications, can lease transmission capacity from OpenNet for the provision of services. Using this model, the Singapore successfully attracted more than SGD 3bn of private capital with SGD 750 million of direct investment. The social resources mobilized are used to benefit the society by accelerating the construction of high-speed broadband network in Singapore, laying a solid foundation to realize such smart scenarios as online studying, video conference and remote healthcare.

4.2.4 Governments should use multiple measures to invigorate the market for developing diversified chip architectures

To create a more digital ecosystem capable of accommodating more diversified application scenarios, governments should encourage market stakeholders to develop chip architecture, devices and applications, form industrial alliance to bridge technical and talent resources in the upstream and downstream of value chains. It can also allocate earmarked fund and introduce enabling laws and regulations. (Figure 4.14).

\textsuperscript{94} KWF: Kreditanstalt Für Wiederaufbau, a stated-own German bank

\textsuperscript{95} OpenNet is a company responsible for equipment operation in Singapore
In terms of making policy planning: The government can help improve the computing power ecosystem by promoting the development of diversified chip architectures, such as the RISC architecture, represented by ARM. In the future, as computing power is required in more application scenarios, it is imperative to have more types of chip architectures in place. Realizing this, governments should, through policy planning, support the development of chip architectures such as the RISC architecture, set clear goals including market shares of architecture solutions in different stages. To form an industrial-wide virtuous cycle, the government should also mention in its planning the role of device manufacturer and application developers. The European Union, for example, has indirectly funded the European Processor Initiative (EPI). The aim of this project is to design and implement a roadmap for a new family of energy-efficient European processors for exascale computing, high-performance Big-Data and a range of emerging applications (such as autonomous driving). EPI has planned to launch its first-generation chip family in 2021, name Rhea, which will include Arm ZEUS architecture general purpose cores and prototypes of high energy-efficient accelerator tiles: RISC-V based (EPAC), Multi-Purpose Processing Array (MPPA), embedded FPGA (eFPGA) and cryptography HW engine. In addition, governments should step up efforts in building smart society so that the ARM architecture can be better used in more mobile smart scenarios. For instance, Council of London issued the Smart London Plan, which covers a series of
mobile scenarios like smart transportation and smart life. Such policy guidance will inspire more market players to participate in relevant industrial ecosystems.

In terms of building infrastructure and providing service: Governments can encourage stakeholders to form industrial alliance to bridge technical and talent resources in the upstream and downstream of the value chain. For instance, in April 2016, with the support of Ministry of Industry and Information Technology of China, the Green Computing Consortium was jointly established by leading global server and cloud computing enterprises, Chinese universities and research institutions. The Consortium will support the development of the RISC architecture ecosystem by pooling extensive industrial resources and talents. To be specific, by the end of October 2019, the Consortium has a total of 55 members that covers the entire RISC server value chain. There are traditional server manufacturers (HP, Dell, Huawei, Lenovo), internet companies (Alibaba, Baidu), chip manufacturers (ARM, Phytium, Hxt-semitech), and software companies (ChinaSoft, etc.). Marvell, as a manufacturer of ARM-based server processor, joined the Green Computing Industry Consortium this October. It has announced that it would contribute to key application development for big data, enterprise and cloud computing on platforms based on the ARM architecture. The Consortium also pools talents in the sector. For example, the Green Computing Consortium initiated the ARM HPC joint innovation project with Shanghai Jiaotong University in 2018. Under this project, experts from members of the Consortium are invited to give lectures and attend seminars on ARM HPC technologies, contributing to ARM 64-related scientific research and talent cultivation. EU’s Horizon 2020 programme funded EPI Consortium through a Framework Partner Agreement to implement EPI. EPI Consortium gathers 27 partners from 10 European countries to develop the processor and ensure that the key competence of high-end chip design remains in Europe. Members of EPI Consortium include, among others, top computing R&D institutions (ETH Zurich, E4 Computer Engineering), chip providers (Bull, Extoll), and automobile makers (BMW, Rolls-Royce).

In terms of enhancing legal framework and supervision: Governments can broaden market access for more to participate in the computing power ecosystem (chip design and manufacturing, devices and equipment, software, etc.) and streamline bureaucratic approval procedures. For example, in an effort to adapt to the fast-paced FinTech software industry, Singapore launched the FinTech Fast Track Initiative to accelerate patent approval and lower threshold for market access. From April 26, 2018, the approval time for FinTech patents was shortened to six months from the previous two years.

With comprehensive support from the government as mentioned above, the ecosystem of computing power will become more diversified and balanced. (Figure 4.15). The chip market will be shared by multiple architectures including X86, RISC and others. Different architecture will also have their own compatible devices, systems and applications. As such, computing power of chips based on all architectures need to be at full play.
An ideal, sufficiently diversified computing power ecosystem

Figure 4.15 A more diversified computing power ecosystem should feature multiple chip architectures with each of them supported by the whole industrial value chain.

4.2.5 Governments should tackle computing power safety concerns via policy planning and fiscal support

Governments should become better able to prevent computing power-related risks. On policy-making level, governments of different countries should come together and formulate international security standards; on the fiscal support front, governments can allocate designated fund for risk identification and warning by professional team; on the legal and regulatory level, governments should introduce a sound legal framework to protect the security of computing power. (Figure 4.16).
In terms of policy planning, governments of different countries should work together to formulate a set of international security standards, which will serve as the safety net for computing power to develop across the world. The security framework of computing power has four dimensions: network, software, device and chip (Figure 4.17). Each dimension needs a universal set of security standards for support.

At the chip dimension, ISO (International Standardization Organization) and IEC (International Electrotechnical Commission) have jointly formulated multiple sets of standards including ISO/IEC 9798 and ISO/IEC 11770 to offer guidance for entity authentication, secret key management and encryption algorithms. For instance, by September 2018, ISO-9796 RSA encryption algorithm has been the most influential public key encryption algorithm and has withstood all cryptographic attacks as of the date.

At the equipment dimension, SHARE\textsuperscript{96} initiated by IBM has developed the SHARE78 international standard of disaster recovery and covered in it seven tiers of practical disaster recovery plans, including equipment security, recovery and backup, and log audit.

At the software dimension, IEEE (the Institute of Electrical and Electronic Engineers) issued the Standard Dictionary of Measures of the Software Aspects of Dependability, with which users may understand and evaluate the dependability of a software.

At the network dimension, 3GPP (3rd Generation Partnership Project) published SA3-5G security standard. The standard suggests using secret key-based technical solution to ensure network slice security, satisfying the differentiated security demand in various application scenarios. A universal set of standards will eliminate the need to customize solutions for different regional markets, thus facilitating the evolution of the computing power ecosystem.

\textsuperscript{96} SHARE is an independent information technology association initiated by IBM
Computing power security architecture

<table>
<thead>
<tr>
<th>Network</th>
<th>Security content</th>
<th>Norms and mechanisms</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Access authentication</td>
<td>Establishment of international safety standards</td>
</tr>
<tr>
<td></td>
<td>Access control</td>
<td>2 Regular notification of security vulnerabilities</td>
</tr>
<tr>
<td></td>
<td>Key management</td>
<td>3 Construction of criminal responsibility mechanism</td>
</tr>
<tr>
<td>Software</td>
<td>Vulnerability management</td>
<td>4 Construction of third-party supervision mechanism</td>
</tr>
<tr>
<td></td>
<td>Data isolation</td>
<td>5 Establishment of security and privacy protection regulations</td>
</tr>
<tr>
<td></td>
<td>Patch management</td>
<td></td>
</tr>
<tr>
<td>Device</td>
<td>Device security</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Disaster recovery backup</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Log audit</td>
<td></td>
</tr>
<tr>
<td>Chip</td>
<td>Data encryption</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Random number self-check</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Safe startup</td>
<td></td>
</tr>
</tbody>
</table>

Source: Desktop research, Roland Berger

Figure 4.17 Governments should improve the computing power security architecture to prevent risks

In terms of extending fiscal support, government can allocate earmarked fund, establish security monitoring authority and engage third-party players to facilitate continuous monitoring – all can help identify and prevent potential security risks. In Germany, every year, the federal and local governments invest EUR 45 million and EUR 5 million respectively to support the Helmholtz Centre for Information Security to combat the increasing risks of cyber-crime. The fund supports about 800 industry experts to carry out academic research on network security and privacy risk, including developing cognitive system of semantic recognition technology to identifying false information and spotting the threat intelligence. The academic achievements of the centre can effectively protect data generated by automation systems, autonomous driving systems and medical statistics. They can also be used to detect potential risks in a transaction and upend cybercrime economy.

In terms of enhancing legal framework and supervision, governments can establish a sound accountability mechanism so that practitioners will pay attention to and comply with security regulations on computing power. In Germany, Federal Data Protection Act provides that unauthorized transmission to third party personal information of large group of individuals, or make such information publicly available for commercial purposes will lead to legal sanctions and penalties. Furthermore, the German Cyber Security Law clearly provides that network operators will be fined if they fail to fix security loopholes promptly after receiving risk warnings from the Federal Office for Information Security.

To conclude, governments should be fully prepared to tackle the challenges in developing computing power by making policy planning, building infrastructure and providing services, extending fiscal support, enhancing legal framework and supervision, etc. Only by doing so can a nation stand ready to embrace the future smart society and occupy the command heights of national comprehensive power competition.
GENERAL DISCLAIMER

THE INFORMATION IN THIS DOCUMENT MAY CONTAIN PREDICTIVE STATEMENT, INCLUDING BUT NOT LIMITED TO, STATEMENTS REGARDING FUTURE FINANCIAL RESULTS, OPERATING RESULTS, FUTURE PRODUCT PORTFOLIOS, AND NEW TECHNOLOGIES. THERE ARE A NUMBER OF FACTORS THAT COULD CAUSE ACTUAL RESULTS AND DEVELOPMENTS TO DIFFER MATERIALLY FROM THOSE EXPRESSED OR IMPLIED IN THE PREDICTIVE STATEMENTS. THEREFORE, SUCH INFORMATION IS PROVIDED FOR REFERENCE PURPOSES ONLY AND CONSTITUTES NEITHER AN OFFER NOR A COMMITMENT. HUAWEI MAY CHANGE THE INFORMATION AT ANY TIME WITHOUT NOTICE, AND IS NOT RESPONSIBLE FOR ANY LIABILITIES ARISING FROM YOUR USE OF ANY OF THE INFORMATION PROVIDED HEREIN.