

BY HAI JIN/HUAZHONG UNIVERSITY OF SCIENCE AND TECHNOLOGY, HAIBO CHEN/SHANGHAI JIAO TONG UNIVERSITY, HONG GAO/HARBIN INSTITUTE OF TECHNOLOGY, XIANG-YANG LI/UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA, SONG WU/HUAZHONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Cloud Bursting for the World's Largest Consumer Market

CLOUD INFRASTRUCTURE IS information technology consisting of various hardware resources and software technologies. It enables ubiquitous access to shared pools of configurable system resources and higher-level services that can be delivered with minimal management effort, often through the Internet. Cloud infrastructure today is a critical platform for many applications, providing basic support for the development of emerging areas, including big data, the Internet of Things (IoT),

and artificial intelligence (AI). In 2016, International Data Corporation's *Cloud Computing Survey* reported cloud technology is becoming a staple of organization infrastructure, as 70% of organizations have at least one application in the cloud, and 56% of organizations are still identifying IT operations as candidates for cloud hosting.¹ In 2017, IDC predicted that by 2021, spending on cloud infrastructure and cloud-supported hardware, software, and services would double to more than \$530 billion.²

For China, the overall market for cloud computing in 2016 was 51.49 billion RMB (China's currency), with an overall annual growth rate of 35.9% in 2016, which was significantly greater than the global average. It is expected that the China cloud computing market will continue to grow significantly over the next two years, reaching 136.6 billion RMB by 2020.³

The development and popularization of cloud computing, especially in emerging domains, brings great convenience. It also poses new challenges involving design and construction of modern cloud infrastructure. Cloud computing in China is also quite different from other countries, as it includes special requirements for the related infrastructure.

Here, we first explore the background of cloud infrastructure in China, then turn to its characteristics, and conclude with an outlook on development.

Background

Greatest number of netizens and IT employees. As of December 2017, the number of netizens in China was 772 million, or 55.8% of the total Chinese population, the number of online shoppers was 533 million, with annual growth of 14.3% over 2016, and the number of online payers was 531 million, including 527 million who pay through their smartphones.⁴ As of January 2018, the total number of employees in computer, communications, and other electronic-equipment-man-



The 13th Five-Year Plan identified cloud computing as an important emerging national strategic industry.

ufacturing industries in China was 8.264 million.⁵ Development of the cloud infrastructure in China involves both challenges and opportunities to meet the needs of such a large number of users and developers.

Best cellular infrastructure. The cellular infrastructure in China facilitates development of cloud infrastructure and services. As of September 2017, the total number of base stations in China was 6.04 million, including 4.47 million 3G and 4G base stations.²⁴ The base stations covered over 95% of the country, including even small villages with only dozens of residents. Moreover, as of December 2017, the number of smartphone netizens in China was 753 million and rising, with an annual growth rate of 8.2% over 2016, reflecting deep penetration of the mobile Internet.²⁵ As of November 2017, there were 3.91 million active smartphone apps available to Chinese consumers, along with more than 2.24 million local third-party apps, surpassing the number of apps (1.78 million) provided by Apple's app store in China.⁴ Some unicorn apps, including Didi Taxi for a Chinese taxi-hailing service, create demand for unprecedented computing power.

Greatest demand for applications. With regard to construction of the cloud infrastructure, the most significant difference between China and other countries is that China has the largest and fastest-growing application demand due to having the greatest number of users. For example, the number of WeChat "HongBao"^a sending and receiving activities on Chinese New Year's Eve in 2014 was 16 million, followed by a dramatic increase to 14.2 billion in 2017.⁶ There were 46 billion "HongBao" activities over six days during the 2017 Spring Festival, with a 43.3% increase over 2016.⁷ Another example of dramatic growth is that for the official website of the China Railway Corporation, "12306," the average number of page views per day was 55.67 billion and 81.34 billion during the peak period, and more than 10.2 million train tickets were

sold through the site on January 11, 2018.⁸ Moreover, on the 2017 "11.11"^b sales day, peak database throughput for Alibaba was 472 million per second and the peak transactions traffic was 256,000 per second, a 2.1× increase over 2016.⁹

Innovation in the mobile Internet industry. China has seen rapid innovation in information industries and services in recent years. In terms of smart transportation, Didi Taxi, for example, had 450 million users and 20 billion daily planning requests and processed 4,500TB data per day in 2017. By accessing information from traffic cameras on smart traffic lights and applying cloud and big data analysis, Didi Taxi was able to improve its activity by 10%–20% in 2017.¹⁰ Meanwhile, the number of active users of shared-bike applications was 221 million, or 28.6% of all Chinese netizens, as of February 2017.⁴

In digital gaming, as of December 2017, market penetration of mobile games in China was 76.1%, with each game user installing 3.35 mobile game apps on average.¹¹ In May 2017, more than 200 million people were playing "King Glory," a popular mobile game in China, with 54.1 million playing it daily.¹² The Chinese Electronic Commerce Research Center reported that during the "11.11" Shopping Festival of 2017, the total online trading volume in China was 253.97 billion RMB, a 45% increase over the same period in 2016. Moreover, major e-commerce enterprises reported 850 million express logistics orders on November 11, 2017, with courier volume during this "11.11" Shopping Festival of more than 1.5 billion, historic peak levels.¹³ The data shows the cloud infrastructure in China is able to serve billions of users concurrently.

Characteristics

Cloud infrastructures worldwide share certain characteristics, including resource-on-demand, elasticity, and geo-distribution. Due to the deep penetration of the mobile Internet and proliferation of mobile apps in China, China's cloud infrastructure

a HongBao is a red envelope with money inside traditionally given by older people to young people as a gift during important festivals; it went digital via WeChat and Alipay.

b 11.11 is an online shopping carnival (such as Taobao and TMall) run annually by Alibaba on November 11.

largely centers on the app ecosystem, as characterized in the following ways:

(Super) app-oriented infrastructure.

Consider the top two largest cloud service providers in China: Alibaba Cloud and Tencent Cloud. Alibaba Cloud originally sought to address the need for massive computing resources for the “11.11” Shopping Festival, with 90% of sales from mobile apps like Taobao and TMall. Tencent Cloud sought to address the computing infrastructure for Tencent’s flagship mobile apps: QQ and WeChat. Its cloud infrastructure thus centered on such super apps, serving hundreds of millions of people daily. The cloud infrastructure cannot be developed simply by reusing open source cloud stacks like OpenStack due to the need of unparalleled concurrency from such apps. To this end, major cloud service providers in China tend to build their own infrastructures, with heavy optimizations tailored for their super apps.

WeChat’s infrastructure evolution.

Here, we consider the evolution of WeChat as an example of how Tencent builds and customizes its cloud infrastructure.¹⁸ WeChat’s initial goal was to develop a message-oriented chat app, with messages synchronized between a sender and its receivers. China is thus the key market worldwide. To satisfy this extremely large-scale requirement, Tencent customized its own Remote Procedure Call library (called Svrkit), an infrastructure pillar for connected distributed planetary services for WeChat. A key design decision in such synchronization was how to propagate messages. Many such designs have adopted the “pull” mode, whereby receivers pull messages from a sender. In contrast, WeChat adopted a “push” propagation mode, because there was a strict upper limit in chat groups, originally 20, later increased to 100 and today 500. The cost for push propagation is bounded, and the receiver, or each WeChat app, can deliver much lower latency. Due to having to serve a rapidly growing user population, WeChat initially adopted a micro-serviced architecture, including aggressive division of functionalities of business representation layer into multiple logic servers (logicSrv); even the same func-

tionality with different priority is divided. For example, message-delivery functionality is divided into three service modules—message synchronization, voice and text message sending, and figures and videos sending—allowing out-scalability with increased numbers of users.

With its customized infrastructure, WeChat grew from a chat app to a massive digital payment system; “Hong-Bao” can be viewed as a special kind of payment, an enterprise business platform, and a development-and-delivery platform (WeChat Mini apps). With its developing ecosystem, the WeChat platform could become not only a cloud platform itself but also attract a large number of third-party apps to access its services on Tencent Cloud, providing seamless integration and winning strong user loyalty.

Scalable and hybrid infrastructure for bursty loads. Many super apps exhibit strong bursty loads, especially under the extra demand on special days or during certain seasons. The cloud infrastructure needs to not only be able to scale up easily but also scale in afterward. Unlike Amazon and Google, which mainly deploy services in their own large-scale datacenters, major cloud service providers in China tend to rent existing datacenters to scale out their services and cloud-based systems, in addition to their own datacenters. These providers rent datacenters because a larger number of small- to medium-size datacenters were built during the IT revolution of the 2000s but had relatively low utilization. For example, as of 2017, there were more than one million datacenters in China, but most were small, each occupying less than 500 square meters.²¹ Cloud service providers usually rent datacenters to quickly scale up their services under bursty loads, then scale in to avoid possibly wasting the infrastructure cost. To allow quick deployment of services, they usually built their customized infrastructure to allow quick deployment of services and increase resource utilization.

Technologies behind Alibaba’s “11.11” Shopping Festival. The unprecedented peak transactions per second requires technologies that are not only from the off-the-shelf open source stack. Alibaba, for example,

uses its own complete customized software stack, from infrastructure software to cloud software.¹⁹ Its current world-record TPC-C result is approximately 500,000 transactions per second,²² while it handles up to 42 million operations/second in its database involving 325,000 and 256,000 NewOrder and Payment Transactions per second.^c To this end, Alibaba has created its own open-source database (called OceanBase) and distributed file system (called TFS). In order to meet the resource requirements of peak traffic, Ali Cloud is able to expand capacity by 100,000 servers in one hour. To quickly deploy such services, it created Pouch, a customized container framework and aggressively deployed and scheduled online services with offline services through its Sigma scale-out scheduler. While hybrid cloud has been advocated for years, Alibaba has pioneered seamless integration of its public cloud with the datacenters of its partners to deliver one-stop handling of individual transactions among multiple service providers.

Deeply integrated/optimized infrastructure. Like other technology giants Amazon, Microsoft, and Google, the growth of cloud scale makes efficiency a top optimization target, as even a single-digit increase in utilization or performance density could save tens of millions of U.S. dollars. This motivates cloud service providers in China to provide deeply integrated and optimized infrastructure.

China’s cloud infrastructure is moving toward hardware/software co-design to improve efficiency and flexibility. Huawei Cloud, another very large cloud service provider in China, publically released its *Service Driven Infrastructure* plan in 2014, including software-defined storage and software-defined networking.²⁰ This allowed offloading key processing functionalities in hardware while retaining software flexibility. Ali Cloud recently released its X-Dragon Cloud Server to aggressively offload VM management services, as well as

^c Note real-world transactions are much more complex than TPC-C, as each user-facing transaction generates a large number of transactions.

customized data-processing services, to a tightly coupled physical installation. It also recently announced it would produce its own neural processing units for AI-related tasks. And UCloud, a top-five cloud service provider, announced its release of near-data-processing infrastructure for big-data applications, yielding improved efficiency.

Outlook


China's cloud infrastructure has made great strides, supporting large-scale applications and millions of users. The rapid development of cloud infrastructure has been promoted both through national research projects and through the corporations involved. The Chinese central and local governments now plan to push development of cloud computing while mainstream enterprises pursue a new round of cloud computing designs.

The government's plan for developing cloud computing. The Chinese central government is emphasizing development of cloud computing and its underlying infrastructure. For example, the 13th Five-Year Plan identified cloud computing as an important emerging national strategic industry.¹⁴ And the Ministry of Industry and Information Technology of China adopted a Three-Year Development Plan for cloud computing, 2017 to 2019, aiming to increase the cloud computing industry in China to 430 billion RMB by 2019.¹⁵ The Chinese central government is also funding a series of projects for cloud computing. In 2017, the "Cloud Computing and Big Data" Special Program of the National Key Research and Development Plan launched 15 projects with total funding of 409 million RMB.¹⁶ In 2018, it plans to start 20 projects with a total budget up to 625 million RMB.¹⁷

Enterprises' plan for developing cloud computing. Chinese enterprises are developing an increasingly powerful cloud infrastructure to provide competitive cloud computing products and services. For example, Inspur and Sugon launched a series of scientific projects to research key technologies in cloud datacenters and servers. And Alibaba expects to use its cloud

unit to carry it through the next decade. According to a Gartner research report in September 2017, Ali Cloud has surpassed Google in IaaS Public Cloud Service and is today the third largest cloud provider in the world.²³ In 2015, Tencent adopted its "Cloud Plus" plan to develop Tencent Cloud, which will invest 10 billion RMB to build a cloud platform and ecosystem over the next five years. Meanwhile, Huawei has established a new business group dedicated to developing Huawei Cloud.

Emerging computing paradigms and cloud computing. Information technology is evolving quickly. Emerging computing paradigms like AI, the IoT, and Cloud-Edge computing have begun to influence the cloud infrastructure and offer opportunities for addressing cloud-related challenges. Machine- and deep-learning algorithms and models for AI are relevant for cloud computing researchers and practitioners. On the one hand, the cloud can benefit from machine and deep learning to support more smart resource management. On the other, machine- and deep-learning requires large-scale computing power, and the cloud is an essential platform for hosting AI services due to its potential for high scalability and ready access to computing resources.

With the rapid development of the mobile Internet and IoT applications in China, the existing centralized cloud computing architecture faces significant challenges. Edge computing is being investigated as a way to better exploit capabilities at the edge of the network to support the IoT. In edge computing, the massive amount of data generated by different kinds of IoT devices can be processed at the network edge instead of having to first transmit it to the centralized cloud infrastructure due to bandwidth- and energy-consumption concerns. Edge computing can thus provide services with quicker response and greater quality compared to traditional cloud infrastructure and is more suitable for being integrated with IoT to provide more efficient and secure services for a vast number of end users. 

Further Reading

1. 2016 IDC Cloud Computing Survey; [https://www.idg.com/tools-for-marketers/2016-idg-enterprise-cloud-](https://www.idg.com/tools-for-marketers/2016-idg-enterprise-cloud-computing-survey/)

2. IDC FutureScape: Worldwide IT Industry 2018 Predictions; <https://www.idc.com/getdoc.jsp?containerId=US43171317>
3. White Paper on Cloud Computing Development in China (2017); <http://www.fx361.com/page/2018/0112/2686558.shtml>
4. The 41st China Statistic Report on Internet Development; <http://www.cnnic.net.cn/hlwzfzj/hlwzbg/hlwzbg/201803/P020180305409870339136.pdf>
5. China Entrepreneur Investment Club; <https://www.ceicdata.com/zh-hans/china/no-of-employee-by-industry-monthly/no-of-employee-computer-communication-other-electronic-equipment>
6. Tencent Tech: HongBao War; <http://new.qq.com/omn/20180215/20180215C0E1MO.html>
7. 2017 WeChat Spring Festival Data Report; <http://tech.qq.com/a/20170203/010341.htm>
8. China Railway Site Sees 5.93 Billion Clicks Per Hour as Busiest Travel Season Starts; <https://technode.com/2018/01/16/chunyun-data/>
9. Alibaba Tech: Fight Peak Data Traffic on 11.11: The Secrets of Alibaba Stream Computing; https://medium.com/@alitech_2017/how-to-cope-with-peak-data-traffic-on-11-11-the-secrets-of-alibaba-stream-computing-17d5e807980c
10. 2017 Annual Urban Transportation Report. DidiChuxing; <http://index.caixin.com/upload/didi2017.pdf>
11. JiGuang. 2017 Mobile Gaming Market Research Report; <https://community.jiguang.cn/t/topic/24810>
12. JiGuang. 2017 King Glory Research Report; <https://www.jiguang.cn/reports/72>
13. 2017 '11.11' E-Commerce Platform Shopping Festival Evaluation Report; http://www.100ec.cn/zt/upload_data/17sh11bg.pdf
14. The development plan of the 13th Five-Year Plan national strategic emerging industry; http://www.gov.cn/zhengce/content/2016-12/19/content_5150090.htm
15. The Three-year Development Plan of Cloud Computing (2017–2019); <http://www.miit.gov.cn/n1146290/n4388791/c5570594/content.html>
16. 2017 Project List of 'Cloud Computing and Big Data' Special Projects in The National Key Research and Development Plan; <http://app.myzaker.com/news/article.php?pk=59a4e2d41bc8e03727000029>
17. 2018 guide for projects of 'Cloud Computing and Big Data' Special Projects in The National Key Research and Development Plan; http://www.stdaily.com/kjzc/top/2017-10/10/content_582554.shtml
18. The evolution of WeChat Infrastructure; <http://www.infoq.com/cn/articles/the-road-of-the-growth-weixin-background>
19. Techniques behind TMall's 11.11 Shopping Festival; <https://jaq.alibaba.com/community/art/show?articleid=1201>
20. Huawei SDI Innovation Architecture; <http://www.cnetnews.com.cn/2014/0918/3034037.shtml>
21. Analysis of 2017 China Datacenter Sector Development and Evolution; <http://www.chyxx.com/industry/201709/564441.html>
22. http://www.tpc.org/tpcc/results/tpcc_results.asp
23. IaaS Public Cloud Service Market Share; <https://www.channel2e.com/news/gartner-public-cloud-iaas-market-share-amazon-aws-microsoft-azure-google-growth/>
24. Number of base stations in China; <http://tech.sina.com.cn/rol/2017-10-22/doc-ifymzksi0587142.shtml>
25. Data analysis on the number of smartphone users nationwide; <http://www.chinabgao.com/k/zhinenshouji/28395.html>

Hai Jin is the Cheung Kung Scholar Chair Professor at Huazhong University of Science and Technology, Wuhan.

Haibo Chen is a professor and Director of the Institute of Parallel and Distributed Systems at Shanghai Jiao Tong University, Shanghai.

Hong Gao is a professor at Harbin Institute of Technology, Harbin.

Xiang-Yang Li is a professor and Executive Dean of the School of Computer Science and Technology at the University of Science and Technology of China, Hefei.

Song Wu is a professor and Director of the Institute of Parallel and Distributed Computing at Huazhong University of Science and Technology, Wuhan.

© 2018 ACM 0001-0782/18/11 \$15.00.