

DATA TRANSMISSION METHOD AND SYSTEM, AND DEVICE AND STORAGE MEDIUM

[0001] This application claims priority to Chinese Patent Application No. 2023101619874, filed with the China National Intellectual Property Administration on February 21, 2023 and entitled "DATA TRANSMISSION METHOD AND SYSTEM, AND DEVICE AND STORAGE MEDIUM", which is incorporated herein by reference in its entirety.

TECHNICAL FIELD

[0002] The present disclosure relates to the field of internet technologies, and in particular, to a data transmission method, a device, a storage medium, and a system.

BACKGROUND

[0003] With the continuous emergence of various cloud service products such as a cloud desktop, cloud rendering, and a cloud game, a user can perform activities such as working and entertainment by using these cloud service products at lower costs and in a more flexible and convenient manner without purchasing an expensive host machine device having higher computing power, and without installing various required application programs on a local host machine. Especially, for some application programs that consume much computing power, only corresponding services provided by cloud service providers need to be used. The cloud service providers have a large quantity of hardware resources, and may virtualize a large quantity of virtual machines by using a virtualization technology and the like. Various different cloud services, for example, the cloud desktop, the cloud rendering, and the cloud game, are deployed on the virtual machines for users to use.

[0004] To use the cloud service, the user needs to install a corresponding client on the local host machine, to remotely access the cloud service through a transmission link (namely, a transmission link between the host machine in which the client is located and the virtual machine in which the cloud service is located) between the client and a corresponding cloud server (that is, a cloud service deployed on a cloud). During actual application, data transmitted between the client and the cloud server may be related to a plurality of different types, including, for example, audio data, peripheral data, image data, and file copy data.

[0005] From a perspective of user experience, the user has different sensitivity to the data. For example, a speed of copying a file is not so obvious to sensual experience of the user. However, if a delay of the audio data is slightly long, the user can clearly hear stuttering, plosive sound, and the like, which brings bad experience to the user.

SUMMARY

[0006] Embodiments of the present disclosure provide a data transmission method, a device, a storage medium, and a system. By adapting to a network status of one transmission link between a client and a cloud server, transmission of different types of data between the client and the cloud server is dynamically scheduled, thereby improving bandwidth utilization and ensuring user experience.

[0007] According to a first aspect, an embodiment of the present disclosure provides a data transmission method, applied to a first device. The method includes: obtaining a plurality of types of data that the first device needs to send to a second device, where one of the first device and the second device corresponds to a cloud server, and the other of the first device and the second device corresponds to a client; correspondingly buffering, according to a plurality of transmission priorities corresponding to the plurality of types of data, the plurality of types of data in a plurality of queues respectively corresponding to the plurality of transmission priorities; obtaining a network quality indicator value of a transmission link between the first device and the second device in response to a current scheduling period being started; determining, according to the network quality indicator value, a duration of the current scheduling period and a data amount that needs to be transmitted in the current scheduling period; and obtaining, according to the transmission priorities respectively corresponding to the plurality of queues, data not exceeding the data amount from the plurality of queues, and sending the obtained data to the second device through the transmission link within the duration.

[0008] According to a second aspect, an embodiment of the present disclosure provides a data transmission apparatus, located in a first device. The apparatus includes: a data buffering module, configured to obtain a plurality of types of data that the first device needs to send to a second device, and correspondingly buffer, according to a plurality of transmission priorities corresponding to the plurality of types of data, the plurality of types of data in a plurality of queues respectively corresponding to the plurality of transmission priorities, where one of the first device and the second device corresponds to a cloud server, and the other of the first device and the second device corresponds to a client; a network detection module, configured to obtain a network quality indicator value of a transmission link between the first device and the second device in response to a current scheduling period being started; and a scheduling processing module, configured to determine, according to the network quality indicator value, a duration of the current scheduling period and a data amount that needs to be transmitted in the current scheduling period; and obtain, according to the transmission priorities respectively corresponding to the plurality of queues, data not exceeding the data amount from the plurality of queues, and send the obtained data to the second device through the transmission link within

the duration.

[0009] According to a third aspect, an embodiment of the present disclosure provides an electronic device, including a memory, a processor, and a communication interface. The memory stores executable codes, and the executable codes, when executed by the processor, causes the processor to perform the data transmission method according to the first aspect.

[0010] According to a fourth aspect, an embodiment of the present disclosure provides a non-transitory machine-readable storage medium. The non-transitory machine-readable storage medium stores executable codes. The executable codes, when executed by a processor of an electronic device, cause the processor to implement at least the data transmission method according to the first aspect.

[0011] According to a fifth aspect, an embodiment of the present disclosure provides a data transmission system, including: a first device and a second device. One of the first device and the second device corresponds to a cloud server, and the other of the first device and the second device corresponds to a client. The first device is configured to: obtain a plurality of types of data that the first device needs to send to the second device; correspondingly buffer, according to a plurality of transmission priorities corresponding to the plurality of types of data, the plurality of types of data in a plurality of queues respectively corresponding to the plurality of transmission priorities; obtain a network quality indicator value of a transmission link between the first device and the second device in response to a current scheduling period being started; determine, according to the network quality indicator value, duration of the current scheduling period and a data amount that needs to be transmitted in the current scheduling period; and obtain, according to the transmission priorities respectively corresponding to the plurality of queues, data not exceeding the data amount from the plurality of queues, and send the obtained data to the second device through the transmission link within the duration.

[0012] The data transmission method provided in this embodiment of the present disclosure is applicable to an application scenario in which data transmission is performed between a cloud server and a corresponding client. In an implementation, one of the first device and the second device with one transmission link established there between corresponds to the cloud server, and the other of the first device and the second device corresponds to the client. For example, if the cloud server is deployed in the first device, the corresponding client is deployed in the second device.

[0013] The data transmission method provided in this embodiment of the present disclosure may be performed by a transmission control service running at an application layer. In the scenario in which data transmission is performed between the first device and the second device

through one transmission link, the transmission control service may be separately deployed in the first device and the second device. The transmission control service deployed in the first device is configured to perform transmission control on a process in which the first device sends data to the second device, and the transmission control service deployed in the second device is configured to perform transmission control on a process in which the second device sends data to the first device. In other words, the transmission control service may run in a device in which the client is located, or may run in a device in which the cloud server is located. In addition, the transmission control service runs at the application layer rather than a network layer or a transport layer.

[0014] A data transmission process is: the transmission control service running at the application layer in the first device obtains a plurality of types of data that the first device needs to send to the second device, and correspondingly buffers, according to a plurality of transmission priorities corresponding to the plurality of types of data, the plurality of types of data in queues respectively corresponding to the plurality of transmission priorities, where the plurality of queues are also located at the application layer. The transmission control service obtains a network quality indicator value of a transmission link between the first device and the second device in response to a current scheduling period being started, and determines, according to the network quality indicator value, duration of the current scheduling period and a data amount that needs to be transmitted in the current scheduling period. Then, the transmission control service obtains, according to the transmission priorities respectively corresponding to the plurality of queues, data not exceeding the data amount from the plurality of queues, and sends the obtained data to the second device through the transmission link within the duration.

BRIEF DESCRIPTION OF THE DRAWINGS

[0015] To describe the technical solutions in embodiments of the present disclosure more clearly, the following briefly describes accompanying drawings required for describing the embodiments. Apparently, the accompanying drawings in the following description show some embodiments of the present disclosure, and a person of ordinary skill in the art may still derive other drawings from these accompanying drawings without creative efforts.

[0016] FIG. 1 is a flowchart of a data transmission method according to an embodiment of the present disclosure.

[0017] FIG. 2 is a flowchart of a data transmission method according to an embodiment of the present disclosure.

[0018] FIG. 3 is a flowchart of a data transmission method according to an embodiment of

the present disclosure.

[0019] FIG. 4 is a flowchart of a data transmission method according to an embodiment of the present disclosure.

[0020] FIG. 5 is a schematic diagram of application of a data transmission method according to an embodiment of the present disclosure.

[0021] FIG. 6 is a schematic structural diagram of a data transmission apparatus according to an embodiment of the present disclosure.

[0022] FIG. 7 is a schematic structural diagram of an electronic device according to an embodiment of the present disclosure.

DETAILED DESCRIPTION

[0023] To make the objectives, technical solutions, and advantages of the embodiments of the present disclosure clearer, the following clearly and completely describes the technical solutions in the embodiments of the present disclosure with reference to the accompanying drawings in the embodiments of the present disclosure. Apparently, the described embodiments are some but not all of the embodiments of the present disclosure. All other embodiments obtained by a person of ordinary skill in the art that are based on the embodiments of the present disclosure without creative efforts shall fall within the protection scope of the present disclosure. In addition, a step sequence in each of the following method embodiments is merely an example, but not a strict limitation.

[0024] A data transmission solution provided in an embodiment of the present disclosure is applicable to a scenario in which data transmission is performed between a cloud server and a client. The cloud server may be, for example, a cloud desktop or various cloud applications. The cloud applications may include, for example, cloud rendering or a cloud game.

[0025] The cloud desktop is a computing-based cloud desktop service, and provides a user with an easy-to-use, secure, and efficient cloud desktop office system by using a desktop as a service (Desktop as a Service, DaaS for short). The cloud desktop runs on a cloud. For example, several virtual machines are constructed on the cloud, and different cloud desktops (which may also be referred to as cloud computers) are built on different virtual machines, to be provided for different users to use. A user connects a thin client or a conventional user terminal (for example, a notebook computer) to the cloud desktop through a public network or a private line.

[0026] Similarly, the cloud application is also an application software running on the cloud. A client of the user does not need to install the application software, and may establish a transmission link with the cloud application on the cloud by using a communication protocol, to use the cloud application.

[0027] An adaptive streaming protocol (Adaptive Streaming Protocol, ASP for short) is a communication protocol, and may be used for data transmission between the cloud desktop on the cloud and the client.

[0028] For the cloud desktop or the cloud application, during actual application, data that needs to be transmitted between the client and the cloud server may include different types of data such as graphic data, audio data, and peripheral data. In a conventional situation in which a transmission link is established between the client and the cloud server, from a perspective of user experience, a user has different sensitivity to the data. For example, a speed of copying a file is not so obvious to sensual experience of the user. However, if a delay of the audio data is slightly long, the user can clearly hear stuttering, plosive sound, and the like, which brings bad experience to the user. Therefore, generally, priorities of these different types of data may be divided.

[0029] In a transmission solution, a plurality of transmission links may be established between the client and the cloud server, and different transmission links are configured with different quality of service (Quality of Service, QoS for short), to respectively transmit data with a plurality of transmission priorities. This transmission solution is implemented at a network layer and a transport layer. Although implementation is simple, there is an application limitation.

[0030] In an implementation, first, transmission links of a plurality of transport layers need to be established, and the transmission links are maintained by a gateway, which reduces a quantity of concurrent connections of the gateway, that is, reduces a quantity of simultaneously connected users. In short, one user occupies a plurality of transmission links, and a total quantity of transmission links that can be simultaneously maintained by the gateway is set. Consequently, a quantity of users that can simultaneously access the gateway is reduced. Second, according to a definition of QoS at the network layer, there is usually only the QoS or no QoS, that is, only two quality of service can be supported. Therefore, only two transmission priorities are supported. In addition, in a self-established network, the QoS may be set freely. However, in a public cloud, QoS setting is controlled by an operator and is not open for free, and the network link is complex, which is generally not feasible in engineering. In other words, a network operator needs to establish transmission links with different QoS between different clients and cloud servers, resulting in high complexity of engineering implementation.

[0031] In view of this, in the embodiment of the present disclosure, data with different transmission priorities is transmitted between the cloud server and the client through only one transmission link. In an implementation, transmission scheduling may be dynamically

performed on the data with different transmission priorities with reference to a dynamic change of a network status of the transmission link. When it is ensured that data with a high transmission priority is transmitted preferentially, it is necessary to ensure that a bandwidth is fully used, and transmission of data with a low transmission priority is considered.

[0032] A data transmission method provided in an embodiment of the present disclosure is applicable to an application scenario in which data transmission is performed between a cloud server and a corresponding client. It is assumed that the cloud server and the client are deployed in a first device and a second device, that is, one of the first device and the second device corresponds to the cloud server, and the other of the first device and the second device corresponds to the client. For example, if the cloud server is deployed in the first device, the corresponding client is deployed in the second device. The first device may be a cloud virtual machine, and the second device may be a thin client or a user terminal such as a notebook computer or a PC.

[0033] The data transmission method provided in this embodiment of the present disclosure may be performed by a transmission control service running at an application layer in the first device and the second device. A transmission link established between the client and the cloud server may be understood as a transmission link established between the first device and the second device.

[0034] The transmission control service deployed in the first device is configured to perform transmission control on a process in which the first device sends data to the second device, and the transmission control service deployed in the second device is configured to perform transmission control on a process in which the second device sends data to the first device. In other words, the transmission control service may run in a device in which the client is located, or may run in a device in which the cloud server is located. In addition, the transmission control service runs at the application layer rather than a network layer or a transport layer.

[0035] FIG. 1 is a flowchart of a data transmission method according to an embodiment of the present disclosure. The method may be performed by a transmission control service running at an application layer in a first device. As shown in FIG. 1, the method includes the following steps.

[0036] 101: Obtain a plurality of types of data that the first device needs to send to a second device, and correspondingly buffer, according to a plurality of transmission priorities corresponding to the plurality of types of data, the plurality of types of data in a plurality of queues respectively corresponding to the plurality of transmission priorities, where one of the first device and the second device corresponds to a cloud server, and the other of the first device

and the second device corresponds to a client.

[0037] 102: Obtain a network quality indicator value of a transmission link between the first device and the second device in response to a current scheduling period being started.

[0038] 103: Determine, according to the network quality indicator value, a duration of the current scheduling period and a data amount that needs to be transmitted in the current scheduling period.

[0039] 104: Obtain, according to the transmission priorities respectively corresponding to the plurality of queues, data not exceeding the data amount from the plurality of queues, and send the obtained data to the second device through the transmission link within the duration.

[0040] Actually, in a bidirectional transmission process in which the client transmits data to the cloud server and the cloud server transmits data to the client, the user sensitivity varies for different types of data.

[0041] For example, in a process in which the cloud server sends data to the client, the data involves a plurality of types, such as audio data, graphic data, and print data, where the transmission priorities corresponding to the three types sequentially decrease. In a process in which the client sends data to the cloud server, the data involves a plurality of types, such as keyboard and mouse data and copy data, where the transmission priorities corresponding to the two types sequentially decrease.

[0042] When the first device corresponds to the cloud server, and the second device corresponds to the client, in the example above, the plurality of types of data that the first device needs to send to the second device include audio data, graphic data, and print data. When the first device corresponds to the client, and the second device corresponds to the cloud server, in the example above, the plurality of types of data that the first device needs to send to the second device include keyboard and mouse data and copy data.

[0043] For ease of description, in this embodiment of the present disclosure, a process in which the first device transmits data to the second device is described, and it is assumed that the first device corresponds to the cloud server, and the second device corresponds to the client.

[0044] After the transmission link between the first device and the second device is established, the transmission control service running at the application layer in the first device may grab, in real time, data that the cloud server running in the first device needs to send to the client, determine a transmission priority corresponding to the currently obtained data according to a data type corresponding to the obtained data and a preset correspondence between different data types and different transmission priorities, and may further buffer, according to a time sequence, the currently obtained data in a queue corresponding to the transmission priority.

[0045] During actual application, the client and the cloud server may set corresponding transmission priorities by default according to types of data to be respectively transmitted by the client and the cloud server, and the user may further adjust transmission priorities corresponding to different types of data.

[0046] Because the transmission control service runs at the application layer, and a plurality of queues corresponding to a plurality of set transmission priorities are maintained by the transmission control service, the plurality of queues are also located at the application layer.

[0047] In an implementation, the plurality of set transmission priorities may be classified into two types. One type is provided with a corresponding buffer queue, and the other type is not provided with a corresponding buffer queue. For ease of description, it is assumed that a first-type transmission priority is represented as P0, and a second-type transmission priority includes P1, P2, and P3. A priority of P0 is higher than other transmission priorities. Based on the example, in an implementation, when the plurality of types of obtained data that the first device needs to send to the second device include data corresponding to a target transmission priority P0, the data corresponding to the target transmission priority P0 may be directly sent to the second device through the transmission link between the first device and the second device, and data corresponding to other transmission priorities (for example, P1, P2, and P3) is correspondingly buffered in corresponding queues according to the respective corresponding transmission priorities.

[0048] It can be learned that data with the highest transmission priority may be preferentially sent to the transmission link for transmission without waiting, and data with other transmission priorities first enters corresponding queues for buffering, and is sent to the transmission link for transmission in subsequent different scheduling periods.

[0049] In an implementation, each time new data is added to a queue, a scheduling process may be triggered. For example, when data is currently stored in at least one of a plurality of set queues, it is determined whether a previous scheduling period ends. If the previous scheduling period ends, it is determined to start a current scheduling period. If the previous scheduling period does not end, processing may end or a timer duration may be set. The timer duration may be set to a difference between current time and end time of the previous scheduling period, that is, set to trigger a scheduling process again when the previous scheduling period ends.

[0050] When it is determined to start the current scheduling period, a current network quality indicator value of the transmission link between the first device and the second device is obtained, and a duration of the current scheduling period and a data amount that needs to be transmitted in the current scheduling period are determined according to the network quality

indicator value.

[0051] In this embodiment of the present disclosure, a scheduling period does not have a fixed duration, but dynamically changes. In other words, the scheduling durations of different scheduling periods are different, and a new scheduling period cannot be started before its previous scheduling period ends. For example, a duration of a corresponding scheduling period is correspondingly set according to a dynamic change of a network status of the transmission link.

[0052] In this embodiment of the present disclosure, the network quality indicator value used for averaging the network status of the transmission link may be an indicator such as a round-trip time (Round-Trip Time, RTT for short) that can reflect a network congestion status.

[0053] Actually, the transmission control service may detect an RTT value of the transmission link at a relatively short collection frequency. In an implementation, assuming that it is determined, at a moment T1, that the current scheduling period is started, a target RTT value whose collection time is closest to the moment T1 may be determined from a plurality of collected RTT values, and the target RTT value is used as a basis for determining the duration of the current scheduling period and the data amount that needs to be transmitted in the current scheduling period. Alternatively, in an implementation, in addition to the target RTT value, a plurality of RTT values collected before the target RTT value may be further obtained, and an average value of the RTT values is used as a basis for determining the duration of the current scheduling period and the data amount that needs to be transmitted in the current scheduling period.

[0054] The foregoing uses only an example of the RTT indicator for description. Actually, other indicators used for evaluating a network status are similar.

[0055] For the duration of the current scheduling period, in summary, a better network quality currently corresponding to the transmission link indicates a shorter duration of the current scheduling period. On the contrary, a poorer network quality currently corresponding to the transmission link indicates a longer duration of the current scheduling period.

[0056] In addition to determining the duration of the current scheduling period, the data amount that needs to be transmitted in the current scheduling period further needs to be determined. In summary, a better network quality currently corresponding to the transmission link indicates a larger data amount that needs to be transmitted in the current scheduling period. On the contrary, a poorer network quality currently corresponding to the transmission link indicates a smaller data amount that needs to be transmitted in the current scheduling period.

[0057] In this embodiment of the present disclosure, after the current scheduling period is

started, it is intended to control a data sending pacing according to a current network quality status, that is, control a data sending bit rate. However, actually, the transmission control service at the application layer cannot directly control the sending bit rate, but can control two factors that can affect the sending bit rate, to indirectly control the sending bit rate. The two factors are the duration of the current scheduling period and the data amount that needs to be sent in the current scheduling period.

[0058] In summary, when the network quality is good, the sending bit rate may be improved. On the contrary, when the network quality is poor, the sending bit rate may be reduced. When the network quality is good, data may be sent faster, but sending time needs to be short, otherwise network congestion occurs quickly. When the network quality is poor, data is sent slower, and the sending time is long, so that there is longer time for recovering the network quality.

[0059] To achieve the objective, when the network quality is good, a shorter duration of the current scheduling period and a larger data amount that needs to be transmitted may be configured, to implement a higher sending bit rate. On the contrary, when the network quality is poor, a longer duration of the current scheduling period and a smaller data amount that needs to be transmitted may be configured, to reduce the sending bit rate.

[0060] In an implementation, a mapping relationship between different value ranges of the network quality indicator value and a duration of a scheduling period as well as a data amount that needs to be transmitted in the scheduling period may be preset. The duration of the current scheduling period and the data amount that needs to be transmitted in the current scheduling period are determined according to the mapping relationship.

[0061] After the duration of the current scheduling period and the data amount that needs to be transmitted in the current scheduling period are determined, data not exceeding the data amount may be transmitted in the current scheduling period through the transmission link between the first device and the second device. The transmitted data is taken out from a plurality of queues corresponding to a plurality of transmission priorities according to a sequence of the transmission priorities. Start time of the current scheduling period is the time (assuming to be $T1$) when the current scheduling period is initiated, and end time $T2$ is determined according to the start time and the duration of the current scheduling period.

[0062] For example, it is assumed that the plurality of queues includes a queue 1, a queue 2, and a queue 3 whose corresponding transmission priorities are in descending order. For ease of description, it is assumed that the determined data amount that needs to be transmitted in the current scheduling period is $S1$.

[0063] Assuming that data buffered in the queue 1 has exceeded $S1$, in an implementation, data whose size is $S1$ may be taken out from the queue 1 according to a sequence of storage time, and the data is sent to the second device through the transmission link between the first device and the second device. Assuming that data buffered in the queue 1 does not reach $S1$, in an implementation, the data (assuming that its data amount is $S2$) stored in the queue 1 may be taken out from the queue 1, and the data is sent to the transmission link. In addition, data whose data amount is $S3$ is taken out from the queue 2, and the data is sent to the second device through the transmission link, where $S2+S3 \leq S1$.

[0064] In addition, after the current scheduling period is started and the duration of the current scheduling period is determined, start time of a next scheduling period may be further set according to the end time of the current scheduling period, that is, the start time of the next scheduling period is set as the end time $T2$ of the current scheduling period.

[0065] In conclusion, in the data transmission solution provided in this embodiment of the present disclosure, only one transmission link needs to be established between the first device and the second device (namely, between the client and the cloud server), division of transmission priorities of data that needs to be transmitted between the two devices and transmission scheduling of the data with different transmission priorities at the application layer do not rely on settings of a network operator, and engineering implementation is relatively simple. In addition, only one transmission link needs to be established between two devices corresponding to one user. In comparison with a plurality of transmission links, a larger quantity of concurrently connected users can be supported. In addition, transmission priorities of data are divided at the application layer, so that a plurality of transmission priorities may be set according to an actual requirement of a user. Compared with control of only two quality of service in QoS at the network layer, more flexible and abundant priority control can be implemented. Finally, the duration of each scheduling period and the data amount that needs to be transmitted are dynamically adjusted according to the network quality status of the transmission link between the first device and the second device, so that the sending pacing can be adaptively controlled according to a dynamic change status of network quality, to preferentially transmit data with a high priority, so as to ensure user experience and improve network bandwidth utilization.

[0066] FIG. 2 is a flowchart of a data transmission method according to an embodiment of the present disclosure. The method may be performed by a transmission control service running at an application layer in a first device. As shown in FIG. 2, the method includes the following steps.

[0067] 201: Obtain a plurality of types of data that the first device needs to send to a second device, and correspondingly buffer, according to a plurality of transmission priorities corresponding to the plurality of types of data, the plurality of types of data in queues respectively corresponding to the plurality of transmission priorities, where one of the first device and the second device corresponds to a cloud server, and the other of the first device and the second device corresponds to a client.

[0068] 202: Obtain a network quality indicator value of a transmission link between the first device and the second device in response to a current scheduling period being started.

[0069] 203: Determine, according to a comparison result between the network quality indicator value and at least one of set thresholds and with reference to a preset correspondence between different threshold ranges and different transmission parameters, a duration of the current scheduling period and a data amount that needs to be transmitted in the current scheduling period, where the different transmission parameters include different scheduling period durations and different data amounts that need to be transmitted in scheduling periods.

[0070] 204: Obtain, according to the transmission priorities respectively corresponding to the plurality of queues, data not exceeding the data amount from the plurality of queues, and send the obtained data to the second device through the transmission link within the duration.

[0071] In this embodiment, when the current scheduling period is started, after the network quality indicator value of the transmission link between the first device and the second device is obtained, the duration of the current scheduling period and the data amount that needs to be transmitted in the current scheduling period may be determined according to the comparison result between the network quality indicator value and the set thresholds.

[0072] In an implementation, a threshold may be set, so that a threshold range less than the threshold and another threshold range greater than the threshold may be obtained. If two thresholds are set, three threshold ranges may be obtained.

[0073] Taking the RTT value as the network quality indicator value as an example, it is assumed that according to a set threshold, a first threshold range less than the threshold and a second threshold range greater than the threshold are obtained. In this case, the network quality reflected by the first threshold range is better than the network quality reflected by the second threshold range.

[0074] For the first threshold range, first scheduling period duration is preset and a data amount that needs to be transmitted in a scheduling period is preset to a first data amount. For the second threshold range, second scheduling period duration is preset and a data amount that needs to be transmitted in a scheduling period is preset to a second data amount. The first

scheduling period duration is less than the second scheduling period duration, and the first data amount is greater than the second data amount, or the first scheduling period duration is equal to the second scheduling period duration, and the first data amount is greater than the second data amount. In other words, a better network quality indicates a shorter scheduling period and a larger data amount transmitted in a scheduling period.

[0075] A situation of dividing three threshold ranges according to two thresholds is similar to the situation of the two threshold ranges. In short, it is assumed that a threshold range a, a threshold range b, and a threshold range c in ascending order are obtained according to two set thresholds. In this case, it is assumed that a network quality status reflected by the threshold range a is: A network is good. A network quality status reflected by the threshold range b is: The network is slightly congested. A network quality status reflected by the threshold range c is: The network is seriously congested.

[0076] For example, a scheduling period duration set for the threshold range a may be 5 milliseconds, and a data amount that needs to be transmitted in a scheduling period is 10×1200 bytes. A scheduling period duration set for the threshold range b may be 5 milliseconds, and a data amount that needs to be transmitted in a scheduling period is 5×1200 bytes. A scheduling period duration set for the threshold range c may be 10 milliseconds, and a data amount that needs to be transmitted in a scheduling period is 1200 bytes.

[0077] In this embodiment, a plurality of network quality statuses are set according to the threshold range to which the network quality indicator value belongs, and different scheduling period durations and data amounts that need to be transmitted in scheduling periods are set for the network quality statuses. The data is buffered in a corresponding transmission priority queue at the application layer, and the sending pacing is controlled for different network quality statuses, instead of queuing data at the network layer.

[0078] For implementation of other steps in this embodiment, reference may be made to the related descriptions in the foregoing other embodiments. Details are not described herein.

[0079] FIG. 3 is a flowchart of a data transmission method according to an embodiment of the present disclosure. The method may be performed by a transmission control service running at an application layer in a first device. As shown in FIG. 3, the method includes the following steps.

[0080] 301: Obtain a plurality of types of data that the first device needs to send to a second device, and correspondingly buffer, according to a plurality of transmission priorities corresponding to the plurality of types of data, the plurality of types of data in queues respectively corresponding to the plurality of transmission priorities, where one of the first

device and the second device corresponds to a cloud server, and the other of the first device and the second device corresponds to a client.

[0081] 302: Obtain a network quality indicator value of a transmission link between the first device and the second device in response to a current scheduling period being started.

[0082] 303: Determine, according to the network quality indicator value, a duration of the current scheduling period and a data amount that needs to be transmitted in the current scheduling period.

[0083] 304: Determine, for each of the plurality of queues, a proportion of the data amount respectively according to transmission priorities respectively corresponding to the plurality of queues and waiting durations of data in the plurality of queues, and send the data obtained from the plurality of queues according to the proportions to the second device through the transmission link within the duration according to a set sequence of the transmission priorities.

[0084] In this embodiment, in a process in which data not exceeding a target data amount is taken out from the plurality of queues according to the transmission priorities corresponding to the plurality of queues in the current scheduling period, and the data is sent to the second device through the transmission link with the second device, in addition to considering the transmission priorities corresponding to the queues, waiting durations of data in each queue may be further considered. The target data amount is a data amount that needs to be transmitted in the current scheduling period.

[0085] For example, it is assumed that two queues of a queue 1 and a queue 2 are set, and a transmission priority corresponding to the queue 1 is higher than a transmission priority corresponding to the queue 2. In an implementation, in the current scheduling period, if data stored in the queue 1 has reached the target data amount, the data of the target data amount may be obtained only from the queue 1, and the data is sent to the transmission link and sent to the second device. If a data amount buffered in the queue 1 is less than the target data amount and there is data buffered in the queue 2, data is read from the queue 2.

[0086] In another implementation, if the data stored in the queue 1 has reached the target data amount, but if a waiting duration of some data stored earlier in the queue 2 is excessively long at this time, it may be determined that data of a first proportion of the target data amount is obtained from the queue 1 and data of a second proportion of the target data amount is obtained from the queue 2. The data of the first proportion obtained from the queue 1 is preferentially sent to the transmission link for transmission, and then the data of the second proportion obtained from the queue 2 is transmitted.

[0087] During actual application, a waiting duration threshold and a data amount proportion

allocation rule may be set to implement step 304.

[0088] For implementation of other steps in this embodiment, reference may be made to the related descriptions in the foregoing other embodiments. Details are not described herein.

[0089] FIG. 4 is a flowchart of a data transmission method according to an embodiment of the present disclosure. The method may be performed by a transmission control service running at an application layer in a first device. As shown in FIG. 4, the method includes the following steps.

[0090] 401: Obtain target data that the first device needs to send to a second device, and store the target data in a queue corresponding to a target transmission priority according to a target data type corresponding to the target data.

[0091] One of the first device and the second device corresponds to a cloud server, and the other of the first device and the second device corresponds to a client. Data transmitted between the client and the cloud server have a plurality of data types, different data types correspond to different transmission priorities, and different queues for buffering data are set for different transmission priorities.

[0092] When the target data that the first device needs to send to the second device is obtained at a current moment, the target data type corresponding to the target data and the target transmission priority corresponding to the target data type are determined, and the target data is stored in the queue corresponding to the target transmission priority.

[0093] 402: Determine whether a previous scheduling period ends; and if the previous scheduling period ends, execute step 403, or if the previous scheduling period does not end, set next scheduling time as end time of the previous scheduling period.

[0094] 403: Determine that a current scheduling period is started.

[0095] In this embodiment, when there are data stored in a plurality of set queues, an execution of a scheduling process may be triggered. First, it is determined whether the previous scheduling period ends, and if the previous scheduling period ends, it is directly determined to start the current scheduling period, that is, current time is determined as start time of the current scheduling period. If the previous scheduling period does not end yet, the next scheduling time is set as the end time of the previous scheduling period.

[0096] During actual application, a dynamic timer may be set to dynamically start a scheduling task.

[0097] In an implementation, if the previous scheduling period does not end yet, the timer duration may be set as a difference between the current time and the end time of the previous scheduling period, that is, how long does the previous scheduling period end. When time

reaches the end time of the previous scheduling period, step 402 may be performed again. In this case, the current scheduling period is to be started.

[0098] In addition, if scheduling period start time that is set according to the end time of the previous scheduling period is reached, it may also be determined to start the current scheduling period.

[0099] 404: Obtain a network quality indicator value of a transmission link between the first device and the second device.

[0100] 405: Determine, according to the network quality indicator value, a duration of the current scheduling period and a data amount that needs to be transmitted in the current scheduling period.

[0101] 406: Set start time of a next scheduling period according to end time of the current scheduling period.

[0102] After the duration of the current scheduling period and the data amount that needs to be transmitted in the current scheduling period are determined according to the currently obtained network quality indicator value, the start time of the next scheduling period may be further set as the end time of the current scheduling period. In an implementation, the timer duration may be set to complete the setting.

[0103] In this way, after the current scheduling period ends, an execution of a next scheduling process may be entered. Actually, in the next scheduling process, the determining in step 402 may also be first performed, and a determining result in this case is that the previous scheduling period has ended. In this case, the previous scheduling period is a scheduling period started before the determining moment.

[0104] 407: Obtain, according to transmission priorities respectively corresponding to a plurality of queues, data not exceeding the data amount from the plurality of queues, and send the obtained data to the second device through the transmission link within the duration.

[0105] In this embodiment, it is assumed that the target data that the first device needs to send to the second device is obtained at a moment T1, and the target data is stored in a corresponding queue for buffering. In this case, because new data is stored in the queue, determining of whether the previous scheduling period ends is triggered. End time T2 of the previous scheduling period is determined when the previous scheduling period is started. If the current moment T1 is earlier than the moment T2, it indicates that the previous scheduling period does not end yet, and determining may be performed until the moment T2. If the moment T1 is after the moment T2, it indicates that the previous scheduling period has ended. In this case, it is determined to start the current scheduling period, that is, it is determined that the start time of

the current scheduling period is the moment T1.

[0106] Then, the network quality indicator value corresponding to the transmission link between the first device and the second device is obtained at this time, and the duration of the current scheduling period and the data amount that needs to be transmitted in the current scheduling period are determined according to the network quality indicator value.

[0107] Then, the start time of the next scheduling period may be set as end time T3 of the current scheduling period. In addition, in the current scheduling period, the data not exceeding the data amount is obtained from the plurality of queues according to a sequence of the transmission priorities, and the data is sent to the second device.

[0108] Based on this, it may be understood that if there are no new data stored in the queue before the moment T3 is reached, when the moment T3 is reached, it may be determined to start the next scheduling period. If there are new data stored in the queue before the moment T3, determining of "whether the previous scheduling period ends" is triggered. In this case, because the current scheduling period defined from the moment T1 to the moment T3 does not end yet, next determining needs to be performed only after the current scheduling period ends. In this case, the determining result is that the previous scheduling period ends, and the next scheduling period is started.

[0109] During actual application, it is assumed that the duration of the current scheduling period is 5 milliseconds, and the data amount that needs to be transmitted is 10*1200 bytes. Actually, the data of 10*1200 bytes may be completely sent within only 1 millisecond, no other data is transmitted within the remaining 4 milliseconds. If there are no new data stored in the plurality of queues within the 4 milliseconds, because it is determined that "the previous scheduling period ends", the next scheduling period may be started after the 4 milliseconds.

[0110] If there are new data stored in a queue within the 4 milliseconds, determining of "whether the previous scheduling period ends" is triggered at this time, and the determining result is "the previous scheduling period does not end yet". The previous scheduling period is the current scheduling period at this time, that is, a scheduling period started before the determining moment. Therefore, the next scheduling period is not started immediately, and the next scheduling period needs to be started after the current scheduling period ends.

[0111] FIG. 5 is a schematic diagram of application of a data transmission method according to an embodiment of the present disclosure. In FIG. 5, it is assumed that a plurality of set transmission priorities include four transmission priorities of P0, P1, P2, and P3 in descending order, buffer queues corresponding to the three transmission priorities of P1, P2, and P3 are set, and there is no buffer queue corresponding to the transmission priority of P0. All the queues

are at the application layer and are maintained by the transmission control service at the application layer.

[0112] Based on the settings above, when grabbing data (where a rectangular block shown in the figure represents one piece of data) that the first device needs to send to the second device, the transmission control service first determines whether a transmission priority corresponding to the data is P0, if the transmission priority corresponding to the data is P0, directly sends the data to the transmission link and preferentially transmits the data to the second device, or if the transmission priority corresponding to the data is not P0, stores the data in a queue corresponding to the corresponding transmission priority. In this way, the data with the transmission priority P0 may be sent to the second device in time, to ensure timeliness of transmission of the data. For example, if audio data sent by the cloud server to the client is set to have the transmission priority P0, the audio data sent by the cloud server to the client is preferentially sent through the transmission link, thereby ensuring user experience. Similarly, if the client transmits data to the cloud server, keyboard and mouse data may be set to have the transmission priority P0, and the keyboard and mouse data is sent to the cloud server in time for responding in time.

[0113] After the data is stored in a queue, it is determined whether the previous scheduling period ends. Assuming that the previous scheduling period has ended, it is determined to start the current scheduling period, then the network quality indicator value of the transmission link is obtained, and a network quality status: good, slightly congested, or seriously congested is determined according to the network quality indicator value, so that the duration of the current scheduling period and the data amount that needs to be transmitted are set according to the network quality status, to control the data sending bit rate.

[0114] For example, as shown in FIG. 5, when the network is good, a short scheduling period length and a large data transmission amount may be set. When the network is slightly congested, a relatively short scheduling period length and a relatively large data transmission amount may be set. When the network is seriously congested, a relatively long scheduling period length and a relatively small data transmission amount may be set.

[0115] In addition, after the duration of the current scheduling period is determined, the start time of the next scheduling period may be set as the end time of the current scheduling period. For example, timing time of the timer is set as the duration of the current scheduling period.

[0116] It may be understood that for the queues of the three transmission priorities of P1, P2, and P3, data with the same transmission priority is stored in a queue corresponding to the same transmission priority according to a sequence of entering time.

[0117] After the duration of the current scheduling period and the data amount that needs to be transmitted are set, data may be obtained from a plurality of queues based on a set polling algorithm, and the data is transmitted to the transmission link. The obtained data amount does not exceed a set data amount. Only the sequence of the transmission priorities may be considered in the polling algorithm, or the waiting durations of data in the queues may be considered.

[0118] In the data transmission solutions provided in the foregoing embodiments, only one transmission link needs to be established between the client and the cloud server that communicate with each other, and transmission priorities of different types of data are distinguished at the application layer, and buffering and transmission control are performed on the different types of data based on the transmission priorities corresponding to the different types of data. This does not rely on the network operator, and is highly feasible on a public cloud. For different network congestion situations, the data is buffered in the queue corresponding to the transmission priority at the application layer, and the sending pacing is controlled instead of queuing data at the network layer, to preferentially transmit data with a high transmission priority. Scheduling period duration and the amount of sent data are dynamically adjusted according to the network status, so that dynamical adjustment can be performed in time according to the network status, to fully use network bandwidth.

[0119] The following describes a data transmission apparatus of one or more embodiments of the present disclosure in detail. A person skilled in the art may understand that all the apparatuses can be formed by configuring market-selling hardware components through steps instructed in this solution.

[0120] FIG. 6 is a schematic structural diagram of a data transmission apparatus according to an embodiment of the present disclosure. The apparatus is located in a transmission control service running at an application layer in a first device. As shown in FIG. 6, the apparatus includes: a data buffering module 11, a network detection module 12, and a scheduling processing module 13.

[0121] The data buffering module 11 is configured to obtain a plurality of types of data that the first device needs to send to a second device, and correspondingly buffer, according to a plurality of transmission priorities corresponding to the plurality of types of data, the plurality of types of data in a plurality of queues respectively corresponding to the plurality of transmission priorities, where one of the first device and the second device corresponds to a cloud server, and the other of the first device and the second device corresponds to a client.

[0122] The network detection module 12 is configured to obtain a network quality indicator

value of a transmission link between the first device and the second device in response to a current scheduling period being started.

[0123] The scheduling processing module 13 is configured to determine, according to the network quality indicator value, a duration of the current scheduling period and a data amount that needs to be transmitted in the current scheduling period; and obtain, according to the transmission priorities respectively corresponding to the plurality of queues, data not exceeding the data amount from the plurality of queues, and send the obtained data to the second device through the transmission link within the duration.

[0124] In an implementation, the scheduling processing module 13 is specifically configured to: determine, according to a comparison result between the network quality indicator value and at least one of set thresholds and with reference to a preset correspondence between different threshold ranges and different transmission parameters, the duration of the current scheduling period and the data amount that needs to be transmitted in the current scheduling period, where the different transmission parameters include different scheduling period durations and different data amounts that need to be transmitted in scheduling periods.

[0125] In an implementation, the different threshold ranges include a first threshold range and a second threshold range, and a network quality reflected by the first threshold range is better than a network quality reflected by the second threshold range. A transmission parameter corresponding to the first threshold range includes a first scheduling period duration and a first data amount, and a transmission parameter corresponding to the second threshold range includes a second scheduling period duration and a second data amount. The first scheduling period duration is less than the second scheduling period duration, and the first data amount is greater than the second data amount. Alternatively, the first scheduling period duration is equal to the second scheduling period duration, and the first data amount is greater than the second data amount.

[0126] In an implementation, the scheduling processing module 13 is specifically configured to: determine, for each of the plurality of queues, a proportion of the data amount respectively according to the transmission priorities respectively corresponding to the plurality of queues and waiting durations of data in the plurality of queues; and send the data obtained from the plurality of queues according to the proportions to the second device through the transmission link within the duration of the current scheduling period according to a set sequence of the transmission priorities.

[0127] In an implementation, the data buffering module 11 is specifically configured to: if the plurality of types of data include data corresponding to a target transmission priority, directly

send the data corresponding to the target transmission priority to the second device through the transmission link, where the target transmission priority is higher than other transmission priorities in the plurality of transmission priorities; and correspondingly buffer data other than the data corresponding to the target transmission priority in the plurality of types of data in corresponding queues according to the respective corresponding transmission priorities.

[0128] In an implementation, the scheduling processing module 13 is further configured to: set start time of a next scheduling period according to end time of the current scheduling period.

[0129] In an implementation, the network detection module 12 is further configured to: if scheduling period start time that is set according to end time of a previous scheduling period has been reached, determine to start the current scheduling period.

[0130] In an implementation, the network detection module 12 is further configured to: determine, in response to storing data in at least one of the plurality of queues, whether a previous scheduling period ends; and determine to start the current scheduling period if the previous scheduling period ends.

[0131] In an implementation, the cloud server includes any one of the following: a cloud desktop and a cloud application.

[0132] According to the foregoing solutions, in a process in which the first device transmits the plurality of types of data to the second device, different transmission priorities may be divided according to sensitivity of a user to the plurality of types of data, and corresponding queues for buffering the data are configured for the different transmission priorities. The period length and the data transmission amount of a scheduling period are dynamically set based on the dynamically changed network quality indicator value of the transmission link between the first device and the second device, to control the data sending pacing at the application layer, so that data that has a high transmission priority and that is sensitive to the user is preferentially transmitted, and bandwidth utilization is also ensured. Transmission of data with a low transmission priority is not excessively delayed for transmission of data with the high transmission priority, thereby ensuring user experience.

[0133] The apparatus shown in FIG. 6 may perform the steps in the foregoing embodiments. For detailed execution processes and technical effects, reference may be made to the descriptions in the foregoing embodiments. Details are not described herein again.

[0134] In a possible design, the structure of the data transmission apparatus shown in FIG. 6 may be implemented as an electronic device. As shown in FIG. 7, the electronic device may include a processor 21, a memory 22, and a communication interface 23. The memory 22 stores executable code, and the executable code, when executed by the processor 21, causes the

processor 21 to implement at least the data transmission method provided in the foregoing embodiments.

[0135] In addition, an embodiment of the present disclosure provides a non-transitory machine-readable storage medium. The non-transitory machine-readable storage medium stores executable codes, the executable codes, when executed by a processor of an electronic device, cause the processor to implement at least the data transmission method provided in the foregoing embodiments.

[0136] The apparatus embodiments described above are merely exemplary. The units described as separate parts may or may not be physically separate. Some or all of the modules may be selected according to actual needs to achieve the objectives of the solutions of the embodiments. A person of ordinary skill in the art may understand and implement the solutions without creative efforts.

[0137] Through the description of the foregoing implementations, a person skilled in the art may clearly understand that the implementations may be implemented by software in addition to a necessary universal hardware platform, and may certainly be implemented by a combination of hardware and software. Based on such an understanding, the foregoing technical solutions essentially, or the part contributing to existing technologies, may be implemented in a form of a software product. The present disclosure may use a form of a computer program product that is implemented on one or more computer-usable storage media (including but not limited to a disk memory, a CD-ROM, an optical memory, and the like) that include computer-usable program codes.

[0138] Finally, it should be noted that the foregoing embodiments are merely intended for describing the technical solutions of the present disclosure, but not for limiting the present disclosure. It should be understood by a person of ordinary skill in the art that although the present disclosure has been described in detail with reference to the foregoing embodiments, modifications can be made to the technical solutions described in the foregoing embodiments, or equivalent replacements can be made to some technical features in the technical solutions, As long as such modifications or replacements do not cause the essence of corresponding technical solutions to depart from the spirit and scope of the technical solutions of the embodiments of the present disclosure.

CLAIMS

1. A data transmission method, applied to a first device, and the method comprising:

obtaining a plurality of types of data that the first device needs to send to a second device, wherein one of the first device and the second device corresponds to a cloud server, and the other of the first device and the second device corresponds to a client;

correspondingly buffering, according to a plurality of transmission priorities corresponding to the plurality of types of data, the plurality of types of data in a plurality of queues respectively corresponding to the plurality of transmission priorities;

obtaining a network quality indicator value of a transmission link between the first device and the second device in response to a current scheduling period being started;

determining, according to the network quality indicator value, a duration of the current scheduling period and a data amount that needs to be transmitted in the current scheduling period; and

obtaining, according to the transmission priorities respectively corresponding to the plurality of queues, data whose total amount does not exceed the data amount from the plurality of queues, and sending the obtained data to the second device through the transmission link within the duration.

2. The method according to claim 1, wherein the determining, according to the network quality indicator value, the duration of the current scheduling period and the data amount that needs to be transmitted in the current scheduling period comprises:

determining, according to a comparison result between the network quality indicator value and at least one of set thresholds and with reference to a preset correspondence between different threshold ranges and different transmission parameters, the duration of the current scheduling period and the data amount that needs to be transmitted in the current scheduling period, wherein the different transmission parameters comprise different scheduling period durations and different data amounts that need to be transmitted in scheduling periods.

3. The method according to claim 2, wherein the different threshold ranges comprise a first threshold range and a second threshold range, and a network quality reflected by the first threshold range is better than a network quality reflected by the second threshold range; and

a transmission parameter corresponding to the first threshold range comprises a first scheduling period duration and a first data amount, and a transmission parameter corresponding

to the second threshold range comprises a second scheduling period duration and a second data amount, wherein

the first scheduling period duration is less than the second scheduling period duration, and the first data amount is greater than or equal to the second data amount; or the first scheduling period duration is equal to the second scheduling period duration, and the first data amount is greater than the second data amount.

4. The method according to any one of claims 1 to 3, wherein the obtaining, according to the transmission priorities respectively corresponding to the plurality of queues, data not exceeding the data amount from the plurality of queues, and sending the obtained data to the second device through the transmission link within the duration comprises:

determining, for each of the plurality of queues, a proportion of the data amount respectively according to the transmission priorities respectively corresponding to the plurality of queues and waiting durations of data in the plurality of queues; and

sending the data obtained from the plurality of queues according to the proportions to the second device through the transmission link within the duration of the current scheduling period according to a set sequence of the transmission priorities.

5. The method according to any one of claims 1 to 3, wherein the method further comprises:

in response to the plurality of types of data comprising data corresponding to a target transmission priority, directly sending the data corresponding to the target transmission priority to the second device through the transmission link, wherein the target transmission priority is higher than the plurality of transmission priorities.

6. The method according to claim 1, wherein the method further comprises:

setting start time of a next scheduling period according to end time of the current scheduling period.

7. The method according to claim 6, wherein before the obtaining the network quality indicator value of the transmission link between the first device and the second device in response to the current scheduling period being started, the method further comprises:

determining to start the current scheduling period, in response to scheduling period start time that is set according to end time of a previous scheduling period having been reached.

8. The method according to claim 1, wherein before the obtaining the network quality indicator

value of the transmission link between the first device and the second device in response to the current scheduling period being started, the method further comprises:

determining, in response to storing data in at least one of the plurality of queues, whether a previous scheduling period ends; and

determining to start the current scheduling period, in response to the previous scheduling period having ended.

9. The method according to claim 1, wherein the cloud server comprises any one of: a cloud desktop and a cloud application.

10. An electronic device, comprising: a memory, a processor, and a communication interface, wherein the memory stores executable codes, and the executable codes, when executed by the processor, cause the processor to perform the data transmission method according to any one of claims 1 to 9.

11. A non-transitory machine-readable storage medium, wherein the non-transitory machine-readable storage medium stores executable codes, and the executable codes, when executed by a processor of an electronic device, cause the processor to perform the data transmission method according to any one of claims 1 to 9.

12. A data transmission system, comprising:

a first device and a second device, wherein one of the first device and the second device corresponds to a cloud server, and the other of the first device and the second device corresponds to a client; and the first device is configured to:

obtain a plurality of types of data that the first device needs to send to the second device;

correspondingly buffer, according to a plurality of transmission priorities corresponding to the plurality of types of data, the plurality of types of data in a plurality of queues respectively corresponding to the plurality of transmission priorities;

obtain a network quality indicator value of a transmission link between the first device and the second device in response to a current scheduling period being started;

determine, according to the network quality indicator value, duration of the current scheduling period and a data amount that needs to be transmitted in the current scheduling period; and

obtain, according to the transmission priorities respectively corresponding to the plurality

of queues, data not exceeding the data amount from the plurality of queues, and send the obtained data to the second device through the transmission link within the duration.