WHAT IS CLAIMED IS:

1. A method for optimizing load balancing based on cloud monitoring, comprising:

obtaining: statistics on traffic distribution within a controlled area of a load balancing system; statistics on backend servers within a plurality of Availability Zones (AZ's) of the load balancing system; and monitored network quality of the plurality of AZs;

determining an AZ suitable to serve as a corresponding primary AZ of a load balancing instance based on the traffic distribution statistics, the backend server statistics, and the monitored network quality; and

performing a primary AZ switch of the load balancing instance in accordance with the determined result.

2. The method according to claim 1, wherein determining the AZ suitable to serve as the corresponding primary AZ of the load balancing instance comprises:

obtaining an exit bandwidth of each of the plurality of AZs of the load balancing system and a corresponding average bandwidth used at a public network exit of the AZ based on the traffic distribution statistics, and computing a first recommendation value of each AZ of the load balancing system for serving as the corresponding primary AZ of the load balancing instance based on the exit bandwidth and the average bandwidth used at the public network exit;

computing a second recommendation value of each AZ of the load balancing system for serving as the corresponding primary AZ of the load balancing instance, based on the backend server statistics;

computing a third recommendation value of each AZ of the load balancing system for serving as the corresponding primary AZ of the load balancing instance, based on the monitored network quality;

determining a resulting recommendation value of each AZ of the load balancing system for serving as the corresponding primary AZ of the load balancing instance, based on the first recommendation value, the second recommendation value, the third recommendation value, and respective weights of the three recommendation values; and

determining the AZ suitable to serve as the corresponding primary AZ of the load balancing instance based on the resulting recommendation value.

3. The method according to claim 2, wherein computing the first recommendation value of each AZ of the load balancing system for serving as the corresponding primary AZ of the load balancing system:

dividing the corresponding average bandwidth used at the public network exit of each AZ by the exit bandwidth of the AZ to obtain a traffic usage rate of the AZ;

multiplying a total bandwidth of the AZ by a preset coordination coefficient to obtain a corresponding weight of the AZ; and

determining the first recommendation value of each AZ of the load balancing system for serving as the corresponding primary AZ of the load balancing instance by the following principle: the lower the traffic usage rate the higher the recommendation value, and the heavier the weight the higher the recommendation value.

4. The method according to claim 2, wherein obtaining the backend server statistics within the plurality of AZs of the load balancing system comprises:

determining the number of backend servers operating in their normal state that are present in each AZ by a health check on the backend servers of the AZ; and determining the backend server statistics within the plurality of AZs of the load balancing system based on the numbers; and

wherein computing the second recommendation value of each AZ of the load balancing system for serving as the corresponding primary AZ of the load balancing instance based on the backend server statistics comprises:

determining the second recommendation value of each AZ of the load balancing system for serving as the corresponding primary AZ of the load balancing instance by the following principle: the greater the number, the higher the recommendation value.

5. The method according to claim 2, wherein computing the third recommendation value of each AZ of the load balancing system for serving as the corresponding primary AZ of the load balancing instance based on the monitored network quality comprises:

counting the number of times network quality parameters of each AZ go abnormal during an operation period based on operation data of network quality monitoring; and

determining the third recommendation value of each AZ of the load balancing system for serving as the corresponding primary AZ of the load balancing instance by the following principle: the higher the number of abnormalities, the lower the recommendation value.

6. The method according to claim 2, wherein determining the resulting recommendation value of each AZ of the load balancing system for serving as the corresponding primary AZ of the load balancing instance comprises:

summing a product obtained by multiplying the first recommendation value by a weight of a corresponding item, a product obtained by multiplying the second recommendation value by a weight of a corresponding item, and a product obtained by multiplying the third recommendation value by a weight of the corresponding item, to obtain a sum total; and

determining the sum total as the resulting recommendation value of each AZ of the load balancing system for serving as the corresponding primary AZ of the load balancing instance.

7. The method according to claim 2, wherein determining the AZ suitable to serve as the corresponding primary AZ of the load balancing instance based on the resulting recommendation value comprises:

determining the AZ with the highest resulting recommendation value, or an AZ with the resulting recommendation value greater than a preset threshold recommendation value as the AZ suitable to serve as the corresponding primary AZ of the load balancing instance.

8. The method according to claim 1, further comprising, prior to performing the primary/backup AZ switch of the load balancing instance in accordance with the determined result:

pre-distributing a network segment route from a load balancing device to which the load balancing instance belongs to a corresponding router of the load balancing device, by means of configuring a static equivalent route;

wherein performing the primary AZ switch of the load balancing instance in accordance with the determined result comprises:

switching the primary AZ of the load balancing instance to the AZ indicated in the determined result by adjusting preference values attached to a corresponding 32-bit route of the load balancing instance when routers are distributing the 32-bit route to backbone network routers, where the backbone network routers of different AZs are connected through Datacenter Interconnection (DCI) channels.

9. A device for optimizing load balancing based on cloud monitoring, comprising:

an acquisition unit, configured to obtain: statistics on traffic distribution within a controlled area of a load balancing system; statistics on backend servers within a plurality of Availability Zones (AZ's) of the load balancing system; and monitored network quality of the plurality of AZs;

a determination unit, configured to determine an AZ suitable to serve as a corresponding primary AZ of a load balancing instance based on the traffic distribution statistics, the backend server statistics, and the monitored network quality; and

a switching unit, configured to perform a primary AZ switch of the load balancing instance in accordance with the determined result.

10. The device according to claim 9, wherein the determination unit comprises:

an acquisition module, configured to: obtain an exit bandwidth of each of the plurality of AZs of the load balancing system and a corresponding average bandwidth used at a public network exit of the AZ based on the traffic distribution statistics; and compute a first recommendation value of each AZ of the load balancing system for serving as the corresponding primary AZ of the load balancing instance based on the exit bandwidth and the average bandwidth used at the public network exit;

a computation module, configured to compute a second recommendation value of each AZ of the load balancing system for serving as the corresponding primary AZ of the load balancing instance, based on the backend server statistics; and compute a third recommendation value of each AZ of the load balancing system for serving as the corresponding primary AZ of the load balancing instance, based on the monitored network quality;

wherein the determination module is configured to determine a resulting recommendation value of each AZ of the load balancing system for serving as the corresponding primary AZ of the load balancing instance, based on the first recommendation value, the second recommendation value, the third recommendation value, and respective weights of these three recommendation values; and

determine the AZ suitable to serve as the corresponding primary AZ of the load balancing instance based on the resulting recommendation value.

11. The device according to claim 10, wherein the computation module is configured to:

divide the corresponding average bandwidth used at the public network exit of each AZ by the exit bandwidth of the AZ to obtain a traffic usage rate of the AZ;

multiply a total bandwidth of the AZ by a preset coordination coefficient to obtain a corresponding weight of the AZ; and

determine the first recommendation value of each AZ in the load balancing system for serving as the corresponding primary AZ of the load balancing instance by the following principle: the lower the traffic usage rate the higher the recommendation value, and the heavier the weight the higher the recommendation value.

12. The device according to claim 10, wherein the acquisition unit is configured to: determine the number of backend servers operating in their normal state that are present in each AZ by a health check on the backend servers of the AZ; and determine the backend server statistics within the AZs of the load balancing system based on the numbers; and

wherein the computation module is configured to: determine the second recommendation value of each AZ of the load balancing system for serving as the corresponding primary AZ of the load balancing instance by the following principle: the greater the number, the higher the recommendation value.

13. The device according to claim 10, wherein the computation module is configured to:

count the number of times network quality parameters of each AZ go abnormal during an operation period based on the operation data of network quality monitoring; and

determine the third recommendation value of each AZ of the load balancing system for serving as the corresponding primary AZ of the load balancing instance by the following principle: the higher the number of abnormalities, the lower the recommendation value.

14. The device according to claim 10, wherein the determination module is configured to:

sum a product obtained by multiplying the first recommendation value by a weight of a corresponding item, a product obtained by multiplying the second recommendation value by a weight of a corresponding item, and a product obtained

by multiplying the third recommendation value by a weight of the corresponding item, to obtain a sum total; and

determine the sum total as the resulting recommendation value of each AZ of the load balancing system for serving as the corresponding primary AZ of the load balancing instance.

15. The device according to claim 10, wherein the determination module is configured to:

determine the AZ with the highest resulting recommendation value, or an AZ with the resulting recommendation value greater than a preset threshold recommendation value as the AZ suitable to serve as the corresponding primary AZ of the load balancing instance.

16. The device according to claim 9, further comprising a distribution unit configured to: pre-distribute a network segment route from the load balancing device to which the load balancing instance belongs to the corresponding router of the load balancing device by means of configuring a static equivalent route;

wherein the switching unit is configured to: switch the primary AZ of the load balancing instance to the AZ indicated in the determined result by adjusting preference values attached to a corresponding 32-bit route of the load balancing instance when routers are distributing the 32-bit route to backbone network routers, where the backbone network routers of different AZs are connected through Datacenter Interconnection (DCI) channels.

17. A storage device storing a computer program that when executed by a processor performs a method for optimizing load balancing based on cloud monitoring, the method comprising:

obtaining statistics on traffic distribution within a controlled area of a load balancing system; statistics on backend servers within a plurality of Availability Zones (AZ's) of the load balancing system; and monitored network quality of the plurality of AZs;

determining an AZ suitable to serve as a corresponding primary AZ of a load balancing instance based on the traffic distribution statistics, the backend server statistics, and the monitored network quality; and

performing a primary AZ switch of the load balancing instance in accordance with the determined result.

18. An apparatus for optimizing load balancing based on cloud monitoring, comprising a storage device, a processor, and a computer program stored on the storage device and executable by the processor, wherein the processor is configured to execute the computer program to perform a method for optimizing load balancing based on cloud monitoring, the method comprising:

obtaining: statistics on traffic distribution within a controlled area of a load balancing system; statistics on backend servers within a plurality of Availability Zones (AZ's) of the load balancing system; and monitored network quality of the plurality of AZs;

determining an AZ suitable to serve as a corresponding primary AZ of a load balancing instance based on the traffic distribution statistics, the backend server statistics, and the monitored network quality; and

performing a primary AZ switch of the load balancing instance in accordance with the determined result.

19. The apparatus according to claim 18, wherein the computer program when executed by the processor causes the following operations to be performed in determining the AZ suitable to serve as the corresponding primary AZ of the load balancing instance:

obtaining an exit bandwidth of each of the plurality of AZs of the load balancing system and a corresponding average bandwidth used at a public network exit of the AZ based on the traffic distribution statistics, and computing a first recommendation value of each AZ of the load balancing system for serving as the corresponding primary AZ of the load balancing instance based on the exit bandwidth and the average bandwidth used at the public network exit;

computing a second recommendation value of each AZ of the load balancing system for serving as the corresponding primary AZ of the load balancing instance, based on the backend server statistics;

computing a third recommendation value of each AZ of the load balancing system for serving as the corresponding primary AZ of the load balancing instance, based on the monitored network quality;

determining a resulting recommendation value of each AZ of the load balancing system for serving as the corresponding primary AZ of the load balancing instance, based on the first recommendation value, the second recommendation value, the third recommendation value, and respective weights of the three recommendation values; and

determine the AZ suitable to serve as the corresponding primary AZ of the load balancing instance based on the resulting recommendation value.

20. The apparatus according to claim 19, wherein the computer program when executed by the processor causes the following operations to be performed in computing the first recommendation value of each AZ of the load balancing system for serving as the corresponding primary AZ of the load balancing system:

dividing the corresponding average bandwidth used at the public network exit of each AZ by the exit bandwidth of the AZ to obtain a traffic usage rate of the AZ;

multiplying a total bandwidth of the AZ by a preset coordination coefficient to obtain a corresponding weight of the AZ; and

determining the first recommendation value of each AZ of the load balancing system for serving as the corresponding primary AZ of the load balancing instance by the following principle: the lower the traffic usage rate the higher the recommendation value, and the heavier the weight the higher the recommendation value.