DICOM Load Balancing Configuration

Enterprise Imaging Workflow Unifier

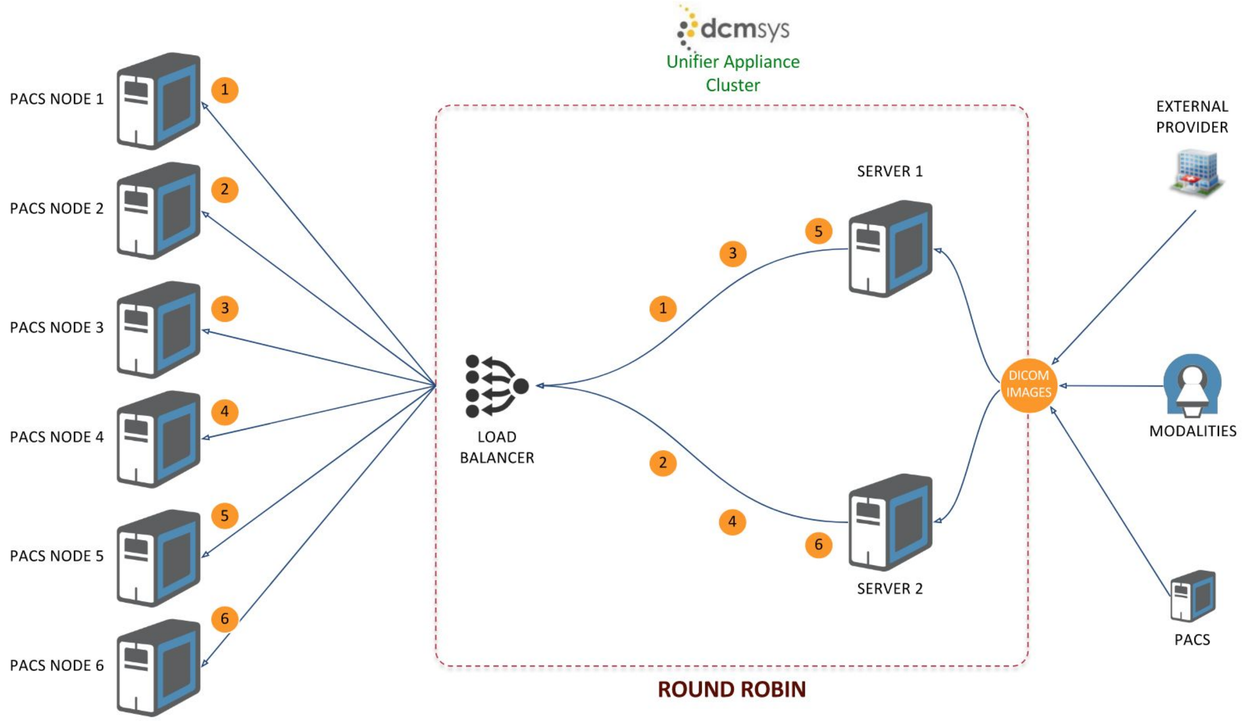
DICOM Load Balancing Configurations

Dicom Systems Unifiers support a variety of highly customizable Load Balancing algorithms. These algorithms are designed specifically for the Load Balancing of DICOM traffic. All images sharing a common Study UID are guaranteed to be routed to the same destination.

Round Robin

Round Robin is undoubtedly the most widely used algorithm. It's easy to implement and easy to understand. Here's how it works. Let's say you have 2 servers waiting for requests behind your load balancer. Once the first request arrives, the load balancer will forward that request to the 1st server. When the 2nd request arrives (presumably from a different client), that request will then be forwarded to the 2nd server.

Because the 2nd server is the last in this cluster, the next request (i.e., the 3rd) will be forwarded back to the 1st server, the 4th request back to the 2nd server, and so on, in a cyclical fashion.



As you can see, the method is very simple. However, it won't do well in certain scenarios.

For example, what if Server 1 had more CPU, RAM, and other specs compared to Server 2? Server 1 should be able to handle a higher workload than Server 2, right?

Unfortunately, a load balancer running on a round robin algorithm won't be able to treat the two servers accordingly. In spite of the two servers' disproportionate capacities, the load balancer will still distribute requests equally. As a result, Server 2 can get overloaded faster and probably even go down. You wouldn't want that to happen.
The Round Robin algorithm is best for clusters consisting of servers with identical specs. For other situations, you might want to look at other algorithms, like the ones below.

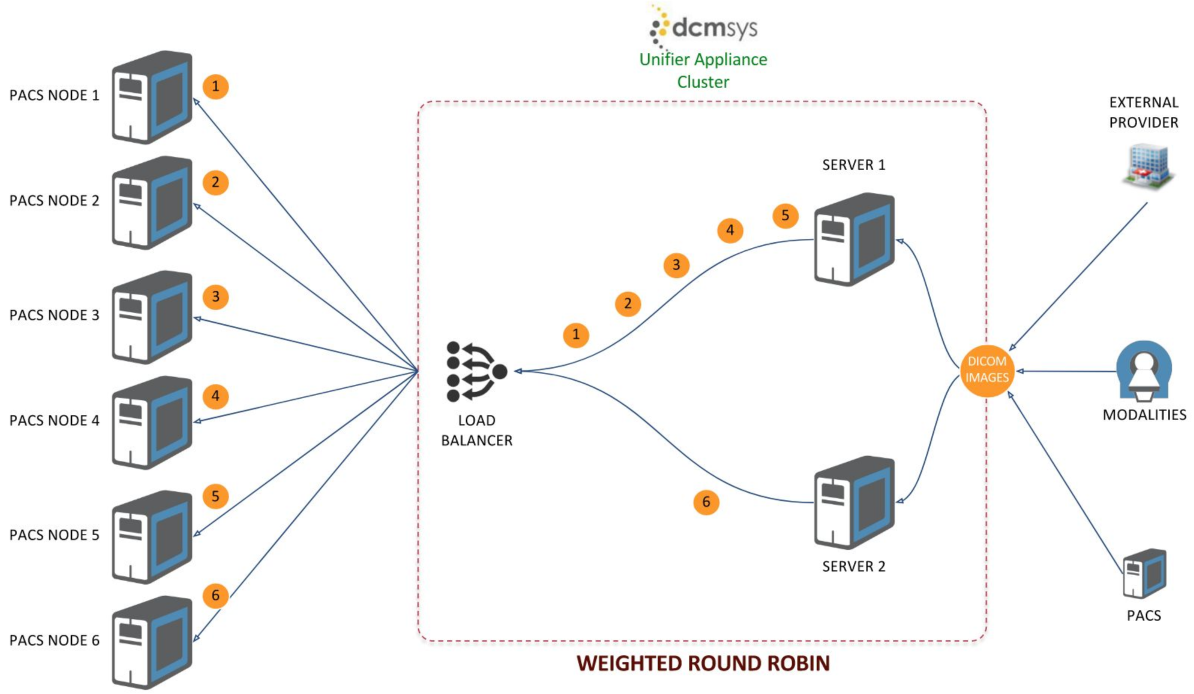
Weighted Round Robin

For the 2nd scenario mentioned above, i.e., Server 1 having higher specs than Server 2, you might prefer an algorithm that assigns more requests to the server with a higher capability of handling greater load. One such algorithm is the Weighted Round Robin.

The Weighted Round Robin is similar to the Round Robin in a sense that the manner by which requests are assigned to the nodes is still cyclical, albeit with a twist. The node with the higher specs will be apportioned a greater number of requests.

But how would the load balancer know which node has a higher capacity? Simple. You tell it beforehand. Basically, when you set up the load balancer, you assign "weights" to each node. The node with the higher specs should of course be given the higher weight.

You usually specify weights in proportion to actual capacities. So, for example, if Server 1's capacity is 5x more than Server 2's, then you can assign it a weight of 5 and Server 2 a weight of 1.

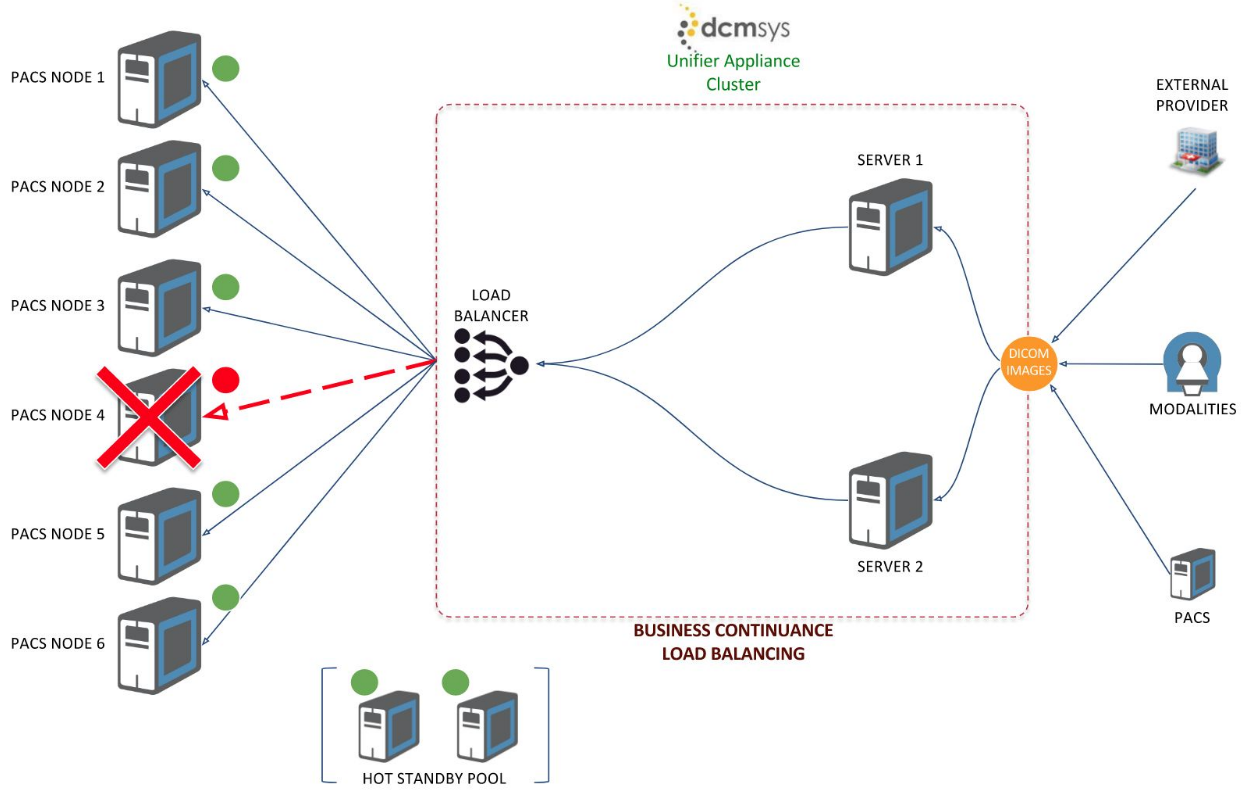


WEIGHTED ROUND ROBIN

So, when clients start coming in, the first 5 will be assigned to node 1 and the 6th to node 2. If more clients come in, the same sequence will be followed. That is, the 7th, 8th, 9th, 10th, and 11th will all go to Server1, and the 12th to Server 2, and so on.
Capacity isn't the only basis for choosing the Weighted Round Robin (WRR) algorithm. Sometimes, you'll want to use it if say you want one server to get a substantially lower number of connections than an equally capable server for the reason that the first server is running other types of mission-critical applications and you don't want it to be easily overloaded.

Load-Balancing based on any combination of DICOM tags, and number of threads per server, or pools of servers

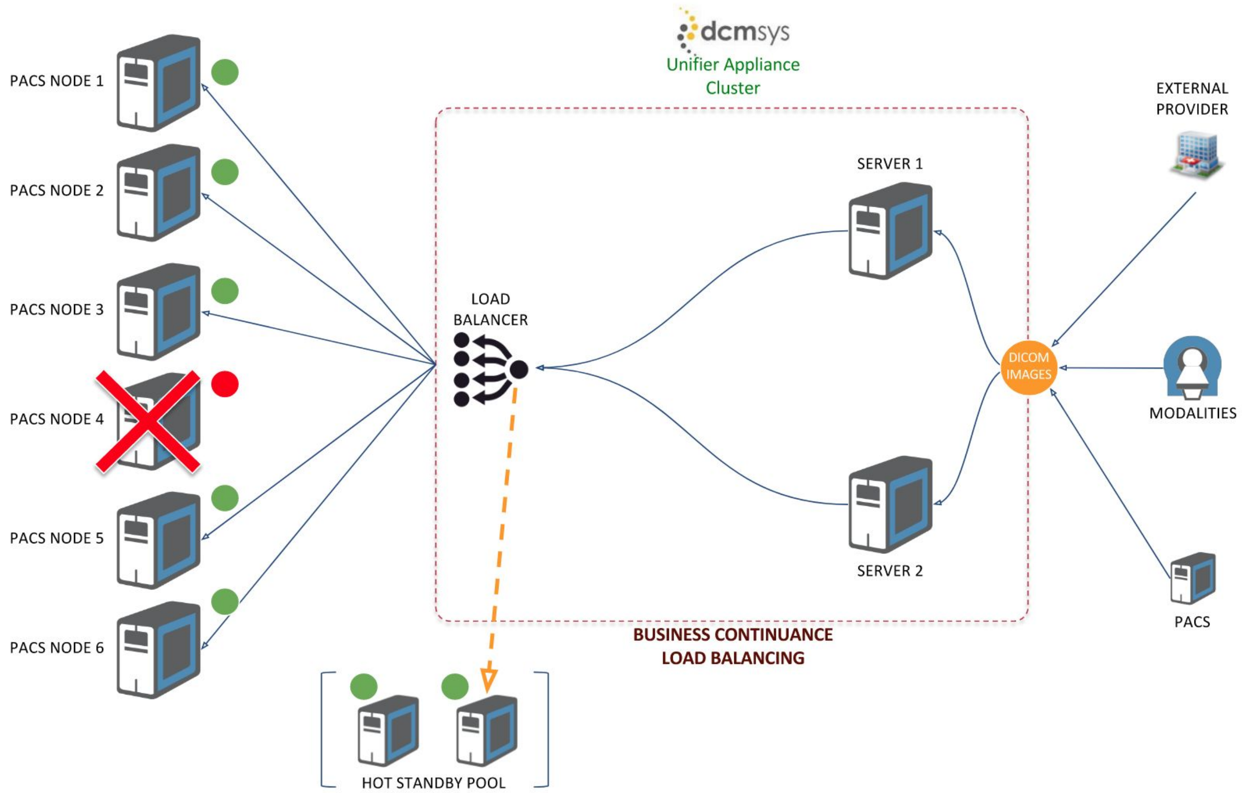
This load-balancing method is specific to the DICOM protocol and DICOM image transfer. A load balancing algorithm can be based on any DICOM tag and/or any combination of DICOM tags, with OR and AND Boolean expressions. For example, you can match any modality, station name and urgent patient to route with higher priority to a specific server, or group of servers, with higher availability. Additionally, the number of simultaneous threads can be defined as an option.



Business Continuance Load-Balancing based on target destination availability

Instead of randomly re-assigning the traffic, we define a HOT standby server, or pool of servers, to accommodate server failure conditions. In the event that one of the primary destination servers fails, and is unavailable, traffic will be redirected to the designated standby server associated with the active server.

A minimum of one standby server is required for this configuration. This configuration minimizes recovery time and data reconciliation in case of any failure. For deployment in Disaster Recovery Data centers, we support dual routing to both Datacenter pools with any of the load-balancing algorithms described above. The standby pool needs to be designed accordingly to make sure there are adequate resources in each data center.



With Unifier, an enterprise can intelligently and proactively balance the transmission of data to and from any system. Load balancing rules customized to specific preferences and needs can contribute to delivering timely, high quality care to patients while also optimizing IT resources, application delivery, and overall scalability.

Learn more by contacting sales@dcmsys.com or call (415) 684-8790