负载均衡速成

追逐时光者 2024年08月08日 07:33 广东

以下文章来源于ByteByteGo ，作者李华

ByteByteGo

硅谷百万粉丝技术大v官方号，官方网站 www.bytebytego.com

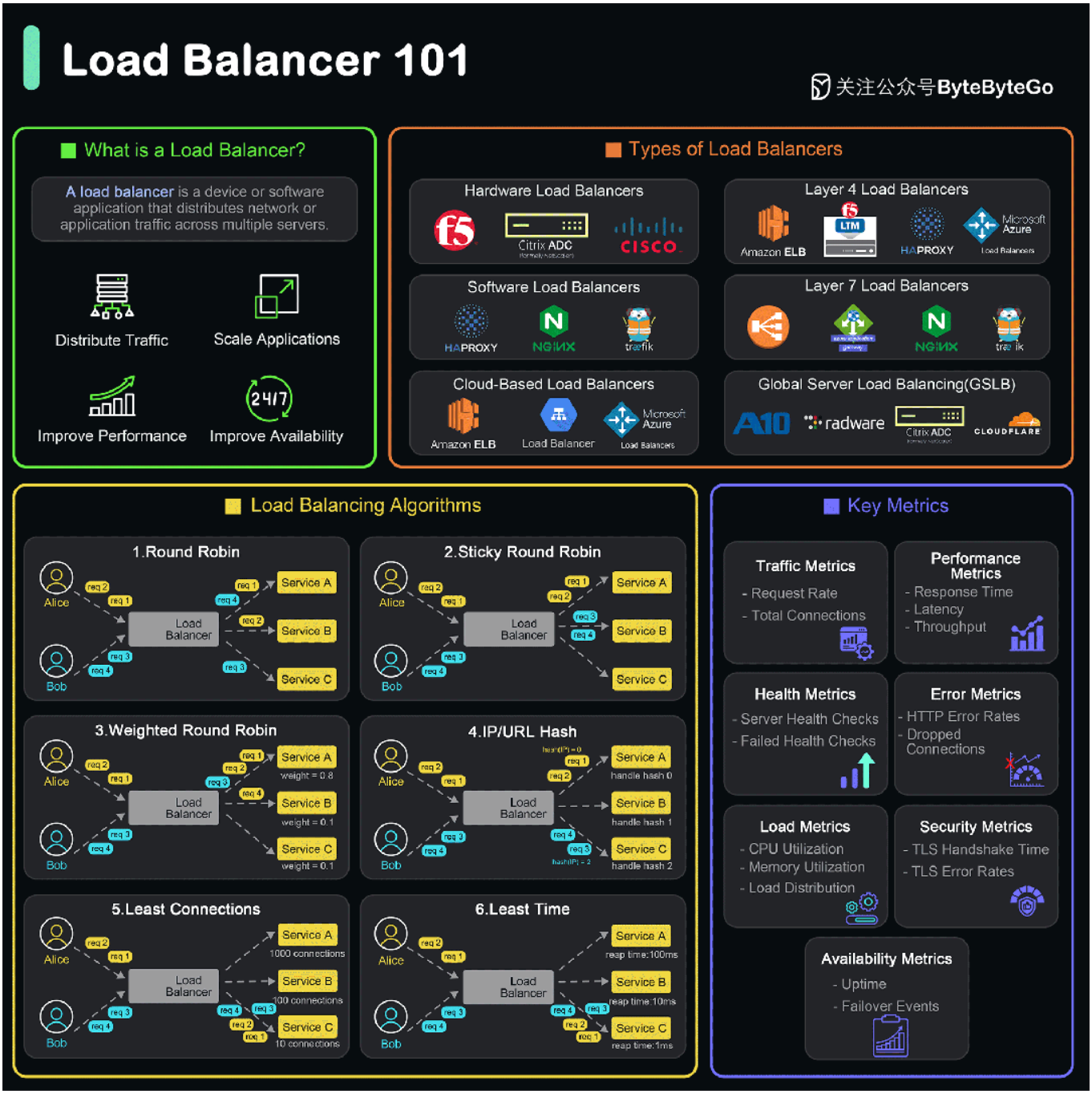
01 什么是负载均衡?

负载均衡是一种在多个服务器之间分配网络或应用程序流量的设备或软件应用程序。

负载均衡的主要目的是确保没有一台服务器承受过多的负载，从而提高响应速度并增加应用程序或网站的可用性。

02 负载均衡的作用是什么?

1. 分配流量：它将传入的网络或应用程序流量分配给多个服务器。
2. 确保可用性和可靠性：通过分配负载，即使一台或多台服务器出现故障，也能确保应用程序的可用性和可靠性。
3. 提高性能：通过平衡负载，有助于优化资源使用、最大限度地提高吞吐量、缩短响应时间，并避免任何一台服务器超负荷运行。
4. 扩展应用：它允许根据需求添加或移除服务器，而不会影响应用程序的性能。



03 负载均衡的类型

1. 硬件负载均衡：这是一种物理设备，用于在服务器之间分配流量。它们通常用于大型数据中心，可提供高性能和低延迟。
2. 软件负载均衡：它们是可安装在标准硬件或虚拟机上的应用程序。与硬件负载均衡相比，软件负载均衡更具灵活性和成本效益。
3. 基于云的负载均衡：由云服务提供商提供，这些负载均衡集成到云基础设施中。例如 AWS Elastic Load Balancer、Google Cloud Load Balancing 和 Azure Load Balancer。

4. 第 4 层负载均衡（传输层）：在传输层（OSI 第 4 层）运行，根据 IP 地址和 TCP/UDP 端口做出转发决定。它们不查看数据包的内容。
5. 第 7 层负载均衡（应用层）：在应用层（OSI 第 7 层）运行，根据信息内容（如 HTTP 标头、cookie、URL 路径）做出更复杂的决定。
6. 全球服务器负载均衡（GSLB）：将流量分配到多个地理位置，以提高全球范围内的冗余和性能。

04 负载均衡中使用的常见算法

1. 轮循：在服务器之间按顺序分配请求。
2. 最少连接：向活动连接最少的服务器发送流量。
3. 最少响应时间：将流量导向响应时间最快的服务器。
4. IP 哈希值：使用客户端的 IP 地址来确定接收请求的服务器。
5. 加权循环/最后连接：将更多流量分配给容量更大的服务器。

05 监控的关键指标

1. 流量指标

- 请求率：负载均衡每秒处理的请求数。有助于了解流量模式和高峰使用时间。
- 连接速率：每秒建立的新连接数。有助于检测流量的突然峰值。
- 总连接数：任何给定时间内的活动连接总数。表示总体负载，有助于容量规划。

2. 性能指标

- 响应时间：响应请求所需的平均时间。响应时间过长可能表明存在性能问题。
- 延迟：请求从客户端传输到服务器再返回所需的时间。高延迟会影响用户体验。
- 吞吐量：每秒通过负载均衡传输的数据量。对于了解数据负载和确保带宽容量非常重要。

3. 健康指标

- 服务器健康检查：对后端服务器进行健康检查的次数和状态。有助于识别故障或性能不佳的服务器。
- 失败的健康检查：健康检查失败的次数。表明后端服务器存在潜在问题。

4. 错误指标

- HTTP 错误率：服务器返回的 HTTP 错误（如 4xx、5xx）数量。错误率高可能表明应用程序或服务器存在问题。

- 中断连接：负载均衡中断的连接数。表明服务器可用性或容量存在潜在问题。

5. 负载指标

- CPU 使用率：负载均衡的 CPU 使用率。CPU 使用率高可能表明负载均衡负载过重或过小。
- 内存使用率：负载均衡的内存使用情况。内存使用率过高会影响性能和稳定性。
- 负载分布：后端服务器的流量分布。确保流量分布均匀，没有服务器超载。

6. 安全指标

- SSL/TLS 握手时间：建立安全连接所需的时间。对于了解加密带来的开销非常重要。
- SSL/TLS 出错率：遇到的 SSL/TLS 错误数量。表明证书或加密配置存在潜在问题。

7. 可用性指标

- 正常运行时间：负载均衡的总体可用性。对于了解负载均衡的可靠性至关重要。
- 故障转移事件：流量被重定向到另一个负载均衡的故障转移事件数量。

-------- 往期好文 --------

什么是 API 网关?

全栈开发要掌握什么技术?

面试官：网页太慢了怎么排查?

万字长文详解低时延股票交易系统的设计