(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 109819057 B
(45)授权公告日 2020.09.11

(21)申请号 201910275749.X

(22)申请日 2019.04.08

(65)同一申请的已公布的文献号
申请公布号 CN 109819057 A

(43)申请公布日 2019.05.28

(73)专利权人 科大讯飞股份有限公司
地址 230088 安徽省合肥市高新区望江西路666号

(72)发明人 刘坤 龙明康 王逸群

(74)专利代理机构 北京集佳知识产权代理有限公司 11227
代理人 王云晓 王宝筠

(51)Int.Cl.
H04L 29/08(2006.01)
H04L 12/803(2013.01)
H04L 12/851(2013.01)

(56)对比文件
CN 103246546 A,2013.08.14
CN 103118142 A,2013.05.22
CN 101115016 A,2008.01.30
CN 105227602 A,2016.01.06
CN 109462647 A,2019.03.12
EP 3113539 A1,2017.01.04
WO 03012667 A1,2003.02.13

审查员 王勇

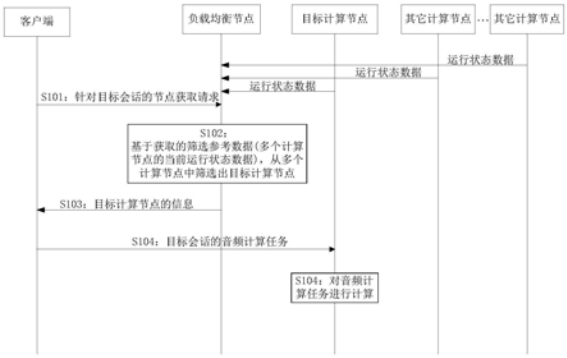
权利要求书3页 说明书12页 附图3页

(54)发明名称

一种负载均衡方法及系统

(57)摘要

本申请提供了一种负载均衡方法及系统，负载均衡方法应用于负载均衡系统，负载均衡系统包括负载均衡节点和多个计算节点，方法包括：负载均衡节点接收客户端针对一条目标会话发送的节点获取请求；负载均衡节点基于获取的筛选参考数据，从多个计算节点中筛选出目标计算节点，并将目标计算节点的信息发送至客户端，其中，筛选参考数据至少包括多个计算节点的当前运行状态数据，目标计算节点接收客户端发送的目标会话的音频计算任务，并对目标会话的音频计算任务进行计算。本申请提供的负载均衡方法提高了系统整体资源的使用率，降低了系统性能波动，且解决了负载均衡节点的流量瓶颈问题。



CN 109819057 B

1.一种负载均衡方法,其特征在于,应用于负载均衡系统中的负载均衡节点,所述方法包括:

接收客户端针对一条目标会话发送的节点获取请求,所述节点获取请求用于请求获取所述负载均衡系统中处理所述目标会话的目标计算节点;

基于获取的筛选参考数据,从所述负载均衡系统中的多个计算节点中筛选出所述目标计算节点,其中,所述筛选参考数据至少包括所述多个计算节点的当前运行状态数据,一计算节点的当前运行状态数据至少包括当前最佳任务并发量,所述当前最佳任务并发量由该计算节点处理完上一音频计算任务后,根据以其处理的上一音频计算任务是否为首个音频计算任务为依据确定的最佳任务并发量,以及处理上一音频计算任务时的实时率和响应时间确定;

将所述目标计算节点的信息发送至所述客户端,以使所述客户端基于所述目标计算节点的信息将所述目标会话的音频计算任务直接发送至所述目标计算节点进行计算。

2.根据权利要求1所述的负载均衡方法,其特征在于,所述筛选参考数据还包括:所述目标会话的会话时长预测值和发包频率预测值;

其中,所述目标会话的会话时长预测值和发包频率预测值采用与所述目标会话对应的场景标识和用户标识所对应的会话参数预测模型预测得到。

3.根据权利要求1所述的负载均衡方法,其特征在于,还包括:

按预设的采集周期采集各计算节点的运行状态数据并记录;或者,接收各计算节点按预设的上报周期上报的自身的运行状态数据并记录;

任一计算节点的当前运行状态数据为所述负载均衡节点记录的该计算节点的最新运行状态数据。

4.根据权利要求1～3中任意一项所述的负载均衡方法,其特征在于,任一计算节点的当前运行状态数据还包括以下数据中的一种或多种:

当前CPU使用率、当前GPU使用率、当前的任务并发量。

5.根据权利要求4所述的负载均衡方法,其特征在于,所述筛选参考数据还包括:

各个计算节点当前处理会话的发包频率和/或各个计算节点当前处理会话的发起时刻。

6.一种负载均衡方法,其特征在于,应用于负载均衡系统中的多个计算节点中的目标计算节点,所述目标计算节点为所述负载均衡系统中的负载均衡节点在接收到客户端针对一条目标会话发起的节点获取请求时,基于筛选参考数据从所述负载均衡系统中的多个计算节点中筛选出的计算节点,所述筛选参考数据至少包括所述多个计算节点的当前运行状态数据,一计算节点的当前运行状态数据至少包括当前最佳任务并发量;

所述方法包括:

接收所述客户端发送的、所述目标会话的音频计算任务;

对所述音频计算任务进行计算;

处理完所述音频计算任务时,以所述目标会话的音频计算任务是否为首个任务,以及处理所述目标会话的音频计算任务时的实时率和响应时间为依据,确定当前最佳任务并发量,以提供给所述负载均衡节点。

7.根据权利要求6所述的负载均衡方法,其特征在于,还包括:

在接收到所述目标会话的音频计算任务时，和/或，处理完所述目标会话的音频计算任务时，获取自身的当前运行状态数据；

将获取的当前运行状态数据上报至所述负载均衡节点，以使所述负载均衡节点对其针对所述目标计算节点记录的运行状态数据进行更新。

8.根据权利要求7所述的负载均衡方法，其特征在于，所述当前运行状态数据还包括以下数据中的一种或多种：

当前CPU使用率、当前GPU使用率、当前的任务并发量。

9.根据权利要求7或8所述的负载均衡方法，其特征在于，还包括：

在接收到所述目标会话的音频计算任务时，和/或，处理完所述目标会话的音频计算任务时，获取自身当前处理会话的发包频率和/或当前处理会话的发起时刻；

将自身当前处理会话的发包频率和/或当前处理会话的发起时刻上报至所述负载均衡节点。

10.根据权利要求6所述的负载均衡方法，其特征在于，所述以所述目标会话的音频计算任务是否为首个任务，以及处理所述目标会话的音频计算任务时的实时率和响应时间为依据，确定当前最佳任务并发量，包括：

若所述目标会话的音频计算任务为首个任务，则获取预先确定的初始最佳任务并发量作为当前的最佳任务并发量；若所述目标会话的音频计算任务并非首个任务，则获取其处理完上一音频计算任务后确定的最佳任务并发量，作为当前的最佳任务并发量；

基于所述当前的最佳任务并发量、当前的任务并发量和处理所述目标会话的音频计算任务时的实时率和响应时间，确定是否需要对所述当前的最佳任务并发量进行调整；

若需要对所述当前的最佳任务并发量进行调整，则基于预设的调整步长逐步对所述当前的最佳任务并发量进行调整；调整后的最佳任务并发量作为当前最终的最佳任务并发量。

11.根据权利要求10所述的负载均衡方法，其特征在于，所述基于所述当前的最佳任务并发量、当前的任务并发量和处理所述目标会话的音频计算任务时的实时率和响应时间，确定是否需要对所述当前的最佳任务并发量进行调整，包括：

当所述当前的任务并发量与所述当前的最佳任务并发量的差值的绝对值小于预设的第一阈值时，若处理所述目标会话的音频计算任务时的实时率与基准实时率的差值的绝对值大于预设的第二阈值，和/或，处理所述目标会话的音频计算任务时的响应时间与基准响应时间的差值的绝对值大于预设的第三阈值，则确定需要对所述当前的最佳任务并发量进行调整。

12.根据权利要求10所述的负载均衡方法，其特征在于，确定所述初始最佳任务并发量的过程包括：

获取自身的CPU型号、CPU实际核数、GPU型号、音频处理模型大小，以及自身未执行音频计算任务时的内存可用空间大小和显存可用空间大小；

基于自身的CPU型号、CPU实际核数、GPU型号、音频处理模型大小，以及自身未执行音频计算任务时的内存可用空间大小和显存可用空间大小，获取基于CPU算力和GPU算力评估的第一最大任务并发量、基于内存空间评估的第二最大任务并发量、以及基于显存空间评估的第三最大任务并发量；

将所述第一最大任务并发量、所述第二最大任务并发量和所述第三最大任务并发量中的最小并发量确定为所述初始最佳任务并发量。

13.一种负载均衡系统，其特征在于，包括：负载均衡节点和多个计算节点；

所述负载均衡节点，用于接收客户端针对一条目标会话发送的节点获取请求，所述节点获取请求用于请求获取处理所述目标会话的目标计算节点；基于获取的筛选参考数据，从所述多个计算节点中筛选出目标计算节点，并将所述目标计算节点的信息发送至所述客户端，其中，所述筛选参考数据至少包括所述多个计算节点的当前运行状态数据，一计算节点的当前运行状态数据至少包括当前最佳任务并发量，所述当前最佳任务并发量由该计算节点处理完上一音频计算任务后，根据以其处理的上一音频计算任务是否为首个音频计算任务为依据确定的最佳任务并发量，以及处理上一音频计算任务时的实时率和响应时间确定；

所述多个计算节点中的所述目标计算节点，用于接收所述客户端发送的所述目标会话的音频计算任务，并对所述音频计算任务进行计算，以及，处理完所述音频计算任务时，以所述目标会话的音频计算任务是否为首个任务，以及处理所述目标会话的音频计算任务时的实时率和响应时间为依据，确定当前最佳任务并发量，以提供给所述负载均衡节点。

14.根据权利要求13所述的负载均衡系统，其特征在于，所述筛选参考数据还包括：所述目标会话的会话时长预测值和发包频率预测值；

其中，所述目标会话的会话时长预测值和发包频率预测值通过与所述目标会话对应的场景标识和用户标识所对应的会话参数预测模型预测得到。

15.根据权利要求13或14所述的负载均衡系统，其特征在于，所述目标计算节点，还用于在接收到所述目标会话的音频计算任务时，和/或，处理完所述目标会话的音频计算任务时，获取自身的当前运行状态数据；将自身的当前运行状态数据上报至所述负载均衡节点，以使所述负载均衡节点对其针对所述目标计算节点记录的运行状态数据进行更新。

16.根据权利要求13所述的负载均衡系统，其特征在于，所述目标计算节点在处理完所述音频计算任务时，以所述目标会话的音频计算任务是否为首个任务，以及处理所述目标会话的音频计算任务时的实时率和响应时间为依据，确定当前最佳任务并发量时，具体用于当所述目标会话的音频计算任务为首个任务时，获取预先确定的初始最佳任务并发量作为当前的最佳任务并发量，当所述目标会话的音频计算任务并非首个任务时，获取其处理完上一音频计算任务后确定的最佳任务并发量，作为当前的最佳任务并发量；基于所述当前的最佳任务并发量、当前的任务并发量和处理所述目标会话的音频计算任务时的实时率和响应时间，确定是否需要对所述当前的最佳任务并发量进行调整；若需要对所述当前的最佳任务并发量进行调整，则基于预设的调整步长逐步对所述当前的最佳任务并发量进行调整；调整后的最佳任务并发量作为当前最终的最佳任务并发量。

一种负载均衡方法及系统

技术领域

[0001] 本申请涉及分布式计算技术领域，更具体地说，涉及一种负载均衡方法及系统。

背景技术

[0002] 应用系统数据量的增长，使得数据处理请求和计算强度相应地增长，而日益增长的数据处理请求和计算强度对系统的规模和处理能力提出了更高的要求。为了应对这样的趋势，出现了分布式系统。分布式系统即利用大量计算节点完成单个节点无法完成的计算、存储任务，分布式系统中大量计算节点的存在使得负载均衡变得尤为重要。

[0003] 负载均衡指的是，如果一组计算节点提供同质化的服务，那么对服务的请求就应该均匀分摊到这些节点上。负载均衡的意义在于，让所有的节点以最小的代价、最好的状态对外提供服务。负载均衡提高了系统的可靠性，降低了单个节点过载、宕机之后对整个系统的影响。

[0004] 实时语音计算系统是目前比较常用的一种分布式的应用系统，由于实时语音计算系统处理的是音频流，因此，其具有流式密集计算的特点。目前的负载均衡方法并没有考虑实时语音计算系统的特点，即目前的负载均衡方法在实时语音计算系统上的应用效果并不理想。

发明内容

[0005] 有鉴于此，本申请提供了一种负载均衡方法及系统，用以提供一种面向实时语音计算的负载均衡方案，其技术方案如下：

[0006] 一种负载均衡方法，应用于负载均衡系统中的负载均衡节点，所述方法包括：

[0007] 接收客户端针对一条目标会话发送的节点获取请求，所述节点获取请求用于请求获取所述负载均衡系统中处理所述目标会话的目标计算节点；

[0008] 基于获取的筛选参考数据，从所述负载均衡系统中的多个计算节点中筛选出所述目标计算节点，其中，所述筛选参考数据至少包括所述多个计算节点的当前运行状态数据；

[0009] 将所述目标计算节点的信息发送至所述客户端，以使所述客户端基于所述目标计算节点的信息将所述目标会话的音频计算任务直接发送至所述目标计算节点进行计算。

[0010] 可选的，所述筛选参考数据还包括：所述目标会话的会话时长预测值和发包频率预测值；

[0011] 其中，所述目标会话的会话时长预测值和发包频率预测值采用与所述目标会话对应的场景标识和用户标识所对应的会话参数预测模型预测得到。

[0012] 可选的，所述负载均衡方法还包括：

[0013] 按预设的采集周期采集各计算节点的运行状态数据并记录；或者，接收各计算节点按预设的上报周期上报的自身的运行状态数据并记录；

[0014] 任一计算节点的当前运行状态数据为所述负载均衡节点记录的该计算节点的最新运行状态数据。

[0015] 可选的，任一计算节点的当前运行状态数据包括以下数据中的一种或多种：当前CPU使用率、当前GPU使用率、当前的任务并发量、当前的最佳任务并发量。

[0016] 可选的，所述筛选参考数据还包括：各个计算节点当前处理会话的发包频率和/或各个计算节点当前处理会话的发起时刻。

[0017] 一种负载均衡方法，应用于负载均衡系统中的多个计算节点中的目标计算节点，所述目标计算节点为所述负载均衡系统中的负载均衡节点在接收到客户端针对一条目标会话发起的节点获取请求时，基于筛选参考数据从所述负载均衡系统中的多个计算节点中筛选出的计算节点，所述筛选参考数据至少包括所述多个计算节点的当前运行状态数据；

[0018] 所述方法包括：

[0019] 接收所述客户端发送的、所述目标会话的音频计算任务；

[0020] 对所述音频计算任务进行计算。

[0021] 可选的，所述负载均衡方法还包括：

[0022] 在接收到所述目标会话的音频计算任务时，和/或，处理完所述目标会话的音频计算任务时，获取自身的当前运行状态数据；

[0023] 将获取的当前运行状态数据上报至所述负载均衡节点，以使所述负载均衡节点对其针对所述目标计算节点记录的运行状态数据进行更新。

[0024] 可选的，所述当前运行状态数据包括以下数据中的一种或多种：

[0025] 当前CPU使用率、当前GPU使用率、当前的任务并发量、当前的最佳任务并发量。

[0026] 可选的，所述的负载均衡方法还包括：

[0027] 在接收到所述目标会话的音频计算任务时，和/或，处理完所述目标会话的音频计算任务时，获取自身当前处理会话的发包频率和/或当前处理会话的发起时刻；

[0028] 将自身当前处理会话的发包频率和/或当前处理会话的发起时刻上报至所述负载均衡节点。

[0029] 可选的，处理完所述音频计算任务时，获取当前最佳任务并发量，包括：

[0030] 若所述目标会话的音频计算任务为首个任务，则获取预先确定的初始最佳任务并发量作为当前的最佳任务并发量；若所述目标会话的音频计算任务并非首个任务，则获取其处理完上一音频计算任务后确定的最佳任务并发量，作为当前的最佳任务并发量；

[0031] 基于所述当前的最佳任务并发量、当前的任务并发量和处理所述目标会话的音频计算任务时的实时率和响应时间，确定是否需要对所述当前的最佳任务并发量进行调整；

[0032] 若需要对所述当前的最佳任务并发量进行调整，则基于预设的调整步长逐步对所述当前的最佳任务并发量进行调整；调整后的最佳任务并发量作为当前最终的最佳任务并发量。

[0033] 可选的，所述基于所述当前的最佳任务并发量、当前的任务并发量和处理所述目标会话的音频计算任务时的实时率和响应时间，确定是否需要对所述当前的最佳任务并发量进行调整，包括：

[0034] 当所述当前的任务并发量与所述当前的最佳任务并发量的差值的绝对值小于预设的第一阈值时，若处理所述目标会话的音频计算任务时的实时率与基准实时率的差值的绝对值大于预设的第二阈值，和/或，处理所述目标会话的音频计算任务时的响应时间与基准响应时间的差值的绝对值大于预设的第三阈值，则确定需要对所述当前的最佳任务并发

量进行调整。

[0035] 可选的，确定所述初始最佳任务并发量的过程包括：

[0036] 获取自身的CPU型号、CPU实际核数、GPU型号、音频处理模型大小，以及自身未执行音频计算任务时的内存可用空间大小和显存可用空间大小；

[0037] 基于自身的CPU型号、CPU实际核数、GPU型号、音频处理模型大小，以及自身未执行音频计算任务时的内存可用空间大小和显存可用空间大小，获取基于CPU算力和GPU算力评估的第一最大任务并发量、基于内存空间评估的第二最大任务并发量、以及基于显存空间评估的第三最大任务并发量；

[0038] 将所述第一最大任务并发量、第二最大任务并发量和第三最大任务并发量中的最小并发量确定为所述初始最佳任务并发量。

[0039] 一种负载均衡系统，包括：负载均衡节点和多个计算节点；

[0040] 所述负载均衡节点，用于接收客户端针对一条目标会话发送的节点获取请求，所述节点获取请求用于请求获取处理所述目标会话的目标计算节点；基于获取的筛选参考数据，从所述多个计算节点中筛选出目标计算节点，并将所述目标计算节点的信息发送至所述客户端，其中，筛选参考数据至少包括所述多个计算节点的当前运行状态数据；

[0041] 所述多个计算节点中的所述目标计算节点，用于接收所述客户端发送的所述目标会话的音频计算任务，并对所述音频计算任务进行计算。

[0042] 可选的，所述筛选参考数据还包括：所述目标会话的会话时长预测值和发包频率预测值；

[0043] 其中，所述目标会话的会话时长预测值和发包频率预测值采用与所述目标会话对应的场景标识和用户标识所对应的会话参数预测模型预测得到。

[0044] 可选的，所述目标计算节点，还用于在接收到所述目标会话的音频计算任务时，和/或，处理完所述目标会话的音频计算任务时，获取自身的当前运行状态数据；将自身的当前运行状态数据上报至所述负载均衡节点，以使所述负载均衡节点对其针对所述目标计算节点记录的运行状态数据进行更新。

[0045] 可选的，所述目标计算节点在处理完所述音频计算任务时，获取自身的当前最佳任务并发量时，具体用于当所述目标会话的音频计算任务为首个任务时，获取预先确定的初始最佳任务并发量作为当前的最佳任务并发量，当所述目标会话的音频计算任务并非首个任务时，获取其处理完上一音频计算任务后确定的最佳任务并发量，作为当前的最佳任务并发量；基于所述当前的最佳任务并发量、当前的任务并发量和处理所述目标会话的音频计算任务时的实时率和响应时间，确定是否需要对所述当前的最佳任务并发量进行调整；若需要对所述当前的最佳任务并发量进行调整，则基于预设的调整步长逐步对所述当前的最佳任务并发量进行调整；调整后的最佳任务并发量作为当前最终的最佳任务并发量。

[0046] 经由上述方案可知，本申请提供的负载均衡方法及系统，考虑到计算节点的压力实时变化，负载均衡节点在接收到客户端针对一条目标会话发起的节点获取请求时，至少基于多个计算节点的当前运行状态从多个计算节点中挑选目标计算节点，为了避免负载均衡节点出现明显流量瓶颈的问题，本申请提供基于旁路模式的均衡策略，即，客户端在获得目标计算节点的信息后，将目标会话的音频计算任务直接发送至目标计算节点，而不再经

负载均衡节点，如此，负载均衡节点的流量大大减少，本申请提供的负载均衡方法提高了系统整体资源的使用率，降低了系统性能波动，且解决了负载均衡节点的流量瓶颈问题。

附图说明

[0047] 为了更清楚地说明本申请实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本申请的实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据提供的附图获得其他的附图。

[0048] 图1为本申请实施例提供的负载均衡方法的流程示意图；

[0049] 图2为本申请实施例提供的目标计算节点处理完目标会话的音频计算任务时，获取自身当前的最佳任务并发量的实现过程的流程示意图；

[0050] 图3为本申请实施例提供的目标计算节点确定初始最佳任务并发量的流程示意图；

[0051] 图4为本申请实施例提供的负载均衡系统的结构示意图。

具体实施方式

[0052] 下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

[0053] 目前的负载均衡方法为：首先，通过配置写入需要负载均衡的计算节点的地址，然后，使用哈希、加权、轮询等方式从各计算节点的地址中筛选出目标地址(即从各个计算节点中筛选出目标节点)，最后，负载均衡节点基于筛选出的目标地址进行消息转发。

[0054] 发明人经研究发现，对于实时语音计算这一应用，在大规模集群下，计算节点的算力以及会话的时长差异(比如，不同用户的说话方式或说话习惯不同，导致有的会话长，有的会话短)明显，且计算节点的压力(数据流量)实时变化，而现有的负载均衡方法在应用于实时语音计算时，并未考虑上述情况，导致现有的负载均衡方法存在如下问题：

[0055] 其一，现有的负载均衡方法采用的是静态转发策略，即，在向计算节点分发请求的时候，并未考虑计算节点的实时状态；

[0056] 其二，现有的负载均衡方法采用的是集中式转发策略，即所有的数据流量都要经过负载均衡节点，音频数据流量占用较高，导致负载均衡节点存在明显的流量瓶颈；

[0057] 其三，在挑选计算节点时，未考虑音频会话时长等影响计算时长的因素；

[0058] 其四，计算节点无法自动评估以及自动调整自身的运算能力，难以适应大规模集群中计算节点的算力差异明显的场景。

[0059] 由于现有的负载均衡方案并没有考虑上述情况，因此，难以保证计算节点压力的平衡，往往会导致系统性能波动，资源利用率较低，且，负载均衡节点存在明显的流量瓶颈。

[0060] 鉴于上述问题，本案发明人进进行了深入研究，最终提出了一种效果较好的、面向实时语音计算的负载均衡方案。接下来通过下述实施例对本申请提供的负载均衡方法。

[0061] 请参阅图1，示出了本申请实施例提供的负载均衡方法的流程示意图，该负载均衡

方法可以包括：

[0062] 步骤S101：负载均衡系统中的负载均衡节点接收客户端针对一条目标会话发送的节点获取请求。

[0063] 其中，一条会话指的是一句话或者连续的多句话，针对一条会话发送的节点获取请求用于请求获取对该会话的音频数据进行计算的最佳计算节点。

[0064] 步骤S102：负载均衡节点基于获取的筛选参考数据，从负载均衡系统中的多个计算节点中筛选出目标计算节点。

[0065] 其中，筛选参考数据至少包括多个计算节点的当前运行状态数据。考虑到计算节点的运行状态是实时变化的，本实施例至少以各个计算节点的当前运行状态数据为筛选依据，从多个计算节点中筛选目标计算节点。

[0066] 任一计算节点的当前运行状态数据可以包括以下数据中的一种或多种：当前CPU使用率、当前GPU使用率、当前的任务并发量和当前的最佳任务并发量，优选的，任一计算节点的当前运行状态数据包括上述的所有数据。

[0067] 优选的，筛选参考数据除了包括上述各个计算节点的当前运行状态数据外，还可以包括各个计算节点当前处理会话的发包频率和/或各个计算节点当前处理会话的发起时刻。

[0068] 在本实施例，负载均衡节点若要基于多个计算节点的当前运行状态数据筛选目标计算节点，首先需要获得多个计算节点的当前运行状态数据，在一种可能的实现方式中，负载均衡节点按预设的采集周期采集各计算节点的运行状态数据、各个计算节点当前处理会话的发包频率和当前处理会话的发起时刻，并记录这些数据，在另一种可能的实现方式中，各计算节点按预设的上报周期主动向负载均衡节点上报自身的运行状态数据、自身当前处理会话的发包频率和当前处理会话的发起时刻，负载均衡节点接收各计算节点上报的数据并记录，筛选参考数据中任一计算节点的当前运行状态数据、当前处理会话的发包频率和当前处理会话的发起时刻为负载均衡节点记录的最新数据。

[0069] 步骤S103：负载均衡节点将目标计算节点的信息发送至客户端。

[0070] 客户端接收到目标计算节点的信息后，将会话的音频计算任务发送至目标计算节点。

[0071] 其中，目标计算节点的信息可以但不限定为该目标计算节点的地址，还可以为目标计算节点的节点标识，此时，客户端可存储节点标识与节点地址的对应关系，在获得目标计算节点的节点标识后，可基于节点标识与节点地址的对应关系获得目标计算节点的地址。客户端在获得目标计算节点的地址后，便可将目标会话的音频计算任务发送至目标计算节点。

[0072] 本实施例中，负载均衡节点确定出目标计算节点，并将目标计算节点的信息反馈至客户端后，客户端直接向目标计算节点发送目标会话的音频计算任务，而不需要再经负载均衡节点，负载均衡节点的数据流量大大减少，从而解决了负载均衡节点流量瓶颈的问题。

[0073] 需要说明的是，客户端在接收到目标会话的首个音频帧时，向负载均衡节点发送节点获取请求，在一种可能的实现方式中，客户端在向负载均衡节点发送节点获取请求时，可在节点获取请求中携带目标会话的首个音频帧，当负载均衡节点接收到节点获取请求

时，确定出目标计算节点，确定出目标计算节点后，一方面将目标计算节点的信息(比如地址)发送至客户端，另一方面将节点获取请求中携带的、目标会话的首个音频帧发送至目标计算节点进行计算，当客户端接收到目标计算节点的信息时，将目标会话的后续音频帧发送至目标计算节点进行计算。在另一种可能的实现方式中，客户端在向负载均衡节点发送节点获取请求时，可不在节点获取请求中携带目标会话的首个音频帧，当客户端获得负载均衡节点确定出的目标计算节点后，再将目标会话的首个音频帧及后续的音频帧陆续发送至目标计算节点进行计算。

[0074] 步骤S104：目标计算节点接收客户端发送的目标会话的音频计算任务进行计算。

[0075] 本申请实施例提供的负载均衡方法，考虑到计算节点的压力实时变化，负载均衡节点在接收到客户端针对目标会话发起的节点获取请求时，至少基于多个计算节点的当前运行状态从多个计算节点中挑选目标计算节点，为了避免负载均衡节点出现明显流量瓶颈的问题，本申请提供基于旁路模式的均衡策略，即，客户端在获得目标计算节点的信息后，将目标会话的音频计算任务直接发送至目标计算节点，而不再经负载均衡节点，如此，负载均衡节点的流量大大减少，本申请提供的负载均衡方法提高了系统整体资源的使用率，降低了系统性能波动，且解决了负载均衡节点的流量瓶颈问题。

[0076] 可以理解的是，在大规模请求接入的情况下，系统中会话的会话时长差异通常较大，会话时长的差异会导致不同时长的会话消耗的计算资源存在差异，比如，会话时长较短的会话消耗的计算资源较少，而会话时长较长的会话消耗的计算资源较多，也就是说，会话时长对系统中各个计算节点的均衡程度会产生影响，如果能够获得会话的会话时长，在负载均衡时将会话时长这一影响因素加以考虑，将提高负载均衡的效果。然而，实时的语音交互场景使得用户在发起语音交互请求时，系统在流式数据传输下无法准确获得当次会话的会话时长。

[0077] 有鉴于此，本申请针对不同的应用场景和不同的用户预先训练出会话参数预测模型，具体的，会话参数预测模型的每一训练样本为一条会话对应的网络类型和发起时刻，样本标签为该会话的实际会话时长、该会话的实际发包频率。其中，一条会话对应的网络类型为传输该条会话的音频数据的网络的类型，比如3G、4G等。

[0078] 在本实施例中，当客户端接收到目标会话的首个音频帧时，会基于目标会话对应的应用场景标识和用户标识获取到对应的会话参数预测模型，然后将目标会话对应的网络类型和发起时刻输入该会话参数预测模型，从而获得该会话参数预测模型输出的，目标会话的会话时长预测值和发包频率预测值。

[0079] 客户端在获得目标会话的会话时长预测值和发包频率预测值后，向负载均衡节点发送节点获取请求，并在节点获取请求中携带会话的会话时长预测值和发包频率预测值。

[0080] 负载均衡节点接收到节点获取请求时，获得目标会话的会话时长预测值和发包频率预测值，然后基于各个计算节点的当前运行状态数据(比如当前CPU使用率、当前GPU使用率、当前的任务并发量、当前的最佳任务并发量)当前处理会话的发包频率分布、当前处理会话的发起时刻分布以及目标会话的会话时长预测值和发包频率预测值，从多个计算节点中筛选出目标计算节点，然后将目标计算节点的信息(比如地址)反馈给客户端，如此，客户端便可直接基于目标计算节点的信息将目标会话的音频计算任务发送至目标计算节点。

[0081] 优选的，目标计算节点在接收到目标会话的音频计算任务时，和/或，处理完目标

会话的音频计算任务时，更新自身的运行状态数据。

[0082] 在一种可能的实现方式中，目标计算节点可在未接收到音频计算任务时，获取自身的运行状态数据并记录，后续在每接收到一音频计算任务和/或每处理完一音频计算任务时，获取自身的当前运行状态数据，对记录的运行状态数据进行更新，如此，不管是使用负载均衡节点采集还是计算节点上报的方式，负载均衡节点均能获得计算节点的最新运行状态数据。

[0083] 以下对目标计算节点在接收到目标会话的音频计算任务时，和/或，处理完目标会话的音频计算任务时，获取自身的当前运行状态数据的过程进行介绍：

[0084] 对于当前运行状态数据中的当前CPU使用率、当前GPU使用率，目标计算节点可调用操作系统的应用程序编程接口(Application Programming Interface，API)获得。

[0085] 对于当前并发任务数、当前处理会话的发包频率、当前处理会话的发起时刻，目标计算节点可通过对其实时记录的数据访问情况进行统计分析获得。

[0086] 对于当前运行状态数据中的当前最佳任务并发量，目标计算节点在接收到目标会话的音频计算任务后时，可获取其处理前一音频计算任务后确定的最佳任务并发量作为当前最佳任务并发量，而目标计算节点在处理完目标会话的音频计算任务时，可首先获取其处理前一音频计算任务(目标会话对应的音频计算任务的前一音频计算任务)后确定的最佳任务并发量作为当前最佳任务并发量，然后基于处理目标会话对应的音频计算任务时的实时率和响应时间，确定是否需要对当前最佳任务并发量进行调整，若需要调整，则基于预设的调整规则对当前最佳任务并发量进行调整，调整后的最佳任务并发量作为当前最终的最佳任务并发量。

[0087] 需要说明的是，在大规模集群中，计算节点的资源分配是实时变化的，因此，计算节点的最佳任务并发量也需要根据实际的运行情况进行调整，基于此，本实施例中的计算节点在每处理完一音频计算任务后，均会基于处理该音频计算任务时的实时率和响应时间对最佳任务并发量进行是否需要调整的判别。

[0088] 下面对目标计算节点处理完目标会话的音频计算任务时，获取自身当前的最佳任务并发量进行介绍，请参阅图2，示出了其实现过程的流程示意图，可以包括：

[0089] 步骤S201a：若目标会话的音频计算任务为首个音频计算任务，则获取预先确定的初始最佳任务并发量作为当前的最佳任务并发量。

[0090] 步骤S201b：若目标会话的音频计算任务并非首个任务，则获取其处理完上一音频计算任务后确定的最佳任务并发量，作为当前的最佳任务并发量。

[0091] 步骤S202：基于当前的最佳任务并发量、当前的任务并发量和处理目标会话的音频计算任务时的实时率和响应时间，判断是否需要对当前的最佳任务并发量进行调整，若是，则执行步骤S203，若否，则不对当前的最佳任务并发量进行调整。

[0092] 具体的，当当前的任务并发量与当前的最佳任务并发量的差值的绝对值小于预设的第一阈值时(即当前的任务并发量接近当前的最佳任务并发量)，若处理目标会话的音频计算任务时的实时率与基准实时率的差值的绝对值大于预设的第二阈值，和/或，处理目标会话的音频计算任务时的响应时间与基准响应时间的差值的绝对值大于预设的第三阈值，则确定需要对当前的最佳任务并发量进行调整。

[0093] 步骤S203：基于预设的调整步长逐步对当前的最佳任务并发量进行调整，调整后

的最佳任务并发量作为当前最终的最佳任务并发量。

[0094] 具体的，可基于下式对当前的最佳任务并发量进行调整：

[0095] $B' = (1 \pm ns)B$ (1)

[0096] 其中，n为调整次数，s为调整步长，B为调整前的最佳任务并发量，B′为调整后的最佳任务并发量。

[0097] 需要说明的是，目标计算节点首次进行调整时，B为预先确定出的、初始最佳任务并发量。可以理解的是，目标计算节点在每处理完一个音频计算任务后，都会基于其处理该音频计算任务时的实时率和响应时间判断前一次确定的最佳任务并发量是否需要调整，如此，就需要有一个初始最佳任务并发量，也就是说，后续的最佳任务并发量是在初始最佳任务并发量的基础上调整得到的。

[0098] 以下对目标计算节点确定初始的最佳任务并发量的过程进行介绍。

[0099] 请参阅图3，示出了目标计算节点确定初始最佳任务并发量的流程示意图，可以包括：

[0100] 步骤S301：获取自身的CPU型号、CPU实际核数、GPU型号、音频处理模型大小，以及自身未执行音频计算任务时的内存可用空间大小和显存可用空间大小。

[0101] 具体的，目标计算节点可调用操作系统的API接口获取自身的CPU型号、CPU实际核数、GPU型号、音频处理模型大小，以及自身未执行音频计算任务时的内存可用空间大小和显存可用空间大小。

[0102] 步骤S302：基于自身的CPU型号、CPU实际核数、GPU型号、音频处理模型大小，以及自身未执行音频计算任务时的内存可用空间大小和显存可用空间大小，获取基于CPU算力和GPU算力评估的第一最大任务并发量、基于内存空间评估的第二最大任务并发量、以及基于显存空间评估的第三最大任务并发量。

[0103] 具体的，步骤S302的实现过程可以包括：

[0104] 步骤S3021：基于CPU型号、CPU实际核数确定CPU实际算力，基于GPU型号确定GPU实际算力，并基于CPU实际算力和GPU实际算力确定第一最大任务并发量。

[0105] 其中，基于获取的GPU型号确定GPU实际算力的过程可以包括：基于获取的GPU型号确定GPU标准算力，作为CPU实际算力。具体的，目标计算节点可基于预先获得的GPU型号与GPU标准算力的对应关系表，确定与获取的GPU型号对应的GPU标准算力。

[0106] 其中，基于获取的基于CPU型号、CPU实际核数确定CPU实际算力的过程可以包括：基于CPU型号获得CPU标准算力和CPU标准核数，基于CPU标准算力、CPU标准核数和CPU实际核数，确定CPU实际算力。具体的，可基于预先获得的CPU型号与CPU标准核数和CPU标准算力的对应关系表，确定与获取的CPU型号对应的CPU标准算力、CPU标准核数。在获得CPU标准算力、CPU标准核数和CPU实际核数后，可通过下式确定CPU实际算力：

[0107] $$C_c = \frac{C \cdot P}{C_k} \quad (2)$$

[0108] 其中，C为CPU标准算力，P为CPU实际核数，C_k为CPU标准核数，C_c为CPU实际算力。

[0109] 在获得GPU实际算力和CPU实际算力后，可基于GPU实际算力和CPU实际算力，利用预先获得的关系函数f，确定该节点的最大并发量作为第一并发量，其中，关系函数以CPU算

力和GPU算力为自变量，以最大任务并发量为因变量，具体的：

[0110] $B_1=f(C_c,G_c)$ (3)

[0111] 其中，C_c为CPU实际算力，G_c为GPU实际算力，B_1为基于CPU实际算力和CPU实际算力确定的第一最大任务并发量。

[0112] 步骤S3022：基于内存可用空间大小、每条会话占用内存空间大小和音频处理模型大小确定第二最大任务并发量。

[0113] 本步骤的目的在于，确定计算节点内存受限情况下的最大任务并发量。

[0114] 由于音频处理模型会占用一些内存空间，每条会话也会占用一些内存空间，通过内存可用空间大小和音频处理模型所占内存空间大小可获得会话可用内存空间大小，基于会话可用内存空间大小以及每路会话占用内存空间大小，可确定出可处理的会话数，即最大任务并发量。

[0115] 具体的，可基于内存可用空间大小、每条会话占用内存空间大小和音频处理模型大小，利用下式确定第二最大任务并发量：

[0116] $$B_2=\frac{M-A_m r_1}{S_m} \quad (4)$$

[0117] 其中，M为内存可用空间大小，A_m为音频处理模型大小，S_m为每条会话所占内存空间大小，r_1为音频处理模型大小与音频处理模型所占内存空间大小的比值，B_2即为第二最大任务并发量。

[0118] 步骤S3023：基于显存可用空间大小、每条会话占用显存空间大小和音频处理模型大小，确定第三最大任务并发量。

[0119] 由于音频处理模型会占用一些显存空间，每条会话也会占用一些显存空间，通过显存可用空间大小和音频处理模型所占显存空间大小可获得会话可用显存空间大小，基于会话可用显存空间大小以及每路会话占用显存空间大小，可确定出可处理的会话数，即最大任务并发量。

[0120] 具体的，可基于显存可用空间大小、每条会话占用显存空间大小和音频处理模型大小，利用下式确定第三最大任务并发量：

[0121] $$B_3=\frac{G_m-A_m r_2}{S_{gm}} \quad (5)$$

[0122] 其中，G_m为显存可用空间大小，A_m为音频处理模型大小，S_{gm}为每条会话占用显存空间大小，r_2为音频处理模型大小与音频处理模型所占显存空间大小的比值，B_3即为第三最大任务并发量。

[0123] 步骤S303：将第一最大任务并发量、第二最大任务并发量和第三最大任务并发量中的最小并发量确定为初始最佳任务并发量。

[0124] 需要说明的是，上述的关系函数f、每条会话所占内存空间大小S_m、音频处理模型大小与音频处理模型所占内存空间大小的比值r_1、每条会话所占的显存空间大小S_{gm}，音频处理模型大小与音频处理模型所占显存空间大小的比值r_2预先获得。

[0125] 在一种可能的实现方式中，可在计算节点上线之前，将计算节点接入内存不受限

的基准测试环境，在该基准测试环境中，针对每个计算节点，测试满足基准实时率和基准响应时间要求的最大任务并发量并记录，同时记录CPU算力和GPU算力，即针对每个计算节点，获得一组数据(包括最大任务并发量、CPU算力和GPU算力)，如此可获得多组数据，基于获得的多组数据可构建出以CPU算力和GPU算力为自变量、以最大任务并发量为因变量的关系函数。

[0126] 在测试的过程中，获取计算节点将音频处理模型加载到内存时所占用内存空间的大小，即音频处理模型所占内存空间大小，然后计算音频处理模型大小与音频处理模型所占内存空间大小的比值，即得到上述的r_1；同样的，获得计算节点将音频处理模型加载到显存时所占用显存空间的大小，即音频处理模型所占显存空间大小，然后计算音频处理模型大小与音频处理模型所占显存空间大小的比值，即得到上述的r_2。

[0127] 在测试的过程中，当发起单路会话时，获取内存占用增长量，作为每路会话所占用内存空间大小S_m，同样的，当发起单路会话时，获取显存占用增长量，作为每路会话所占用显存空间大小S_{gm}。

[0128] 本申请实施例提供的负载均衡方法，一方面，在挑选目标计算节点时，综合考虑计算节点的运行状态、会话时长等影响因素，另一方面，为了避免负载均衡节点出现明显流量瓶颈的问题，采用基于旁路模式的均衡策略，即，客户端在获得目标计算节点的信息后，将目标会话的音频计算任务直接发送至目标计算节点，而不再经负载均衡节点，如此，负载均衡节点的流量大大减少，再一方面，计算节点能够基于自身的运行情况实时调节自身的最佳并发能力，从而使自身能够以最好的状态对外提供服务，综上，本申请实施例提供的负载均衡方法提高了系统整体资源的使用率，降低了系统性能波动，且解决了负载均衡节点的流量瓶颈问题。

[0129] 本申请实施例还提供了一种负载均衡系统，请参阅图4，示出了该负载均衡系统的结构示意图，其可以包括：负载均衡节点401和多个计算节点。

[0130] 负载均衡节点401，用于接收客户端针对一条目标会话发送的节点获取请求，所述节点获取请求用于请求获取所述负载均衡系统中处理所述目标会话的目标计算节点；基于获取的筛选参考数据，从所述多个计算节点中筛选出目标计算节点402，并将目标计算节点402的信息发送至客户端400，其中，筛选参考数据至少包括多个计算节点的当前运行状态数据。

[0131] 负载均衡节点401，还用于按预设的采集周期采集各计算节点的运行状态数据并记录；或者，接收各计算节点按预设的上报周期上报的自身的运行状态数据并记录；任一计算节点的当前运行状态数据为负载均衡节点记录的该计算节点的最新运行状态数据。

[0132] 多个计算节点中的目标计算节点402，用于接收客户端发送的目标会话的音频计算任务，并对音频计算任务进行计算。

[0133] 考虑到计算节点的压力实时变化，本申请实施例提供的负载均衡系统中，负载均衡节点在接收到客户端针对目标会话发起的节点获取请求时，至少基于多个计算节点的当前运行状态从多个计算节点中挑选目标计算节点，为了避免负载均衡节点出现明显流量瓶颈的问题，本申请提供基于旁路模式的均衡策略，即，客户端在获得目标计算节点的信息后，将目标会话的音频计算任务直接发送至目标计算节点，而不再经负载均衡节点，如此，负载均衡节点的流量大大减少，本申请提供的负载均衡系统提高了系统整体资源的使用

率,降低了系统性能波动,且解决了负载均衡节点的流量瓶颈问题。

[0134] 在一种可能的实现方式中,负载均衡节点401获取的筛选参考数据还包括:目标会话的会话时长预测值和发包频率预测值。

[0135] 其中,目标会话的会话时长预测值和发包频率预测值采用与目标会话对应的场景标识和用户标识所对应的会话参数预测模型预测得到。

[0136] 在一种可能的实现方式中,负载均衡节点401获取的筛选参考数据中,任一计算节点的当前运行状态数据包括以下数据中的一种或多种:当前CPU使用率、当前GPU使用率、当前的任务并发量、当前的最佳任务并发量。

[0137] 在一种可能的实现方式中,所述筛选参考数据还包括:各个计算节点当前处理会话的发包频率和/或各个计算节点当前处理会话的发起时刻。

[0138] 在一种可能的实现方式中,目标计算节点402,还用于在接收到目标会话的音频计算任务时,和/或,处理完目标会话的音频计算任务时,获取自身的当前运行状态数据;将自身的当前运行状态数据上报至所述负载均衡节点,以使负载均衡节点对其针对目标计算节点记录的运行状态数据进行更新。

[0139] 在一种可能的实现方式中,目标计算节点402在处理完所述音频计算任务时,获取自身的当前最佳任务并发量时,具体用于:若所述目标会话的音频计算任务为首个任务,则获取预先确定的初始最佳任务并发量作为当前的最佳任务并发量;若所述目标会话的音频计算任务并非首个任务,则获取其处理完上一音频计算任务后确定的最佳任务并发量,作为当前的最佳任务并发量;基于所述当前的最佳任务并发量、当前的任务并发量和处理所述目标会话的音频计算任务时的实时率和响应时间,确定是否需要对所述当前的最佳任务并发量进行调整;若需要对所述当前的最佳任务并发量进行调整,则基于预设的调整步长逐步对所述当前的最佳任务并发量进行调整;调整后的最佳任务并发量作为当前最终的最佳任务并发量。

[0140] 在一种可能的实现方式中,目标计算节点402在基于所述当前的最佳任务并发量、当前的任务并发量和处理所述目标会话的音频计算任务时的实时率和响应时间,确定是否需要对所述当前的最佳任务并发量进行调整时,具体用于:当所述当前的任务并发量与所述当前的最佳任务并发量的差值的绝对值小于预设的第一阈值时,若处理所述目标会话的音频计算任务时的实时率与基准实时率的差值的绝对值大于预设的第二阈值,和/或,处理所述目标会话的音频计算任务时的响应时间与基准响应时间的差值的绝对值大于预设的第三阈值,则确定需要对所述当前的最佳任务并发量进行调整。

[0141] 在一种可能的实现方式中,目标计算节点402在确定初始最佳任务并发量时,具体用于获取自身的CPU型号、CPU实际核数、GPU型号、音频处理模型大小,以及自身未执行音频计算任务时的内存可用空间大小和显存可用空间大小;基于自身的CPU型号、CPU实际核数、GPU型号、音频处理模型大小,以及自身未执行音频计算任务时的内存可用空间大小和显存可用空间大小,获取基于CPU算力和GPU算力评估的第一最大任务并发量、基于内存空间评估的第二最大任务并发量、以及基于显存空间评估的第三最大任务并发量;将所述第一最大任务并发量、所述第二最大任务并发量和所述第三最大任务并发量中的最小并发量确定为所述初始最佳任务并发量。

[0142] 最后,还需要说明的是,在本文中,诸如第一和第二等之类的关系术语仅仅用来将

一个实体或者操作与另一个实体或操作区分开来，而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括所述要素的过程、方法、物品或者设备中还存在另外的相同要素。

[0143] 本说明书中各个实施例采用递进的方式描述，每个实施例重点说明的都是与其他实施例的不同之处，各个实施例之间相同相似部分互相参见即可。

[0144] 对所公开的实施例的上述说明，使本领域专业技术人员能够实现或使用本申请。对这些实施例的多种修改对本领域的专业技术人员来说将是显而易见的，本文中所定义的一般原理可以在不脱离本申请的精神或范围的情况下，在其它实施例中实现。因此，本申请将不会被限制于本文所示的这些实施例，而是要符合与本文所公开的原理和新颖特点相一致的最宽的范围。

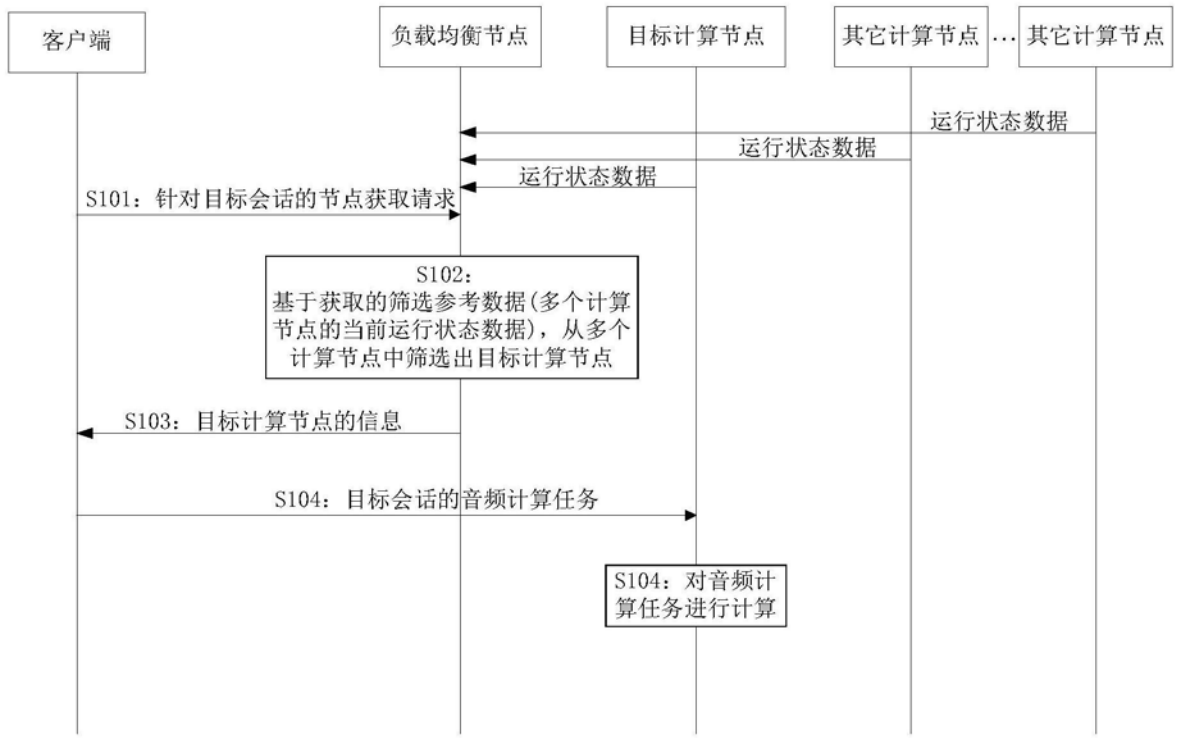

客户端
负载均衡节点
目标计算节点
其它计算节点
其它计算节点
运行状态数据
运行状态数据
运行状态数据
S101：针对目标会话的节点获取请求
S102：
基于获取的筛选参考数据(多个计算节点的当前运行状态数据)，从多个计算节点中筛选出目标计算节点
S103：目标计算节点的信息
S104：目标会话的音频计算任务
S104：对音频计算任务进行计算

图1

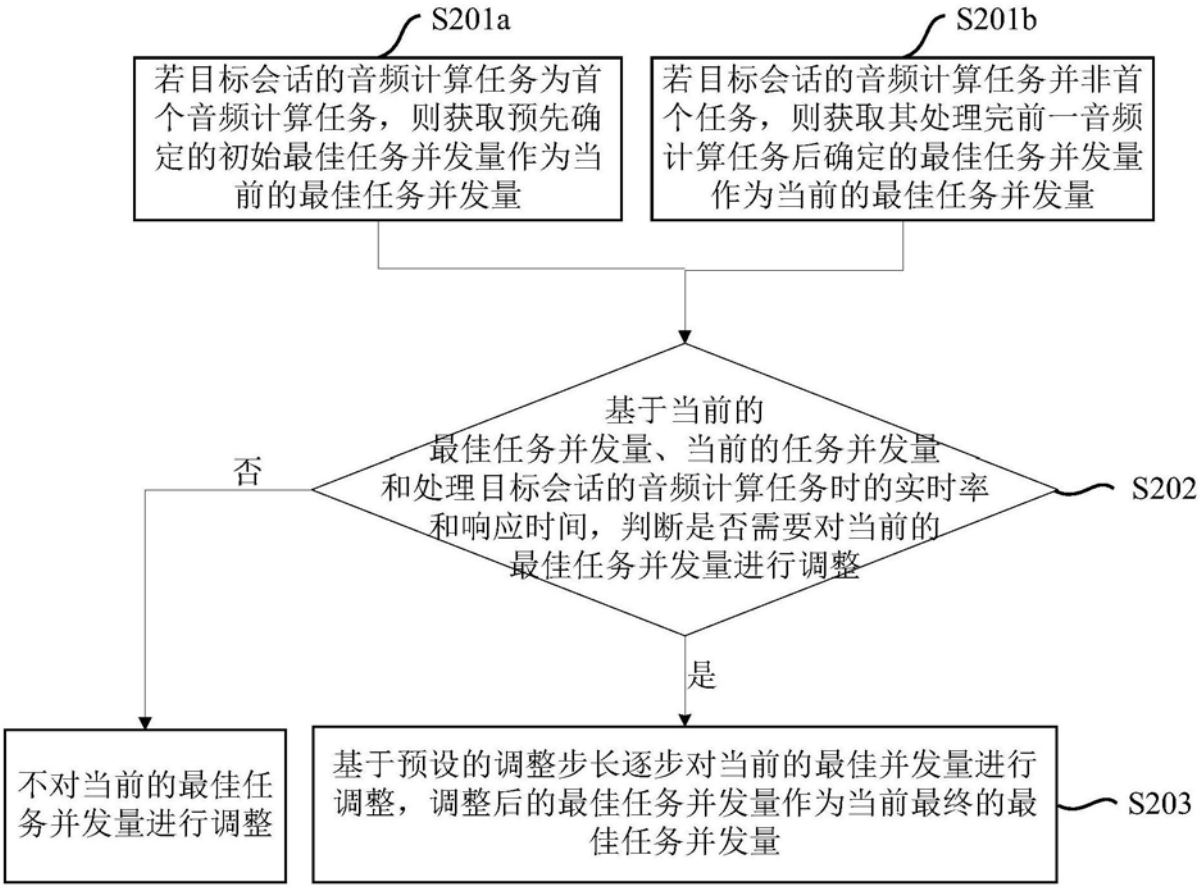

S201a
若目标会话的音频计算任务为首个音频计算任务，则获取预先确定的初始最佳任务并发量作为当前的最佳任务并发量
S201b
若目标会话的音频计算任务并非首个任务，则获取其处理完前一音频计算任务后确定的最佳任务并发量作为当前的最佳任务并发量
基于当前的最佳任务并发量、当前的任务并发量和处理目标会话的音频计算任务时的实时率和响应时间，判断是否需要对当前的最佳任务并发量进行调整
S202
否
是
不对当前的最佳任务并发量进行调整
基于预设的调整步长逐步对当前的最佳并发量进行调整，调整后的最佳任务并发量作为当前最终的最佳任务并发量
S203

图2

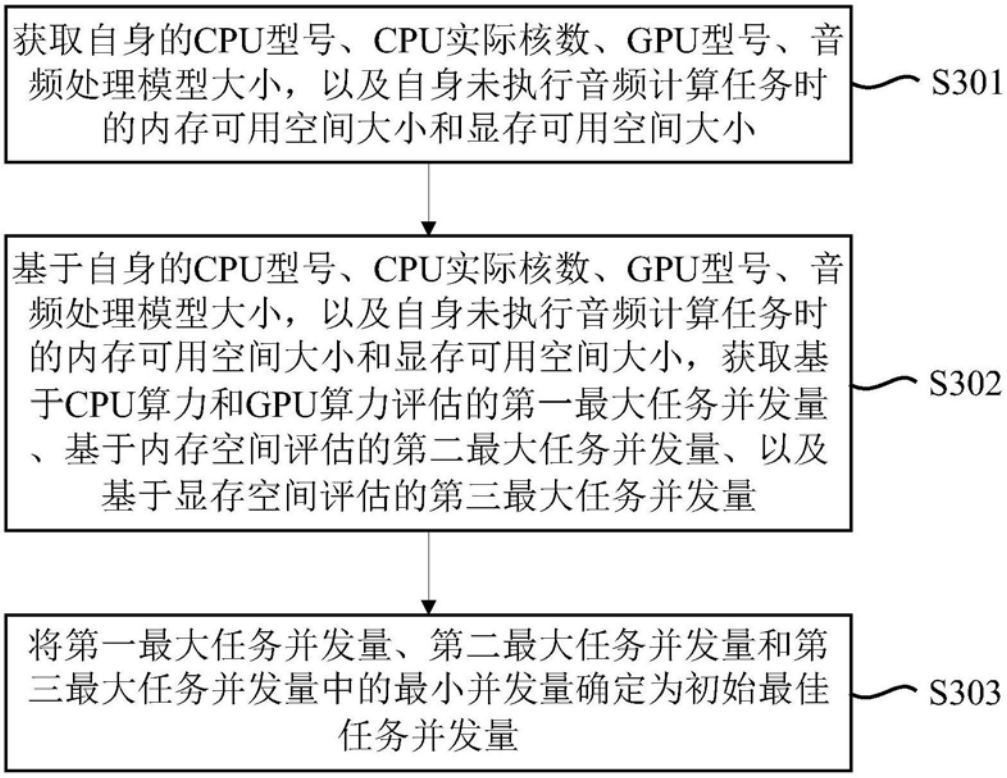

获取自身的CPU型号、CPU实际核数、GPU型号、音频处理模型大小，以及自身未执行音频计算任务时的内存可用空间大小和显存可用空间大小
S301
基于自身的CPU型号、CPU实际核数、GPU型号、音频处理模型大小，以及自身未执行音频计算任务时的内存可用空间大小和显存可用空间大小，获取基于CPU算力和GPU算力评估的第一最大任务并发量、基于内存空间评估的第二最大任务并发量、以及基于显存空间评估的第三最大任务并发量
S302
将第一最大任务并发量、第二最大任务并发量和第三最大任务并发量中的最小并发量确定为初始最佳任务并发量
S303

图3

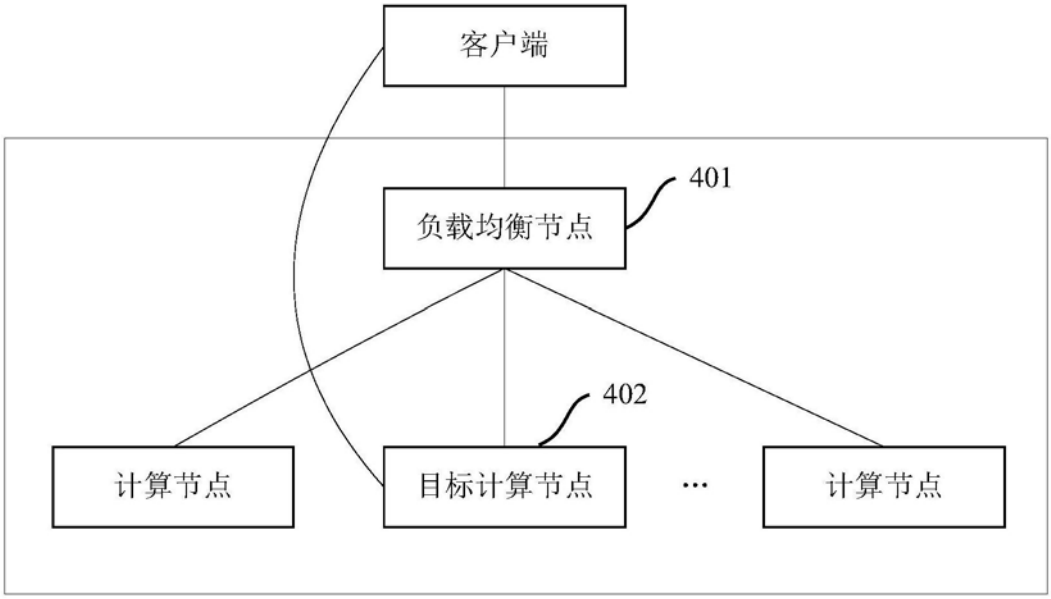

客户端
401
负载均衡节点
402
计算节点
目标计算节点
...
计算节点

图4