# 绪论

## 课题背景与研究意义

资料显示，自互联网技术诞生以来，尤其是移动互联网的迅猛发展，目前，仅以手机终端为例，全球移动手机终端已经达到xx台。因而，如何满足数以几十亿计的用户访问请求，并保证用户服务的高性能、高可用与高可拓展性，给提供用户服务的后端服务器结构设计带来了巨大的挑战。

为解决这一问题，为用户提供及时、可靠、高效的网络服务，同时保证后端服务器的较低消耗，很多服务器结构设计方案被提出。根据用户请求规模和实际业务场景的不同，主要有单一服务器和服务器集群这两种方案。

为应对高并发用户访问请求，单一服务器方案主要通过升级服务器硬件配置，提升单台服务器性能这一方式来解决。例如，使用性能更高的CPU、容量更大的内存、更快的数据传输协议等等。显然，通过升级硬件配置的方式来提升单台服务器设备的性能这一方法是有限制的，在高并发用户请求场景中，该方法显然不能奏效。除此之外，对互联网企业而言，如何在满足用户请求的前提下尽可能降低服务器消耗成本是其要考虑的核心问题。显然，提升单一服务器端的硬件性能无法满足这一要求。

针对上述单一服务器方案存在的诸多弊端，服务器集群方案与服务器集群技术应运而生。服务器集群方案主要有以下几个特点。第一，服务器集群对单个服务器端的性能要求不高，不需要为每台服务器配置最佳的硬件性能；第二，服务器集群借助用户请求分配方案，将不同用户请求分发至不同的后端服务器，以满足不同用户请求；第三，服务器集群需要合理、高效的集群管理方案和架构设计，统一管理集群中用户请求分发、服务器资源迁移等工作，以保证服务器集群的高性能、高可用与高可拓展性。基于此，集群负载均衡是服务器集群管理方案中一项重要的功能和性能要求。负载均衡的核心思想是，通过一台“中转”路由服务器，根据后端集群中不同服务器的性能以及用户请求等信息，将最适合当前用户请求的后端服务器分配给该用户请求，从而实现集群中不同服务器之间的负载和性能均衡。

但是，目前服务器集群负载均衡存在以下两方面的瓶颈：

第一，高并发用户请求导致网络链路频频发生拥堵，致使数据传输过程中发生数据包延迟甚至丢失，由此导致整体网络为用户提供的服务能力大大降低。因而即使在服务器处理能力足够的情况下也可能因为网络链路拥塞的问题降低整体效率，甚至出现某些服务器空载的情况，造成服务器资源的严重浪费。

第二，当集群系统中服务器在计算速度、通信能力以及存储容量等自身性能方面存在较大差异时，不能充分考虑服务器性能对集群进行动态负载调度会产生不合理的任务分配，导致部分服务器和局部网络链路过载的同时，某些服务器和链路处于空载甚至空闲状态。

就软件方面而言，目前常用的负载均衡算法可以分为基于传统软件的方法和基于流量预测的方法。基于软件方法的负载均衡调度策略主要有基于随机选择任务移动节点的概率调度算法、根据负载变化差额而基于梯度模型的调度算法以及自适应的近邻契约算法等；基于流量预测的动态负载均衡调度策略主要有基于用户请求流量的负载均衡算法和基于服务器负载流量预测的负载均衡方法，这些方法根据预测结果制定负载均衡策略。

## 国内外研究现状

### 集群与集群技术发展

在以云计算为代表的高并发用户请求场景中，通过提高CPU主频、增加内存容量以及拓展总线带宽等方式提升单台服务器的性能显然无法应对大量、高频用户访问请求。服务器集群技术的发展，为高并发应用场景提供了一套更好的解决方案。云计算是分布式计算的一种，是集群技术的一种典型应用场景。

集群是一组相互独立的、通过网络互联的计算机，它们构成了一个组，并以单一系统的模式加以管理。集群系统中的每一个服务器节点都是一台独立的物理设备，其他节点的状态变化不会影响该节点的正常运作。在集群系统运行过程中，若单个服务器节点出现宕机等故障导致其不能继续提供服务，集群系统会选出下一个节点作为该业务运行的替代载体，以保证集群服务的高可用性。因此，用户与集群相互作用时，对于用户而言，一个集群相当于一台单独的服务器。一个可靠的服务器集群系统应具备高性能、高可用和高可拓展等特性。

集群技术是一种服务器架构技术，其通过硬件或软件技术将一些独立的服务器组织在一起，共同处理高并发用户请求。集群技术可以有效提高集群系统的高性能、高可用和高可拓展等性能。集群技术解决了单个服务器界定啊运算能力不足、IO性能不足等问题，提高了集群服务的可靠性，使集群获得可拓展能力，降低集群整体的运维成本。根据组成集群系统的计算机之间的体系结构是否相同，集群可分为同构集群与异构集群。根据业务场景和技术点的不同，集群可分为三种类型，即高可用集群、高性能集群和负载均衡集群。每种集群的介绍如下。

高可用集群一般是指当集群中某个节点发生失效时，其上的任务会自动转移到其他正常的节点上。还指可以将集群中的某节点进行离线维护再上线，该过程并不影响整个集群的运行。当节点中运行任务的应用程序出现故障，或者系统硬件如网络出现故障时，集群可以将任务自动、快速地从一个节点切换到另一个节点，从而保证集群持续、不间断地对外提供服务。

高性能计算集群将计算任务分配到集群的不同节点而提高计算能力，因而主要应用在科学计算领域。比较流行的高性能计算集群采用Linux操作系统和其他一些免费软件来完成并行运算。这一集群配置通常被称为Beowulf集群。这类集群通常运行特定的程序以发挥高性能计算集群的并行能力。这类程序一般应用特定的运行库，比如专为科学计算设计的MPI库。高性能计算集群适用于在计算中各计算节点之间发生大量数据通讯的计算作业，比如一个节点的中间结果或影响到其他节点计算结果的情况。

负载均衡集群由两台或者两台以上的服务器组成，分为前端负载调度和后端服务两个部分。负载调度部分负载把用户的请求按照不同的策略分配给后端服务节点，而后端节点是真正提供应用程序服务的部分。与高可用集群不同的是，负载均衡集群中，所有的后端节点都处于活动状态，它们都对外提供服务，分摊系统的工作负载。

负载均衡集群可以把一个高负载的用户请求分散到多个节点共同完成，适用于高并发、大负载访问的应用系统。但是它也有不足的地方：当一个节点出现故障时，前端调度系统并不知道此节点已经不能提供服务，仍然会把用户请求调度到故障节点上来，这样访问就会失败。为了解决这个问题，负载调度系统一般都引入了节点监控系统。节点监控系统位于前端负载调度机上，负责监控下面的服务节点。当某个节点出现故障后，节点监控系统会自动将故障节点从集群中剔除；当此节点恢复正常后，节点监控系统又会自动将其加入集群中，而这一切对用户来说是完全透明的。

### 负载均衡技术发展

参考：<https://kns.cnki.net/KXReader/Detail?invoice=r5o1Rkq2g3R0YDW28ISmR%2FhFTd0nuBw9fLsgOOf8RZ0kcqo0ZzUdvkkpt7StZLqgru5G%2Fw%2BQ%2FyxIE6H9DVxdL%2F2h0WXyucHVQBJqAH%2FtoK3Sy28SmzznMrbtn76zmDuaxABfDzXCv6vo6HCaEWOJ7A1LowAlve9kEi6EOO%2F%2B6DI%3D&DBCODE=CJFD&FileName=HBYD202111030&TABLEName=cjfdlast2021&nonce=C210A84FEAE343B69610C02A6FDF1162&uid=&TIMESTAMP=1669278085628>

负载均衡技术定义：

负载均衡技术于1996年由Foundry公司提出，是现代计算机领域的基础技术之一。其基本原理是通过运行在前端的负载均衡服务器，根据执行的负载均衡算法，将用户请求分配到后端服务器节点上，从而提高整个系统的扩展能力，实现服务的并行扩展。同时，负载均衡技术还可以起到对外网屏蔽内网服务器的作用，从而提高系统的可用性。一般来说，负载均衡技术具有两个方面的含义：一方面，负载均衡技术对用户请求进行了合理分配，后端多台服务器设备共同处理任务，使得整个集群系统的处理能力大大提高；另一方面，由单台服务器优化为多台服务器处理，缩短了集群系统响应和用户等待的时间。（参考：<https://kns.cnki.net/KXReader/Detail?invoice=QclEwPxFxYuDpZc%2BjqB%2BW%2FHhO1DyWCWsqbLIZ%2F0YiADnfWl2ctl%2FUJVnfPbwjrhAfSSDJ7IJ6XBVVmCr8YUzMTRrSQ7OCu0ABwy1CNcF2TCPJKI0MOIYAJcJthSWpuv4dLh5P8XmmK0u0gR00Rz9eWYTdnhH9cxUnO5SEUIP5HI%3D&DBCODE=CJFD&FileName=KJNJ202204027&TABLEName=cjfdlast2022&nonce=DB9B30888EC14751A4E4C009B772CBC2&uid=&TIMESTAMP=1669277945920>）

在云计算场景中，负载均衡技术主要包含两层含义：其第一层含义中负载均衡主要体现在执行任务的虚拟机层当中；另外一层含义中负载均衡体现在实体机层当中。（参考：https://kns.cnki.net/KXReader/Detail?invoice=mCBLdm0oIxYpivXqSEzVT636rACcQBtyck8W249GkQ26lxvkA6GPyZt5MBjLZepoRVd30V%2FOV5zBtHuquZGtjYg1eDxDhod3a9LVjr2NSkzt1lgklS9IH0gbAvaItjnt1%2BUlfeJGkJvzGS2nxkuu7pqZxDKbjuCgH6gzgInexIU%3D&DBCODE=CJFD&FileName=DZRU202123061&TABLEName=cjfdlast2022&nonce=83DADBCEC15A4AAB88405103667CBD1B&uid=&TIMESTAMP=1669277632332）

负载均衡技术的分类：（参考：https://blog.csdn.net/hui540/article/details/8950927）

针对不同的应用场景，目前有多种不同的负载均衡技术以满足不同的用户请求。根据负载均衡所采用的设备对象、负载均衡的作用范围以及应用的网络层次等三个方面，负载均衡技术可以分为以下几类。

软/硬件负载均衡

软件负载均衡解决方案是指在一台或多台服务器的操作系统上安装一个或多个附加软件来实现负载均衡，该种解决方案基于特定环境，配置简单，使用灵活，成本低廉，可以满足一般的负载均衡需求。

当然，软件解决方案缺点也较多。因为每台服务器上安装额外的软件运行会消耗系统不定量的资源，越是功能强大的模块，消耗得越多。所以当连接请求特别大的时候，软件本身会成为服务器工作成败的一个关键；受操作系统的限制，软件可扩展性并不是很好；另外，操作系统本身存在的一些问题，往往会引起集群安全问题。

硬件负载均衡解决方案是直接在服务器和外部网络间安装负载均衡设备，这种设备通常被称为负载均衡器。由于专门的设备完成专门的任务，硬件负载均衡器独立于操作系统，其整体性能得到极大提高；加上多样化的负载均衡策略，智能化的流量管理，硬件负载均衡解决方案可达到最佳的负载均衡效果。

硬件负载均衡器有多种多样的形式，除了作为独立意义上的负载均衡器外，有些负载均衡器集成在交换设备中，置于服务器与公共网络之间；有些则通过两块网络适配器将这一功能集成到服务器中，一块连接到公共网络上，一块连接到后端服务器群的内部网络上。

整体而言，硬件负载均衡解决方案在功能、性能上优于软件方式，不过成本昂贵。

本地/全局负载均衡

本地负载均衡是指对本地的服务器集群做负载均衡，全局负载均衡是指对分别放置在不同的地理位置、有不同网络结构的服务器集群间作负载均衡。

本地负载均衡能有效地解决数据流量过大、网络负载过重的问题，并且不需花费昂贵开支购置性能卓越的服务器，充分利用现有设备，避免服务器单点故障造成数据流量的损失。其有灵活多样的负载均衡策略把用户请求流量合理地分配给服务器集群内的服务器共同负担。若需要为现有服务器扩充升级，只需简单地增加一个新的服务器到服务集群中，而不需改变现有网络结构、停止现有的服务。

全局负载均衡适用于服务器节点分布在不同区域的集群。该负载均衡方案可以使用户只以一个IP地址或域名就能访问到离自己最近的服务器，从而获得最快的访问速度。该方案也可用于子公司站点分散较广的大公司，通过企业内部互联网来达到资源统一合理分配的目的。

全局负载均衡有以下的特点：

第一，实现地理位置无关性，能够远距离为用户提供完全的透明服务；

第二，除了能避免服务器、数据中心等的单点失效，也能避免由于ISP专线故障引起的单点失效；

第三，解决网络拥塞问题，提高服务器响应速度，服务就近提供，实现更好的访问质量。

不同网络层次的负载均衡

针对网络上负载过重的不同瓶颈所在，从网络的不同层次入手，我们可以采用相应的负载均衡技术来解决现有问题。现代负载均衡技术通常操作于网络的第四层或第七层。

第四层负载均衡将一个公共网络上合法注册的IP地址映射为多个内部服务器的IP地址，对每次TCP连接请求动态使用其中一个内部IP地址，达到负载均衡的目的。在第四层交换机中，此种均衡技术得到广泛的应用，一个目标地址是服务器群VIP（虚拟IP，Virtual IP address）连接请求的数据包流经交换机，交换机根据源端和目的IP地址、TCP或UDP端口号和一定的负载均衡策略，在服务器IP和VIP间进行映射，选取服务器群中性能最佳的服务器来处理连接请求。

第七层负载均衡控制应用层服务的内容，提供了一种对访问流量的高层控制方式，适合对HTTP服务器群的应用。第七层负载均衡技术通过检查流经的HTTP报头，根据报头内的信息来执行负载均衡任务。

云计算中负载均衡技术的应用：（参考：<https://www.yisu.com/ask/7781.html）https://kns.cnki.net/KXReader/Detail?invoice=mCBLdm0oIxYpivXqSEzVT636rACcQBtyck8W249GkQ26lxvkA6GPyZt5MBjLZepoRVd30V%2FOV5zBtHuquZGtjYg1eDxDhod3a9LVjr2NSkzt1lgklS9IH0gbAvaItjnt1%2BUlfeJGkJvzGS2nxkuu7pqZxDKbjuCgH6gzgInexIU%3D&DBCODE=CJFD&FileName=DZRU202123061&TABLEName=cjfdlast2022&nonce=83DADBCEC15A4AAB88405103667CBD1B&uid=&TIMESTAMP=1669277632332>

基于集群负载均衡技术实现的高可用和高可靠特性，负载均衡技术的应用主要有以下几个方面。

第一，能够用来解决高并发问题，主要应用于高访问量的业务；

第二，根据业务发展从而扩展应用程序；

第三，可以在负载均衡实例下添加多台云服务器实例，解决单点故障问题；

第四，在各地域部署多可用区，实现同城容灾；

第五，将域名解析到不同地域的负载均衡实例下，实现全局负载均衡，解决跨地域容灾问题。

### 负载均衡技术研究现状

互联网技术与应用的快速普及，伴随互联网终端用户的快速增长，国内外互联网市场均涌现出众多“头部”互联网企业，其旗下产品一般具备上亿甚至十亿级用户。不同企业产品其应用场景也呈现不同特点。为保证用户体验，为用户提供高可用、高可靠服务，服务器集群负载均衡技术的研究得到了众多科研工作者和互联网厂商的广泛关注，该技术也取得了极大地发展。根据负载均衡技术的应用场景，负载均衡技术覆盖了分布式计算、并行计算、网格计算以及云计算等诸多应用和技术场景。根据负载均衡技术的策略和所引用的系统规模，负载均衡技术的发展呈现从静态向动态，从集中式到分布式的发展趋势和特点。根据负载调节方式的不同，集群负载均衡技术可分为静态策略和动态策略；根据负载控制方式的不同，集群负载均衡技术可分为集中式策略和分布式策略。（参考：<https://kreader.cnki.net/Kreader/CatalogViewPage.aspx?dbCode=CMFD&filename=1014380619.nh&tablename=CMFD201501&compose=&first=1&uid=WEEvREcwSlJHSldSdmVpbEs1TGVZcHpHVU1ZL1RvZzlZZkllVjM3QnVkQT0=$9A4hF_YAuvQ5obgVAqNKPCYcEjKensW4IQMovwHtwkF4VYPoHbKxJw!!）>

国内外对集群负载均衡技术的研究侧重略有不同，下面分别为国内外负载均衡技术的研究现状。

就国内研究现状而言，在负载均衡算法的理论研究层面，算法优化是主要研究方向；在负载均衡算法的应用研究层面，对已有负载均衡软件产品进行改进以提高负载均衡软件的可用性是主要研究方向。文献[x]为了解决服务器集群在处理并发任务请求时请求分配不均衡和任务完成时间较长等问题，提出了一种基于布谷鸟搜索的集群负载均衡多目标优化调度算法。首先，依据服务器集群的任务请求分配特点，通过监控、记录服务器实时负载信息，构建与服务器实时负载信息相关的、以最小化任务完成时间和增强负载均衡有效度为目标函数的优化模型，确定决策变量为任务请求与服务器的匹配集。然后通过引入带精英策略的非支配排序布谷鸟搜索算法对决策变量进行迭代寻优，在适应度函数的选择更新下，找到符合全局最优的Pareto解集，调度机制根据确定最优的匹配集进行任务的调整与转发。（参考：https://kns.cnki.net/KXReader/Detail?invoice=cbnNJjOhhltYMAqPdqdEYbbF4wQd6sQKdrQMi2vj7sJBV7O89lADwOaRQrXrAWCWNGNGlLVWUgtsc%2Fq5A2mTU%2B0mMceriK%2FllddCG6iTRVLa3GHbZ2ubLD0Z4MMCFKRPPLipBEVZip%2BpP9fCLkB1nfFTuIWH2GAWTs0LNe9smPU%3D&DBCODE=CJFD&FileName=JSJA2022S1098&TABLEName=cjfdlast2022&nonce=D740AE982B23480B9342F79561AEEEC4&uid=&TIMESTAMP=1669381334958）文献[x]针对物联网中智能应用快速增长导致的移动网络数据拥塞问题，构建了一种基于雾集群协作的云雾混合计算模型，在考虑集群负载均衡的同时引入权重因子以平衡计算时延和能耗，最终实现系统时延能耗加权和最小。为了解决该混合整数非线性规划问题，将原问题分解后采用库恩塔克(KKT)条件和二分搜索迭代法对资源配置进行优化，提出一种基于分支定界的开销最小化卸载算法(BB-OMOA)获得最优卸载决策。（参考：https://kns.cnki.net/kcms/detail/detail.aspx?dbcode=CAPJ&dbname=CAPJLAST&filename=DZYX20221019000&uniplatform=NZKPT&v=OBbwn4s\_s\_DK98bmQWtwJfHYN2se0p4s6\_CCRxpF8bXurZ3EZ09Z4Vu8UKYaRio3）文献[x]针对传统负载均衡算法不能满足公网数字集群系统高并发用户请求和快速呼叫建立等现状，提出一种基于负载反馈的分布式数字集群动态负载均衡算法，实现公网数字集群系统负载均衡，提高用户容量。首先建立参与MCPTT服务器的静态负载和动态负载监控机制和指标；然后利用加权轮询算法为用户分配参与MCPTT服务器，并通过用户请求的处理获得复合负载参数；根据负载指标的反馈更新参与MCPTT服务器权值以动态调整服务器负载。（参考：https://kns.cnki.net/KXReader/Detail?invoice=wtCQ5BWvkmTYWR45BkKuw4OGpgA8SmtBteBaxMnjPHVGfVjXiucvi4fH%2FyyGsbF5Jf%2BdaYf0b1PGUgna1Ads8MNU%2BFne5OJhyYZ9Yst1KGVjw9cmsgpM%2Bqm%2BACn%2FO6%2FVA3FteTHN1BeRtlXnK%2B41cryT7u%2FwZMEPIuLrNz81qEo%3D&DBCODE=CJFD&FileName=JSYJ202202034&TABLEName=cjfdlast2022&nonce=A1815E0A075547A49DBD0C719431A9BA&uid=&TIMESTAMP=1669381322169）文献[x]为了解决Web集群中的高并发访问和资源异构引发的负载不均衡问题,提出了一种基于改进布谷鸟搜索的负载均衡算法。该算法建立了自适应负载分配权重模型,使用基于目标函数的布谷鸟搜索算法寻找最优权重。首先,根据集群任务调度特性定义了负载分配权重,并建立了集群调度的目标函数。其次,将负载分配权重编码为布谷鸟种群个体,使用布谷鸟搜索算法寻找最优个体,结合目标函数进行评估。另外,通过混沌变异增加初始种群的均匀度和离散度,并通过在布谷鸟搜索中引入反向学习,加速了最优权重的输出,根据最优权重将任务调度至集群中的各节点。（参考：https://kns.cnki.net/kcms/detail/detail.aspx?dbcode=CJFD&dbname=CJFDLAST2020&filename=ZJSG202004014&uniplatform=NZKPT&v=MbeZY5coO7AjsY33nNKyn53\_\_\_Caq\_7JcHxlbbqm1IZwpTX4PYWxfaLi-3-LL1gT）文献[x]针对Web集群服务器在处理分配混合页面访问任务时开环负载均衡算法导致局部服务器负载偏高而降低系统效率的问题,建立了一类基于动态调节的闭环负载分配策略。根据服务器处理Web访问页面类型的特点,建立静态页面与动态页面混合处理的负载量模型,形成处理不同服务请求与负载均衡的内在动态映射关系,以此优化静态页面缓存与调用方式,并基于服务器负载率动态预测和均衡指标,采用负载率偏差最小的任务权重最优分配模型,确定服务器集群的最优任务权重,实现了服务器集群处理混合页面访问的负载均衡分配策略和算法。（参考：https://kns.cnki.net/kcms/detail/detail.aspx?dbcode=CJFD&dbname=CJFDLAST2018&filename=JSGG201818015&uniplatform=NZKPT&v=EFV\_LjBRpfmFuKAKvJCD03WtHq0IM6FA\_UToARe3Lu9VAMvlEbO3nfNYgKrM58AA）文献[x]分析Nginx服务器负载均衡的体系架构,研究默认的加权轮询算法,并提出一种改进后的动态负载均衡算法,实时收集负载信息,重新计算并分配权值。（参考：https://kns.cnki.net/KXReader/Detail?invoice=nxuqzbJ2ZCEiiPqwuoSTOElN2LPC%2BCGdqwEWSglxlv9udyIWQoShu%2F8TAKcvnAU1ybc0iK2NBgJBg6gVnTKi24ftbjmwjR8QScsYd%2BXo9vO0m%2FGkqm6fzBPA1LHL32J%2BYSAJdAzEfy%2FUmDb0UTYvEFt%2FGmP1OlwVg%2FCUzs98plw%3D&DBCODE=CJFD&FileName=WJFZ202003014&TABLEName=cjfdlast2020&nonce=048B75ABB0224CD5B85F8644D5A507D2&uid=&TIMESTAMP=1669384678147）文献[x]为了减轻快速增长的网络负载压力,为web后端服务器集群搭建了基于Nginx的负载均衡服务器,将其作为集群的反向代理服务器,使集群具备了负载均衡的功能,对负载均衡算法进行了分析。并针对Nginx自带负载均衡策略的缺陷提出了一种动态自适应负载均衡算法-改进型加权最小连接数算法,同时对其算法进行了设计。（参考：https://kns.cnki.net/kcms/detail/detail.aspx?dbcode=CJFD&dbname=CJFDLAST2017&filename=RJZZ201708004&uniplatform=NZKPT&v=FURYjqmywhl9kF0fbnKDhbaxQHe7dwykQ-krERuviU4uWS5ic8tfqm7jjMvHb3ti）

就国外研究现状而言，在负载均衡技术的理论研究层面，其主要研究方向在于云计算环境下的负载均衡算法优化；在负载均衡算法的应用研究层面同样侧重于对已有负载均衡软件产品进行改进以提高负载均衡软件的可用性。文献[x]提出了一种基于云分区概念的负载平衡模型，该模型具有切换机制，可以针对不同的情况选择不同的策略。该算法将博弈论应用于负载平衡策略，以提高云计算环境中的效率。（参考：https://kns.cnki.net/kcms/detail/detail.aspx?dbcode=CJFD&dbname=CJFD2013&filename=QHDY201301007&uniplatform=NZKPT&v=RDRhQ7Lne2Dczux3-hNMTJ6fpRbR2mGoukJyBwAjKOEyxr7Jr\_Z5gO7PfXDgLNhH）文献[x]提出了一种算法，该算法利用适应度函数和双寡头博弈理论将任务分配给能够处理传入任务的资源需求的物理机器。该算法的核心在于是优化数据中心的负载平衡。当优化发生时，任何物理机器都不会过载或利用不足，从而导致资源的有效利用。（参考：https://schlr.cnki.net/en/Detail/index/GARJ2021\_1/SPQD53572CED01943FD84936A52B9E420359）文献[x]提出了一种基于双阈值的功率感知蜜蜂负载平衡算法，用于将传入的用户请求公平、均匀地分配给所有虚拟机。该算法实现了消耗最少的资源来满足服务需求。（参考：https://schlr.cnki.net/en/Detail/index/GARJ2021\_2/SSJDE5CB873398D02F96EAF2B050A18D84E8）

### 负载预测技术研究现状

在服务器集群系统中，负载均衡和资源分配是实现集群高可用和高可靠性的两项关键技术。其中资源管理和分配是集群系统中控制成本和合理分配服务器计算能力的重要算法，集群系统进行资源管理和分配时一项很重要的参考指标便是各服务器节点的负载。随着国内外研究人员对负载均衡和资源管理技术的深入研究，以及人工智能的快速发展，集群资源管理和分配方案性能参考指标由原来的服务器节点静态负载逐渐转向对服务器节点动态负载。服务器节点动态负载的获取一个重要的方式便是负载预测。实时性要求较高的用户请求，需要集群系统做出高效、准确的任务分配和调度，此时，准确的服务器节点负载预测起到至关重要的作用。基于负载预测的动态负载均衡对于集群高效任务分配和调度而言具有重要意义。因此，越来越多的集群资源管理和分配方案都在利用负载预测技术来提升服务器节点的动态负载获取准确度，进而提升整个集群负载均衡和资源管理的质量和准确度。

集群服务器节点负载预测一直是国内外研究人员的研究热点，目前，国内外在该领域的研究主要集中在服务器节点负载时序流量的预测。根据研究方法的不同，对于时序数据的预测主要有三类方法，分别为基于传统线性回归模型的负载预测方法，基于传统机器学习的负载预测方法和基于深度学习的负载预测方法。

第一类基于传统线性回归模型的负载预测方法主要有自回归[x](Autoregressive,AR)、滑动平均(Moving Average,MA)、自回归移动平均[x](Autoregressive Moving Average, ARMA)以及差分整合移动平均自回归[x](Autoregressive Integrated Moving Average, ARIMA)等模型。文献[x]提出了一种使用包括聚类技术和ARIMA模型的电力负载预测方法。该新方法包括对一整年的数据进行聚类，使用K-means聚类的预测日，并使用结果预测电力的峰值负载。该方法也可以在需求响应中实施，以通过避免在高电价时段期间使用电力来减少电费。（参考：https://onlinelibrary.wiley.com/doi/full/10.1002/2475-8876.12135#:~:text=The%20combination%20of%20clustering%20and%20the%20ARIMA%20model,time%20to%20design%20strategies%20for%20peak%20load%20reduction.）文献[x]应用两个机器学习模型，即AR（自回归）和ARIMA（自回归综合移动平均）来预测（软件定义网络）SDN的流量。SDN流量被视为一个时间序列。从平均绝对百分比误差（MAPE）来看，ARIMA的预测精度高于AR。（参考：https://link.springer.com/chapter/10.1007/978-981-15-5258-8\_75）这些模型在复杂度低、线性关系较强的数据中可以实现较好的预测效果，因此此类方法存在对数据的限制较多，且需要人工调整模型参数等方面的不足。

第二类基于传统机器学习的负载预测方法主要有马尔科夫模型[x]、贝叶斯模型[x]、支持向量回归[x](Support Vector Regression,SVR)模型，以及决策树和传统人工神经网络[x](Artificial Neural Networks, ANN)等模型。文献[x]为了其提高负载短期预测的预测精度, 提出一种改进灰狼搜索算法优化支持向量机（SVM）的短期云计算资源负载预测模型 (EGWO-SVM) 。该模型能够更加准确地刻画云计算短期资源负载的复杂变化趋势, 从而有效提升云计算资源负载短期预测的精度。（参考：https://kns.cnki.net/KXReader/Detail?invoice=Xi2GWXPD7x9%2B7r3VqptQ4sx66hsJ%2BwIQCQzofM%2B%2Bc1cZLDaRZXtaqyNKKCzw%2B2HJqFFMXR57KUQkbswvyyr%2FygPPCEmm63qR45FjhwxRxnabVzc150m7QkD4wvMjO344rkRub5b2UtffVfThDe2vhwOr8Wc%2Befhjwdo1sMmSM0E%3D&DBCODE=CJFD&FileName=JSGG201707011&TABLEName=cjfdlast2017&nonce=0995427E48E349B8B4615CDB8B86FBEF&uid=&TIMESTAMP=1669471418460）文献[x]讨论了人工神经网络（ANN）在负载预测中的应用和训练，以及使用人工神经网络进行短期负荷预测的可能性。（参考：https://www.osti.gov/biblio/6015589）此类方法能够提取时序数据中的短期非线性特征，但无法捕获数据中的长期依赖关系，因此其在长期预测方面存在较大不足。

第三类基于深度学习的预测方法在时序数据预测方面取得了较好的发展。Fargana[x]等人利用LSTM的长期依赖挖掘能力，在此基础上使用Encoder-Decoder模型架构对时序数据进行特征提取和分解，提高了集群负载预测准确率；MinXian Xu[x]等人提出一种多特征负载预测模型，利用集群负载数据中不同特征之间的相互作用关系，使用多维特征进行负载预测。但是由于不同特征之间的相关性不均等且作用效果有限，不同目标特征的预测效果存在较大差异；为了解决集群长期负载存在的模式转换和振幅波动问题，更好挖掘负载数据的不同周期模式，Jiaming Huang[x]等人提出一种多尺度注意力机制，根据不同的周期模式设置不同的注意力权重，提高了集群负载的长期预测能力。但是，该模型在短期预测方面存在较大不足；为提高模型的短期预测能力，Wenyan Chen[x]等人提出将TCN时序神经网络用于集群负载预测，同时利用多维特征进行目标特征预测。该方法在短期预测方面具备较好表现，但是长期预测能力存在很大不足。

## 存在的挑战

互联网技术的快速普及使得互联网用户在过去十几年中实现了高速增长，随之而来的是海量用户请求。为保证用户服务质量，实现高可用、高可靠服务器集群系统，负载均衡与负载预测技术取得了长足发展。例如，文献[x]提出了一种基于注意机制的LSTM编码器-解码器机制，该方法在云计算等混合工作负载预测中实现了较好的性能表现；文献[x]提出一种基于负载反馈的分布式数字集群动态负载均衡算法,实现公网数字集群系统负载均衡,提高用户容量。该算法负载均衡度更小,用户请求响应延迟更低。然而，目前服务器集群负载均衡存在以下两方面的瓶颈：

集群负载呈现如下两个特点：1）短时间跨度内，负载变化呈现无周期性和波动性；2）长时间跨度内，负载变化呈现周期性特点，且不同时间跨度呈现不同的周期模式。因此，如何提高负载预测的准确度，同时兼顾短期预测和长期预测，是负载预测中需要解决的一个关键问题。

服务器集群将多台服务器节点连接到一起，在减轻单台服务器压力的同时为用户提供高质量的服务。单台服务器完成负载预测后，集群如何利用负载预测结果制定负载均衡策略，兼顾全局任务分配和局部负载调度，在实现为服务器节点合理分配任务的同时，避免集群负载资源浪费，保证集群负载均衡解决方案的有效性和系统性是另一个很关键的问题。

## 主要研究工作

根据以上挑战，本文研究了基于双时序流量预测的自响应动态负载均衡技术。主要包括基于用户请求流量和服务器负载流量的双时序流量预测方法、基于预测自响应的全局任务分配方法、基于集群服务器自索取的局部动态负载调度方法。并最终根据对上述技术的研究与分析，实现了一个基于双时序流量预测的自响应动态负载均衡集群系统。本文研究内容分为四个部分：

研究内容一，基于用户业务请求流量和服务器节点工作负载流量的双时序流量预测分别对用户请求流量和服务器负载流量建模时序预测模型，并分别预测出用户预测请求和服务器预测负载；

将用户预测请求和服务器预测负载作为研究内容二基于预测自响应的全局任务分配的输入，通过用户请求、服务器负载和服务器性能之间的作用和响应模型计算出服务器实时自响应负载，然后根据集群服务器自响应实时负载序列，通过加权最小负载分配策略为用户请求选择目标服务器[18]，从而确定用户请求分配方案；

在局部动态负载调度方面，研究内容三基于预测自响应的全局任务分配建立基于接受者主动的服务器自索取动态任务调度方案，协调局部相邻服务器节点之间的任务分配关系，平衡各服务器节点之间的负载。其中研究内容二基于预测自响应的全局任务分配和研究内容一输入相互协同关系，分别处理新用户请求全局分配和局部相邻服务器之间的任务调度重分配关系。

最后将三项研究内容进行整合，集成出一套基于双时序流量预测的自响应动态负载均衡集群系统。

## 论文组织结构

本文主要解决了服务器集群系统中基于负载预测实现高可用集群负载均衡和高效资源管理的问题，对基于用户请求流量和服务器负载流量的双时序流量预测技术、基于预测自响应的全局任务分配技术以及基于集群服务器自索取的局部动态负载调度技术进行了深入研究。论文总体组织结构如图1.1所示，总共包括六个章节，每个章节的具体内容如下：

第一章，绪论。首先阐述了本文的研究背景和意义，包括我国医疗卫生资源和人口总数之间的矛盾，传感技术和医疗信息化的发展趋势等方面。然后介绍了电子健康记录数据的应用、基于机器学习的疾病诊断算法研究和多任务学习应用的研究现状，总结分析现有算法的优势和不足，最后对本文的研究内容和组织架构进行概括说明。

第二章，基于双时序流量预测的自响应动态负载均衡集群系统的总体设计。本章首先分析了多源健康感知疾病诊断系统的功能需求和性能需求，然后根据多源健康感知数据结构形式和疾病诊断系统实现方式进行系统结构设计和流程设计，并详细讲解了系统实现所采用的关键算法。

第三章，基于用户请求流量和服务器负载流量的双时序流量预测模型。针对电子健康记录中健康感知数据来源多样、数据结构复杂、不同类型数据之间相关性难以挖掘的问题，本章提出了一种多源健康感知数据动静态关系融合模型，添加掩码向量作用于医学感测数据，在门控循环单元网络架构的基础上对动态医学感测数据、静态个人信息数据和疾病相关关系数据进行融合学习，挖掘多源异构数据的隐藏特征，实现多类别疾病诊断。对于疾病诊断问题中普遍存在的个人信息数据对医学感测数据的影响关系问题，本文采用数据动静态关系融合算法在提取异构数据的关系特征的同时保留了原始数据特征，帮助模型根据不同患者的体质状况对其所患疾病进行差异性诊断。对于多类别疾病诊断任务的疾病相关和疾病互斥问题，本文构建了关联矩阵来提取疾病间关系，将疾病关系数据与医学感测数据、个人信息数据特征进行融合，提高了模型的分类准确度。



图1.1 论文组织结构图

第四章，基于预测自响应和集群服务器自索取的全局任务分配与局部动态负载调度模型。针对疾病诊断任务间存在相似性和差异性的问题，本文提出了一种基于多尺度时间模式挖掘的多任务疾病诊断模型MTMS，该模型将疾病按照生理特征变化的不同划分为两类任务分别进行疾病诊断。模型采用了和第三章相同的数据预处理和动静态关系融合方法，构建了包含底层共享GRU网络和任务私有的基于多尺度扩张卷积的时间注意力网络进行多类别疾病诊断，充分挖掘同类型疾病的相似性关系和不同类型疾病的差异性特征。在任务的私有网络结构中，模型选取多个含有不同扩张因子的扩张卷积网络学习数据在不同时间间隔尺度上的关系，使用GRU网络学习生理特征时间依赖关系，通过注意力机制提取时间权重信息；模型的底层共享网络采用GRU进行时间相关性建模，挖掘历史数据中的趋势特征。在MIMIC-III真实数据集上的实验结果表明，基于多尺度时间模式挖掘的多任务疾病诊断模型可以对多类疾病进行准确诊断。

第五章，基于双时序流量预测的自响应动态负载均衡集群系统的实现。首先介绍了系统实现的软硬件开发环境，并在文中详细地介绍了多源健康感知疾病诊断系统所需的数据结构和功能模块，完成了多源健康感知疾病诊断系统各个模块的代码编写，最后从诊断结果准确性和诊断时间耗费方面对系统性能进行了相应的测试工作。

第六章，总结与展望。对本文已有研究工作进行总结归纳，分析了当前工作的不足之处，并对未来的研究方向进行展望。

# 多源健康感知疾病诊断系统设计

## 基于双时序流量预测的自响应动态负载均衡集群系统需求分析

在服务器集群负载均衡和资源管理过程中，由于用户服务的高可用和高可靠性以及集群系统的复杂性，如何统筹分析用户请求和集群负载信息，利用时序流量预测技术对用户请求和集群负载进行高效、准确的预测，保证集群系统全局任务分配和局部负载调度相统一，是检验集群系统高可用、高可靠性的重要评价标准。为保证集群系统实现负载均衡与资源管理，本文提出了双时序流量预测模型，对用户请求流量和集群负载流量进行流量预测，借助预测自响应的全局任务分配方法和集群服务器自索取的局部动态负载调度方法保证集群系统全局任务分配和局部负载调度相统一。本章首先分析了系统在功能和性能方面需要实现的目标，阐明本系统的实际意义及价值；然后对系统的物理结构和逻辑结构进行设计，介绍了各个功能模块的作用，规划了系统实现流程；最后对系统中所用到的流量预测算法以及任务分配与负载调度算法进行了阐述说明。

### 系统功能需求分析

本系统从实现集群负载均衡与资源管理的功能角度可以划分为两个大的模块，分别为基于用户请求流量和集群工作负载的时序流量预测模块以及基于预测的全局任务分配和局部动态负载调度模块。由于用户请求的高并发性与集群系统的复杂性，为实现时序流量预测，需要相应的数据采集、存储与预处理单元为流量预测提供的数据准备工作；与此同时，还需要模型设计与训练、全局任务分配、局部动态负载调度等功能单元。为保证集群系统实现负载均衡与资源管理，本文设计的基于双时序流量预测的自响应动态负载均衡集群系统将从以下几个角度进行功能需求分析。

1. 数据统计与存储

数据是时序流量预测模型的基础，因此，数据采集、统计与存储单元是必不可少的。

一个完整的服务器集群系统，其数据来源主要有用户请求数据和集群服务器节点的实时负载数据。从数据特征来看，用户请求数据和集群负载数据均为时序数据，是一种时间强相关的数据类型。

在数据统计与存储单元中，首先，系统需要统计用户请求和集群负载信息。一方面，系统需要实时记录与统计来自客户端的用户请求，准确记录用户请求对应的集群资源消耗量，例如CPU、内存、磁盘、网络IO等资源消耗情况；另一方面，系统需要实时记录和统计集群中各服务器节点的负载情况，同样包括服务器节点的CPU、内存、磁盘以及网络等资源信息。其次，系统需要对上述用户请求和集群负载等数据进行合理存储。另外，为保证数据的安全性，系统需要设置有效的数据备份机制。

1. 数据预处理

由于用户请求波动和集群中服务器节点故障，可能会存在部分数据丢失、记录时间不匹配、重复记录等问题，且进行数据统计与存储的服务器节点可能发生某些技术错误，因此需要对冗余数据进行去重、剔除错误数据，填充缺失数据，经过处理后的数据才能用于用户请求与集群负载预测。

1. 用户请求与集群负载预测

用户请求与集群负载预测是本系统的核心功能之一。

本系统需要充分挖掘用户业务请求流量和集群服务器工作负载流量的时序特性，并分别对两种时序流量数据建立有效的流量预测模型，以对用户请求和集群负载进行准确预测，为集群全局任务分配和局部动态负载均衡调度提供可靠依据。

1. 全局任务分配与局部动态负载调度

集群全局任务分配与局部动态负载调度是本系统的另一核心功能。

一方面，在双时序流量预测的基础上，挖掘实时用户业务请求、实时服务器工作负载与预测用户请求、预测服务器工作负载以及服务器性能之间的相互作用与响应关系，并建立合适的全局任务分配模型，实现基于预测自响应的全局任务分配。

另一方面，协调局部相邻服务器节点之间的任务分配关系，平衡各服务器节点之间的负载，减轻负载均衡器压力，降低集群通信开销，减少服务器集群整体资源消耗，实现基于服务器自索取的局部动态负载调度。

### 系统性能需求分析

为了保证负载均衡和资源管理系统的实时性、准确性与低消耗性，本文将从以下四个方面进行系统性能分析：

1. 实时性

解决传统软件方法的负载均衡算法无法实时获取集群服务器工作负载导致负载均衡滞后效果明显，但频繁对服务器进行负载采样以获取实时负载会导致增加服务器压力这一矛盾。本文充分利用用户请求与集群负载预测得到的用户请求与集群负载信息，制定合理的全局任务分配和局部动态负载调度策略，以提高集群负载均衡调度的实时性。

1. 准确性

解决基于流量预测方法的负载均衡算法只针对用户求情或服务器负载中的某一流量做流量预测且未考虑用户请求和服务器负载之间的相互作用这一缺陷。本文在双时序流量预测的基础上，挖掘实时用户业务请求、实时服务器工作负载与预测用户请求、预测服务器工作负载以及服务器性能之间的相互作用与响应关系，并建立合适的全局任务分配模型，实现基于预测自响应的全局任务分配，以提高集群负载均衡调度和资源管理的准确性。

1. 低消耗性

解决传统方法主要依赖负载均衡调度器实现集群服务器进行任务调度，导致用户请求响应慢、服务器通信开销大这一弊端，提高集群的用户响应速度，降低服务器集群的资源消耗。

1. 高可用性

高可用性是衡量集群系统实用价值的一个关键指标，良好的稳定性与高可用性可以为用户提供可靠的使用体验，因此，提高基于双时序流量预测的自响应动态负载均衡集群系统的稳定性和高可用性是系统设计中需要考虑的重要一环。当发生用户请求流量波动、集群节点故障等异常情况时，集群系统仍然能够为用户提供连续、高质量服务，保证集群业务处理稳定，集群运行正常。

## 基于双时序流量预测的自响应动态负载均衡集群系统总体架构设计

### 基于双时序流量预测的自响应动态负载均衡集群系统架构设计

图x.x展示了基于双时序流量预测的自响应动态负载均衡集群系统架构，一共分为数据层、数据处理层、时序流量预测层、全局任务分配和局部动态负载调度（待优化）层以及显示控制层等五个部分。

数据层主要负责集群系统的数据和资源管理工作，为系统提供数据和资源支持。该层由用户请求数据库、集群负载数据库、可用节点信息库、节点资源信息库以及系统配置信息库组成。用户请求数据库包含历史时间内访问集群系统的用户请求时序数据集，包括数据记录时间、任务编号、服务器节点编号、资源消耗率、响应时间等信息；集群负载数据库包含历史时间内各个服务器节点的负载时序数据集，包括服务器节点编号、数据记录时间、资源消耗率等信息；可用节点信息库存储集群当前可用服务器节点的各项基本信息，包括节点编号、负载评分等信息；节点资源信息库存储集群中各服务器节点的资源配置信息，包括节点编号、资源利用率等信息；系统配置信息库存储集群系统的基础配置信息，包括集群架构、节点数量、网络拓扑、通信速率等信息。

数据处理层主要负责对数据层接收到的数据的处理工作，将其转换成时序流量预测层可直接使用的数据。该层包括用户请求和集群负载数据的预处理模块、数据切分模块和多变量联合特征选择模块。其中预处理模块将收集到的数据进行归纳整理，包括数据去重、异常值检查与删除、数据填充操作；数据切分模块对时序流量数据在时间维度上进行切分，将时序数据切分成一段历史训练窗口和未来的预测窗口，对于预测窗口中的每一条样本，基于训练窗口中的历史信息构建特征，转化为一个监督学习预测问题进行求解；多变量联合特征选择模块通过计算不同资源变量特征之间的相关性，为目的变量特征选择多个相关变量特征，将单一变量时序预测问题转化为多变量时序预测问题。

时序流量预测层是基于双时序流量预测的自响应动态负载均衡集群系统的核心层之一，负责用户请求和集群负载的预测任务。该层可分为两个子层，第一子层为用户请求预测层，该子层负责根据历史用户请求时序流量信息对未来一段时间内的用户请求流量进行预测，包含用户请求流量聚类模块和请求预测模块；（补充用户请求预测子层中各模块的作用）第二子层为集群负载预测层，该子层负责对服务器节点未来一段时间内的负载信息进行预测，包含基于一维全卷积的短时特征提取模块、基于LSTM的长时特征提取模块、加权长短时特征融合模块和解码模块。分为数据关系提取模块和疾病诊断模块两部分。其中，短时特征提取模块中在对负载数据进行LSTM编码之前，先使用一维全卷积神经网络对原时序负载数据进行一维全卷积操作，得到短时特征向量；长时特征提取模块将经一维全卷积短时特征提取后的短时特征数据输入LSTM进行长时特征提取，得到负载时序序列的长时特征向量；加权长短时特征融合模块先将经一维卷积后的短时特征向量与经LSTM长时特征提取模块得到的长时特征向量进行拼接融合，得到长短时融合特征向量，然后借助注意力机制对每个预测步的长短时融合特征向量进行注意力加权处理，得到每个预测步的加权长短时融合特征向量；解码模块依次读取加权长短时融合特征、更新其神经元状态和隐藏状态，输出当前时刻的预测值。

全局任务分配和局部动态负载调度（待优化）层是集群系统的另一核心层，负责制定集群任务分配方案和协调局部相邻服务器节点之间的任务分配关系。该层同样可分为两个子层，第一子层为全局任务分配层，该子层负责建立合适的全局任务分配模型，为集群制定任务分配方案；第二子层为局部动态负载调度层，该子层基于服务器自索取机制，实现集群局部动态负载调度。全局任务分配层子层包括自响应实时负载计算模块、加权最小负载全局任务分配模块和目标节点选择模块。。。。（待完善）局部动态负载调度子层包括服务器负载上下限比较模块、集群超载任务管理模块和集群转移任务管理模块。。。。（待完善）

显示控制层用于实现集群的显示功能，包含结果用户请求列表模块、服务器节点列表模块以及用户请求分配和局部负载调度模块。用户请求列表模块用于显示集群系统当前等待处理和分配的用户请求列表；服务器节点列表模块用于显示集群后端当前服务器节点的负载状态；用户请求分配和局部负载调度模块负责为用户请求分配后端服务器节点，以及后端服务器节点之间的局部负载调度。

### 基于双时序流量预测的自响应动态负载均衡集群系统逻辑结构设计

图x.x展示了基于双时序流量预测的自响应动态负载均衡集群系统的逻辑结构，核心功能模块为时序流量预测模块和任务分配与负载调度模块（这个名字需要优化，与上面保持一致），其中时序流量预测模块又分为用户请求预测模块和集群负载预测模块；任务分配与负载调度模块又分为全局任务分配模块和局部负载调度模块。除此之外，还有数据库、数据输入处理模块以及结果显示模块。

时序流量预测模块是基于双时序流量预测的自响应动态负载均衡集群系统的核心功能模块之一，具体又可划分为两个子模块，分别为用户请求预测模块和集群负载预测模块。用户请求预测模块主要包含聚类和预测两部分（待完善）。集群负载预测模块负责对服务器节点未来一段时间内的负载信息进行预测，包含联合特征选择、长短时特征提取、负载预测等部分。在完成数据切分与联合特征选择之后，短时特征提取模块中在对负载数据进行LSTM编码之前，先使用一维全卷积神经网络对原时序负载数据进行一维全卷积操作，得到短时特征向量；然后，长时特征提取模块借助LSTM网络对得到的短时特征向量进行长时特征提取，得到长时特征向量；然后加权长短时特征融合模块结合注意力机制，将短时特征向量与长时特征向量进行拼接融合并做注意力加权处理加权长短时融合特征向量；最后，解码模块依次读取加权长短时融合特征、更新其神经元状态和隐藏状态，得到未来时刻的预测值。

基于双时序流量预测的自响应动态负载均衡集群系统的另一核心功能模块是任务分配与负载调度模块，该功能模块包含全局任务分配和局部动态负载调度两个子模块，分别负责制定集群任务分配方案和协调局部相邻服务器节点之间的任务分配关系。全局任务分配子模块包括自响应实时负载计算模块、加权最小负载全局任务分配模块和目标节点选择模块。。。。（待完善）局部动态负载调度子模块包括服务器负载上下限比较模块、集群超载任务管理模块和集群转移任务管理模块。。。。（待完善）

数据输入处理模块和结果显示模块共同组成了显示控制模块，主要负责对用户请求的接收、时序流量预测和任务分配结果的展示。数据输入处理模块需要实现用户和集群系统的交互，并实时记录并存储用户请求的资源信息。用户请求和集群负载预测完成后，时序流量预测结果显示模块显示未来一段时间内用户请求和集群负载的资源消耗情况，供用户查看中间预测结果；全局任务分配和局部负载调度完成后，结果显示模块会显示集群系统为特定用户请求制定的任务分配方案，以及局部负载调度的结果信息。

该集群系统的用户请求预测模块和集群负载预测模块中的模型训练和流量预测均为离线部分，模型训练会随着历史数据的累积周期性更新，时序流量预测可以根据用户请求的频次进行多次预测，并通过显示控制层对将评估与诊断结果进行可视化展示。

2.3 基于双时序流量预测的自响应动态负载均衡集群系统总体流程设计

本文设计的基于双时序流量预测的自响应动态负载均衡集群系统的流程分为数据预处理阶段、双时序流量预测阶段和任务分配与动态负载调度阶段。数据预处理阶段将采集到的数据进行缺失值处理、归一化处理、数据切分、多变量联合特征处理等操作。时序流量预测阶段进行用户请求和集群负载两种时序流量的预测任务，分析用户请求和集群负载两种时序数据的特征，预测未来一段时间内两种流量的变化趋势，为任务分配与负载调度阶段提供方案制定依据。任务分配与动态负载调度阶段为来自客户端的用户请求制定任务分配方案，同时负责协调局部相邻服务器节点之间的任务分配关系，平衡各服务器节点之间的负载，减少服务器集群整体资源消耗。帮助用户进行疾病诊断与健康评估。基于双时序流量预测的自响应动态负载均衡集群系统的流程如图x.x所示，主要分为数据预处理、双时序流量预测和多任务分配与动态负载调度三个部分，具体流程如下。

（1）数据预处理阶段

数据预处理阶段首先将集群系统中采集和存储的数据进行初步处理，去除重复记录和记录不全的数据；其次进行数据清洗工作，对数据进行缺失值和异常值检查，使用均值填充法对缺失数据进行补全，删除因集群故障导致的异常值；然后对时序流量数据在时间维度上进行切分，将时序数据切分成一段历史训练窗口和未来的预测窗口；最后进行多变量联合特征处理，通过计算不同资源变量特征之间的相关性，为目的变量特征选择多个相关变量特征，将单一变量时序预测问题转化为多变量时序预测问题。

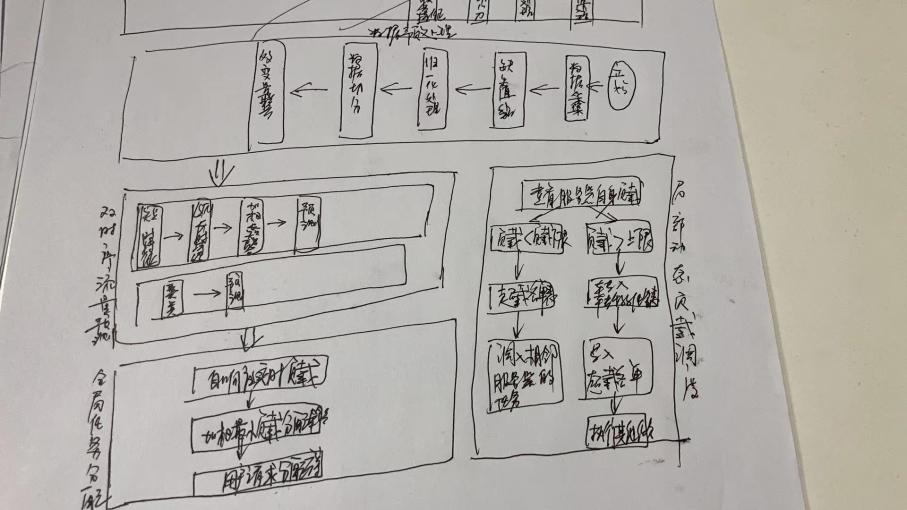


图2.3 多源健康感知疾病诊断系统流程图

（2）双时序流量预测阶段

该阶段分为用户请求流量预测和集群负载流量预测两个任务，两个任务是同时进行的。

用户请求流量任务负责根据历史用户请求时序流量信息对未来一段时间内的用户请求流量进行预测，包含用户请求流量聚类和请求预测两个阶段。集群负载流量预测阶段负责对服务器节点未来一段时间内的负载信息进行预测，分为基于一维全卷积的短时特征提取、基于LSTM的长时特征提取、加权长短时特征融合和解码预测等阶段。

1. 任务分配与动态负载调度阶段

该阶段分为全局任务分配和局部动态负载调度两个任务，其中任务分配任务为触发性任务，其基于用户请求，有新用户请求时运行该任务；动态负载调度为实时任务，其在集群系统的整个生命周期运行。

全局任务分配任务包括自响应实时负载计算模块、加权最小负载全局任务分配模块和目标节点选择模块。。。（待完善）局部动态负载调度任务包括服务器负载上下限比较模块、集群超载任务管理模块和集群转移任务管理模块。。。。（待完善）

（4）结果显示

双时序流量预测阶段完成以后用户请求和集群负载的预测结果通过结果显示模块进行结果的显示；全局任务分配与局部动态负载调度的用户请求分配方案与局部动态负载调度的实时结果也会在结果显示模块进行显示。

2.4 基于双时序流量预测的自响应动态负载均衡集群系统关键算法

本文提出的基于双时序流量预测的自响应动态负载均衡集群系统基于时序流量预测技术，从全局任务分配和局部动态负载调度两个方面着手，实现集群系统的负载均衡和高效资源管理。

在双时序流量预测方面，本文分别对用户请求和集群负载两种时序流量建立时序预测模型。针对用户请求流量，本文提出了一个。。。方法（用户请求流量预测的具体方法名，待完善）针对集群负载流量，本文提出了一个基于长短时特征融合的集群负载流量预测模型。该模型通过多变量联合特征机制，即联合多个资源变量的特征对某一目标资源变量进行预测，充分利用不同资源变量的时序特征以及它们之间的作用关系，提高了目标资源变量的预测准确度。针对复杂集群系统中短期负载预测方面存在的不足，同时增强模型长期预测能力，该模型借助注意力加权长短时特征融合方法，以兼顾长期预测和短期预测。具体而言，为解决现有方法在负载短期预测方面存在的不足，模型借助一维全卷积神经网络（FCN），在进行长期特征提取之前，对数据进行一维全卷积（1D FCN）处理，强化时序负载的短期依赖关系，得到短时特征向量。为增强模型长期预测能力，使用LSTM提取时序负载的长时特征，然后将短时特征与长时特征进行拼接融合，得到长短时融合特征向量；然后利用Attention注意力机制，分别对每一时刻的长短时融合特征向量进行加权，得到对应时刻的加权长时融合特征向量。

在全局任务分配和局部动态负载调度方面，本文分别对全局任务分配和局部动态负载调度两类工作建立不同的处理机制。针对全局任务分配，本文提出了基于预测自响应的全局任务分配模型（待完善方法细节）。针对局部动态负载调度，本文提出了基于集群服务器自索取的局部动态负载调度（待完善方法细节）

2.5 本章小结

本章对基于双时序流量预测的自响应动态负载均衡集群系统进行了总体设计。首先，根据系统应用场景，对集群系统的功能需求和性能需求进行说明；然后，从系统结构设计和逻辑结构设计两个方面来介绍系统总体架构设计，详细解释了每个核心模块的功能；之后，展示了集群系统实现的总体流程，对流程中的每一步都进行了详细说明；最后，介绍了系统中用于双时序流量预测和任务管理的关键算法。

# 基于加权长短时特征融合的~~用户请求流量和集群负载流量的~~双时序流量预测模型

3.1 双时序流量预测问题描述（小论文引言|加热力图或其他图|加用户请求预测部分）

随着互联网终端的快速、大量普及，为满足日益普遍的高并发应用场景，我们对云计算、电网等服务器集群技术提出了更高的要求。针对集群系统常用的适用场景，例如B/S、C/S架构场景，提高用户请求和资源分配准确度，保证任务分配的实时性，降低集群整体资源消耗，是实现高效集群系统的重要参考指标。实现高效的集群负载均衡和资源分配，可以从客户端用户请求和服务端集群负载两个方面着手。

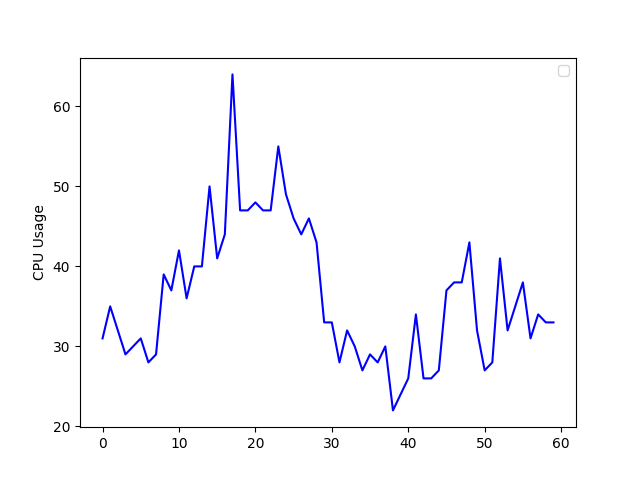
就客户端而言，在服务器集群、云计算、电网等很多应用场景中，我们借助时序流量预测技术，对用户请求流量或网络流量进行预测，并以此为依据来调整和管理集群资源，以辅助运营商进行精细化运营，提高用户请求和资源分配准确度，降低系统资源消耗。

如图x所示，在多数服务器集群的应用场景中，用户请求时序流量或网络时序流量呈现两个显著特点：1）存在大量的用户请求时序流量，某些会呈现一定的模式，但某些可能不会呈现周期性或表现出一定的趋势；2）很多用户请求的持续的时间较短，积累的历史数据很少。

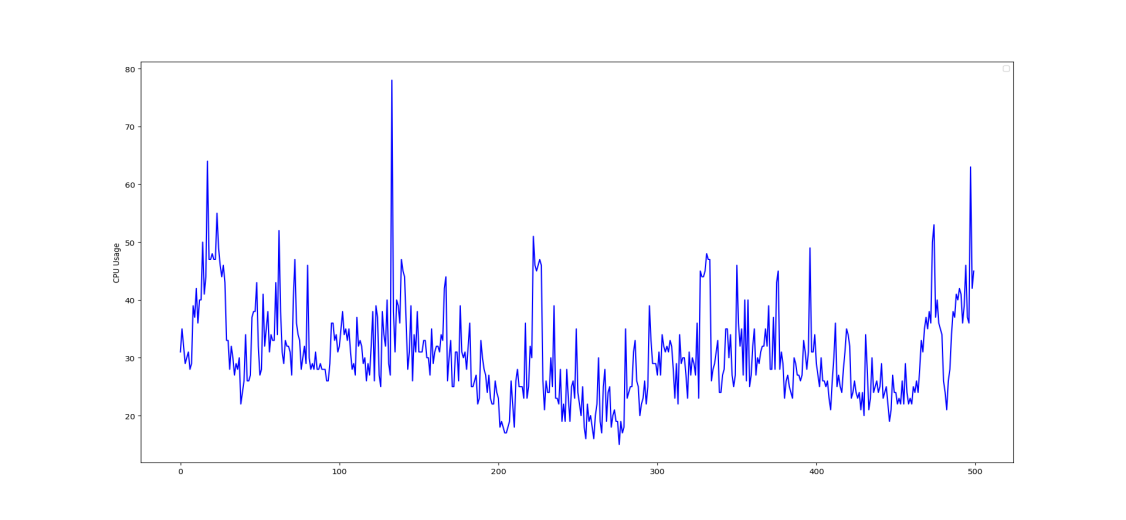
（这里需要加个图）

就服务端而言，在被动响应式集群系统中，资源管理是纯反映式的，系统根据负载变化对集群进行资源分配和调整。配置决策和资源调整的滞后性，容易导致资源配置不足或过度配置问题。为提高系统资源配置的主动性，实现自适应资源分配和管理，我们需要对集群负载进行预测。通过提前预测集群未来一段时间内的负载变化，预先制定资源分配和调整方案，降低配置决策和资源调整滞后性的影响，同时提高集群系统的动态性，进而提高系统的资源利用率。

如图1所示，通过对现有公开集群负载数据集的分析，我们发现，集群负载时序流量呈现如下两个特点：1）短时间跨度内，负载变化呈现无周期性和波动性；2）长时间跨度内，负载变化呈现周期性特点，且不同时间跨度呈现不同的周期模式。因此，如何提高负载预测的准确度，同时兼顾短期预测和长期预测，是负载预测中需要解决的关键问题。



（a）短时间跨度内集群负载变化



（b）长时间跨度内集群负载变化

图1 不同时间跨度内集群负载变化

强调一下，两种时序流量存在一个共性：短时和长时周期呈现不同的变化特征。

。。。

存在的问题：

目前，对于时序数据的预测主要有三类方法，分别为基于传统线性回归模型的负载预测方法，基于传统机器学习的负载预测方法和基于深度学习的负载预测方法。第一类基于传统线性回归模型的负载预测方法主要有自回归[x](Autoregressive,AR)、滑动平均(Moving Average,MA)、自回归移动平均[x](Autoregressive Moving Average, ARMA)以及差分整合移动平均自回归[x](Autoregressive Integrated Moving Average, ARIMA)等模型。这些模型在复杂度低、线性关系较强的数据中可以实现较好的预测效果，因此此类方法存在对数据的限制较多，且需要人工调整模型参数等方面的不足。第二类基于传统机器学习的负载预测方法主要有马尔科夫模型[x]、贝叶斯模型[x]、支持向量回归[x](Support Vector Regression,SVR)模型，以及决策树和传统人工神经网络[x](Artificial Neural Networks, ANN)等模型。此类方法能够提取时序数据中的短期非线性特征，但无法捕获数据中的长期依赖关系，因此其在长期预测方面存在较大不足。近年来，随着深度学习在非线性特征处理方面表现出的显著优势，基于深度学习的预测方法在时序数据预测方面取得了较好的发展。Fargana[x]等人利用LSTM的长期依赖挖掘能力，在此基础上使用Encoder-Decoder模型架构对时序数据进行特征提取和分解，提高了集群负载预测准确率；MinXian Xu[x]等人提出一种多特征负载预测模型，利用集群负载数据中不同特征之间的相互作用关系，使用多维特征进行负载预测。但是由于不同特征之间的相关性不均等且作用效果有限，不同目标特征的预测效果存在较大差异；为了解决集群长期负载存在的模式转换和振幅波动问题，更好挖掘负载数据的不同周期模式，Jiaming Huang[x]等人提出一种多尺度注意力机制，根据不同的周期模式设置不同的注意力权重，提高了集群负载的长期预测能力。但是，该模型在短期预测方面存在较大不足；为提高模型的短期预测能力，Wenyan Chen[x]等人提出将TCN时序神经网络用于集群负载预测，同时利用多维特征进行目标特征预测。该方法在短期预测方面具备较好表现，但是长期预测能力存在很大不足。

现有集群负载预测模型多为基于深度学习的预测模型，为提高模型的预测准确度，现有模型充分利用多特征、注意力机制等方法，并且在短期预测或长期预测的某一方面取得了较好效果。显然，现有模型未能兼顾负载短期预测和长期预测效果，使模型同时具备较好的短期和长期预测能力。~~因此，本文提出利用多变量特征，基于Encoder-Decoder特征提取架构，同时结合注意力机制，以充分挖掘时序数据的长期特征，提高模型长期预测能力。同时，为保证模型的短期预测能力，本文提出在进行特征提取之前，添加一层一维全卷积神经网络（FCN），以更好地挖掘和保留初始时序数据的短期特征，进而提高模型的短期预测能力。~~

本章/文的研究工作：

因此，为解决上述问题，兼顾用户请求流量预测与集群负载预测，提高用户请求和集群负载的预测准确度，本章提出了基于加权长短时特征融合的双时序流量预测模型。具体而言，该模型可分为两大部分，第一部分负责对用户请求流量和集群负载流量进行时序特征提取前的预处理工作，第二部分负责对两类时序数据进行加权长短时特征融合处理。其中，第一部分别对两类时序数据进行聚簇分类（点明，具体化这个聚类方法）和多变量联合特征选择处理；第二部分分别对两类时序数据进行长短时特征提取和加权融合。首先通过一维全卷积短时特征提取模块，利用一维全卷积神经网络（1D FCN）对输入时序负载进行短时特征提取，得到短时特征向量；接着将其输入LSTM长时特征提取模块，进行长时特征提取得到长时特征向量；然后借助注意力加权长短时特征融合模块，将短时特征向量与长时特征向量进行拼接融合，得到长短时融合特征向量；再利用注意力机制，分别对每一时刻的长短时融合特征向量进行加权，得到对应时刻的加权长时融合特征向量；最后利用LSTM解码模块，得到负载预测结果。

本章的主要研究如下：

1）针对集群中用户请求流量呈现出的变化特点，本章提出一种基于K-中心点聚类的DTW动态时间规整算法。。。（待完善）

1）针对集群时序负载呈现出的变化特点，本章提出一种基于多变量联合特征的长短时时序负载预测模型，通过多变量联合特征预测和注意力加权长短时特征融合方法，提高模型的负载预测能力，同时兼顾模型的长时预测和短时预测准确度。

~~2）基于集群中不同负载资源之间存在相互影响和作用关系这一现状，本文提出多变量联合特征预测，即联合多个资源变量的特征对某一目标资源变量进行预测。该方法可以充分利用不同资源变量的时序特征以及它们之间的作用关系，以提高目标资源变量的预测准确度。~~

3）针对集群场景下短期时序预测方面存在的不足，同时增强模型长期预测能力，本章提出注意力加权长短时特征融合方法，以兼顾长期预测和短期预测。针对现有方法在负载短期预测方面存在的不足，本文借助一维全卷积神经网络（FCN），在进行长期特征提取之前，对数据进行一维全卷积（1D FCN）处理，强化时序负载的短期依赖关系，得到短时特征向量。为增强模型长期预测能力，使用LSTM提取时序负载的长时特征，然后将短时特征与长时特征进行拼接融合，得到长短时融合特征向量；然后利用Attention注意力机制，分别对每一时刻的长短时融合特征向量进行加权，得到对应时刻的加权长时融合特征向量。

3.2 基于加权长短时特征融合的双时序流量预测模型框架

3.2.1 双时序数据预处理（待优化）

（1）基于聚类的用户请求流量类别划分

（2）集群负载流量多变量联合特征选择

3.2.2 短时特征提取

3.2.3 长时特征提取

3.2.4 加权长短时特征融合

3.2.5 解码与预测

3.3 基于加权长短时特征融合的双时序流量预测模型实现

3.4 算法描述

3.5 实验与分析

（1）数据集

（2）实验设置

（3）对比算法与评价指标

（4）实验结果对比

（5）消融实验

（6）参数设置

（7）case study

3.6 本章小结