**2023“CCF-蚂蚁绿色计算专项科研基金”**

**项 目 名 称**

**（根据实际课题填写）**

**申报课题：面向云资源动态伸缩调度的时序基础模型研究**

|  |  |
| --- | --- |
| **申请者姓名** | 张永敏 |
| **CCF会员码** | F5385M |
| **所属机构** | 中南大学 |
| **提交日期** | 2023年11月26日 |

**目录**

**1. 项目介绍**

1.1. 基金题目

1.2. 背景及研究意义

1.3. 研究目标

1.4. 研究方法

**2. 研究计划**

2.1. 项目期限（一般是一年期）

2.2. 项目计划（一般分为三或四个阶段）

**3. 预期成果**

**4. 人力、设备等投入及项目预算**

4.1. 人力投入

4.2. 设备投入

4.3. 项目总预算

**5. 教授个人简历**

**1. 项目介绍**

* 1. 题目

**面向云资源动态伸缩调度的时序基础模型研究**

* 1. 研究背景及研究意义

在风险管理和决策的关键任务中，时间序列数据的预测在许多领域中发挥了重要作用，包括统计学、机器学习、数据挖掘、计量经济学、运筹学等等，时间序列预测为决策者提供了未来趋势和模式的信息，帮助他们做出更准确、有效的决策。例如，预测产品的供需可用于优化库存管理、车辆调度和拓扑规划，这对于供应链优化的大多数方面都至关重要。

传统的时间序列预测模型包括ARIMA模型、指数平滑方法和状态空间模型（SSM）等，其中ARIMA模型、指数平滑方法都可以表示成状态空间模型的形式，SSM提供了时间序列建模的通用框架，由状态方程和观察值方程组成。

在现代预测应用中，传统的SSM模型无法从相似的时间序列数据集合中推断出共享模式，这就产生了繁重的计算任务和大量人力劳动需求。因此，深度神经网络（Deep Neural Networks，DNN）凭借其提取高阶特质的能力走进了人们的视野。通过深度神经网络，可以识别时间序列内部和时间序列之间的复杂模式，并且能够从原始时间序列的数据集合中进行识别，所需的人力要少得多。然而，由于这些模型所作的结构假设较少，它们通常需要更大的训练数据集来学习得到准确的模型。

为了弥补这两种方法各自的不足，将传统统计模型与深度学习融合，产生了一些新的预测方法，例如使用循环神经网络（Recurrent Neural Networks，RNN）在SSMs和变分自编码器（Variational Auto Encoding，VAE）之间建立联系。在预测方面，使用RNN在每个时间步上生成线性高斯状态空间模型（LGSSM） 的参数，提出了深度状态空间模型（Deep State Space Model，DSSM）。对于非线性SSM，Eleftheriadis 等［1］提出非线性高斯过程状态空间模型（Gaussian Process State Space Model，GPSSM）。Salinas 等研究了多元时间序列预测问题，提出深度自回归模型（DeepAR）[2]是建立在对时间序列数据进行深度学习的基础上，为概率预测问题设计了一个类似的基于LSTM（Long Short-Term Memory）的自回归RNN架构。而 Vaswani 等提出的Transformer[3]利用Attention Mechanism来处理数据。与基于RNN的方法不同，Transformer 允许模型访问历史的任何部分，而不考虑距离，这使得它更适合于捕捉具有长期依赖性的循环结构。

徐超等提出一种集成自回归综合移动平均（ARIMA）[4]模型与自适应过滤法的组合预测模型。该组合强调ARIMA模型对时间序列数据特征识别与参数估计的优势，同时引入自适应过滤法的“权数”调整思想，对ARIMA模型的参数进行调整，以减少预测误差，提高预测精度。沈旭东对近年来基于深度学习的时间序列分析方法进行讨论，从应用、网络架构、思想等方面总结了最新的时间序列预测、分类、异常检测等任务的深度学习方法，为了解时间序列深度学习解决方案的技术和发展趋势提供了参考。吴双双利用卷积神经网络、循环神经网络、双通道神经网络对数据进行了预测，并取得了不错的预测效果。权钲杰利用长短期记忆网络和卷积神经网络对数据进行预测，并针对深度神经网络模型训练不稳定的问题，研究了将集成学习方法应用于对深度神经网络预测模型的改进，提出了基于噪声扰动集成方法的深度神经网络集成模型。刘峰等提出了一种组合聚类分析和神经网络的预测方法。王慧健等提出一种新的时间序列短期趋势预测方法，通过对时序数据进行离散化，用字符表示各个时间段数据的范围，并利用神经网络语言模型预测得到下一个字符。李洁等基于真实的民航旅客历史出行记录，根据其时序数据的特征建立基于后向传播算法的循环神经网络（RNN）预测模型，对未来时段的日客流量进行预测。在此基础上考虑到时序数据在不同时间尺度呈现不同的变化规律，建立多时间尺度的预测模型对旅客出行的周期性和趋势性进行建模，提升预测精度。

随着以5G，人工智能为代表的新一代信息技术快速发展，推动了我国信息技术与传统产业深度融合和数字经济蓬勃发展的同时，也导致作为信息系统和算力载体的数据中心规模和能耗大幅增长。因此，如何打造绿色算力以优化算力的平均能耗，成为推动我国实现低碳化，绿色发展和助力我国碳达峰、碳中和战略的重大任务。然而，以云为代表的数据中心，尽管依托资源弹性伸缩优势极大提高了IT基础设施的灵活性、可伸缩性和成本效益，但因计算任务海量实时动态和资源需求不确定性等特征，难以合理调度和分配资源，从而导致云中心面临着能耗高，利用率低，性能保障难等困境。据《2022年版中国数据中心市场深度研究与投资可行性报告》和《云计算白皮书（2023年）》统计，2021年国内云中心的总耗电量高达2166亿千瓦，占据社会用电量的2.6%，但资源利用效率仅为68%。因此，**如何依托云计算资源弹性伸缩优势，优化计算负载预测的同时实现资源高效动态伸缩调度成为推动云中心绿色发展的迫切任务，也是助力国家双碳战略和绿色发展的关键组成。**

现有面向云资源动态伸缩调度的研究主要从传统理论和人工智能技术进行资源优化。**基于传统理论设计的云资源动态伸缩调度研究主要采用阈值法，排队论，控制理论等优化工具开展机制设计。**面向阈值法的相关研究主要是通过设置资源阈值，依托预先设定的资源伸缩规则实现调度[5]。但阈值的合理性在该类研究中通常缺乏有效性证明，这可能导致系统资源伸缩调度频率高、稳定性差、抗压能力低。而基于排队论的相关研究依托假定的计算负载分布特性构建排队模型来预测平均响应时间和计算负载，从而实现服务器间的合理分配[6-8]。然而计算负载的分布特性不一定能精准匹配实时动态特性，从而导致系统性能鲁棒性低。面向控制理论的相关工作则利用PIDPI、I等固定增益控制器或者MIMO自适应控制器调整虚拟机数量、CPU和磁盘IO的使用率[9]。但是控制理论通常依赖于对系统的精确建模，导致外部环境的变化、网络延迟、硬件故障等系统因素难以确定，从而导致控制器的性能下降。 **基于人工智能技术设计的云资源动态伸缩调度研究主要从强化学习和时序模型两个角度开展机制设计。**面向强化学习的相关研究依托SARSA、 Q-learning和 ANN等强化学习算法实现虚拟机性能按需自动调节参数[10]，但是强化学习模型通常需要较长的学习时间且可泛化能力低，从而导致面向动态更新的负载情况时效率低。而面向时序模型的相关工作通常根据时间变化特征，构建ARMA、ARIMA、SARIMA等模型，结合其他设计方法精准预测计算负载，从而对云资源实现更灵活、精确的自动缩放[11]。相比于其他设计方法，基于时序模型的相关研究更适用于实时动态的计算环境，但受计算负载预测精度的影响，系统最终的优化性能难以保障。

此外，任务的多源异构特性使其对CPU、内存等资源需求具有差异化的敏感性，**[这里需要补充点文献说明实时]**。这进一步放大了现有面向云资源动态伸缩调度的研究在多源异构，实时动态计算环境下的局限性。因此，**本项目拟面向多源异构任务探究高精度、强泛化的时序基础模型，实现对跨领域应用计算负载精准预测，并构建异构任务对差异化资源需求的敏感性模型，以进一步设计面向多源异构任务的多维度资源调度优化机制，实现云中心的局部资源弹性伸缩和全局资源的规划调度，进而最大化云资源利用效率的同时最小化能耗，最终推动绿色计算领域高速发展的同时助力我国碳中和战略实施。**

**参考文献**

1. Berntorp K. Online Bayesian inference and learning of Gaussian-process state–space models[J]. Automatica, 2021, 129: 109613.
2. Schaduangrat N, Anuwongcharoen N, Charoenkwan P, et al. DeepAR: a novel deep learning-based hybrid framework for the interpretable prediction of androgen receptor antagonists[J]. Journal of Cheminformatics, 2023, 15(1): 50.
3. Tang B, Matteson D S. Probabilistic transformer for time series analysis[J]. Advances in Neural Information Processing Systems, 2021, 34: 23592-23608.
4. Kontopoulou V I, Panagopoulos A D, Kakkos I, et al. A review of ARIMA vs. machine learning approaches for time series forecasting in data driven networks[J]. Future Internet, 2023, 15(8): 255.

[5]徐胜超, 陈刚. 基于利用率的容器云弹性资源调度算法[J]. 计算机与数字工程, 2022, 50(9): 1986-1990.

[6]张珂, 张利国. 基于移动边缘计算的 GI/GI/1 排队建模与调度算法[J]. 自动化学报, 2022, 48(7): 1737-1746.

[7]李晖, 严文, 刘文亮, 等. 云计算数据中心的虚拟资源自适应供给模型[J]. Journal of Fuzhou University, 2023, 51(3).

[8]马冯超, 陈思溢, 刘锦. 基于云联盟协同机制的利益优化方法[J]. 计算机集成制造系统, 2023, 29(7): 2385.

[9]刘瀚琛, 王冲, 鞠平, 等. 计及统一潮流控制器的电力系统双层协调弹性调度[J]. Electric Power Automation Equipment/Dianli Zidonghua Shebei, 2023, 43(4).

[10]杨君, 朱颖雯. 面向边云协同计算的能耗感知资源调度方法[J]. 应用科学学报, 2023, 41(3): 369-377.

[11]沐磊, 李洪赭, 李赛飞. 一种改进的 Kubernetes 弹性伸缩策略[J]. 计算机与数字工程, 2022, 50(2): 327-331,372.

* 1. 研究目标

1. 提出一种基于多源异构任务的高效时序基础模型，形成一套面向跨场景应用的异构任务负载精准预测体系；
2. 设计面向多源异构任务的多维度资源调度优化机制，并将其应用于计算、存储等领域，将资源利用率提升5%以上。
   1. 研究方法

【说明研究项目拟采用的技术路线与研究方法，突出其创新性和可落地性等优势，不用写那么多公式】

基于本项目的研究目标，拟采用的研究方法技术路线图如图1所示。核心机理是针对多源异构实时动态的任务需求，首先探索基于多源异构任务的高效时序基础模型，实现面向跨领域应用的计算负载精准预测；然后，结合计算任务与资源需求相关关系，从负载预测和调度优化两个层面出发，设计面向多源异构任务的多维度资源调度优化机制，以形成面向云资源的局部弹性伸缩和全局规划调度的优化方法。

图1.研究方法技术路线图**（缺图）**

1. **基于多源异构任务的高效时序基础模型**

针对现有RNN、Transformer等时序基础模型单个时间尺度信息提取能力有限，难以同时捕获趋势、周期和突变等多级模式，并且计算复杂度和内存开销以及训练时长较高等问题，本项目拟提出一种基于自注意力机制（Self-Attention Mechanism）的时间序列预测模型，通过捕捉序列中不同位置之间的依赖关系，和评估不同位置对于序列中其他位置的重要性，优化计算效率的同时解决传统模型的输入时间序列长度受限、内存占用率高等问题。

具体而言，**假定初始时序数据为（），则序列中不同位置间的依赖关系可表示为（）。**

**设计稀疏注意力机制模块**，结合依赖关系，以节点稀疏性得分为衡量不同时序位置对于序列中其他位置的重要性，从而减少非必要的点积操作，进而降低模型计算复杂度。**若。。。。则节点稀疏性得分为。。。**

**知识Distilling（蒸馏）**，通过对主要特征时间步query的优势特征赋予更高的权重，并且在下一层生成聚焦的self-attention feature map（自注意力机制特征图），并利用attention block（自注意力模块）的n-heads权重矩阵大幅度降低输入的时间维度，并且添加一个max-pooling层，以进行最大池化操作，从而逐级减少特征图的大小，进而减少计算空间的内存消耗，最终在保持模型性能的同时减少模型的复杂度和计算资源需求，提高模型泛化能力，实现更快速的推理速度。**若。。。。**则时间序列长期依赖关系的单元λ值的更新公式：

其中包含Multi-Head ProbSparse self-attention（多头稀疏性自注意力机制） 以及 attention block（自注意力模块）的关键操作，表示在时间序列上的一维卷积操作，表示激活函数，用于非线性变换卷积特征，

**设计生成式预测模块**，结合编码器的特征提取和解码器的生成能力，以及上下文信息传递来提高生成的准确性和可靠性。将decoder模块的输入分为两部分，分别是待预测时间点前的一段已知序列，以及带预测序列的placeholder（占位符）序列，placeholder通常设为0,以防止每个位置关注未来的位置，避免了自回归。代表placeholder，代表开始字符，改进后的decoder输入如下：

其中，指将两个向量在相同维度上进行连接，生成一个更长的向量，表示模型维度。此外，选用MSE（Mean Squared Error）平均平方误差作为模型损失函数，用于评估模型对于多步预测任务的性能，并将其结果反向传播到整个模型。

**优化预训练和高效微调**，选择P-Tuning v2并结合使用预训练模型参数约束P-tuning v2过程的方法进行微调。让Prompt Tuning能够在不同参数规模的预训练模型、从而使得不同下游任务的结果上都达到匹敌Fine-tuning的结果。

**迁移和多任务学习（感觉是体现在时序数据的维度上的优化）**

1. **面向多源异构任务的多维度资源调度优化机制**

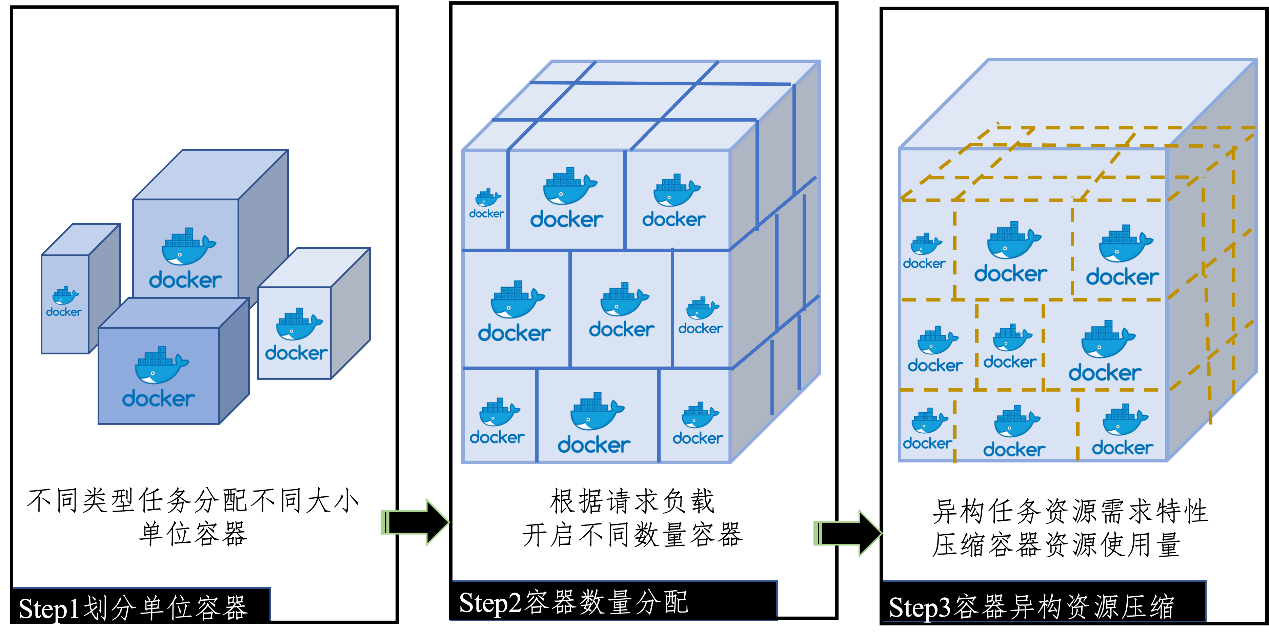
针对任务实时动态且资源需求异构等特征，本项目拟设计一种面向多源异构任务的多维度资源调度优化机制，通过对周期内任务进行时隙切分，并构建任务响应时延与资源配置间的相关关系函数，在服务器物理资源限制下，探究面向多应用集群响应时延和服务器总功耗之间的最佳平衡。

具体而言，**周期任务时隙切分**，将整个优化周期内实时异构的任务划分成T个时隙，并用t表示第个时间片段。对于时隙t内的第i种任务，每种任务的计算需求期望表示为。服务器上为处理任务分配通信资源和计算资源，其中计算资源包含CPU、GPU、内存和I/O等K种异构资源，用集合表示。考虑到不同传输环境下云中心传输速率不同，进而导致传输时延不同，本项目拟基于正交频分多址技术刻画云中心任务传输速率，以保证不同通信对之间不存在干扰。若表示信道带宽，表示来自其他服务区使用同一信道用户所造成的信道干扰，表示任务的传输通信功率，表示噪声功率谱密度，表示信道增益，则传输速率可表示为：。当服务器的数据传输量为且反向时延为，则任务i通过信道传输数据的时延为。而云中心的计算性能因异构任务的差异化资源需求，在相同云中心资源配置下存在差异化表现。因此，我们首先**构建异构任务对CPU、内存、GPU等计算资源的差异需求函数**，通过对异构应用程序在不同容器资源配置下的性能进行详细测试，采用非线性函数拟合的方式拟合——使用均方根误差（RMSE）作为评估指标来评估各模型拟合效果。若表示任务量、计算资源与计算时延的拟合函数**，**则任务的处理速率可表示为:。考虑到通常存在多个用户同时向服务器请求计算任务的情况，本项目拟通过M/M/N排队模型来刻画多个用户同时请求任务的过程，则云中心平均任务队列和平均处理时间可表示为。

其中，表示运行中的物理主机的数量。此外，出于优化云中心能耗的目的，并且由于物理主机活跃度一定程度上影响其能耗。因此，本项目拟刻画基于主机活跃度的能耗模型。即，将数据中心能耗刻画为活跃期和空闲期能耗的累加。若是第个处于活动状态的物理主机的功耗，且满足随应用容器集群的资源分配情况变化；q为处于空闲状态的物理主机的数量，表示第个处于空闲状态的物理主机的功耗；为当前数据中心可用主机的最大数量。因此，云中心的总功耗可描述为如下形式：

其中，，表示各类计算资源的利用率。基于云中心的时延和能耗关于其资源配置的表现，进一步**构建面向多应用集群的资源优化问题**，其优化目标为是在服务器物理资源限制下，寻找各应用集群的响应时延和服务器总功耗的最佳平衡。该问题可刻画为问题P1：

由于服务器的计算资源包含单位容器的计算资源和容器开启的数目，**该问题为混合整数非线性优化问题，本项目拟依托变量分解的方法进行问题分解，并结合凸优化理论、压缩技术和强化学习实现局部资源的动态伸缩管理和全局资源规划调度**。具体路线如图2所示。



**单位容器优化**：结合任务对资源的敏感特性确定出K种应用的最优单位容器计算资源配置。即通过松弛服务器计算和通信资源上限约束，根据任务对资源的敏感特性和服务器功耗曲线构建单位容器配置优化问题P2：

,

其中，表示任务的单位容器处理单位任务的计算时延，表示处理任务的单位容器在计算资源利用率为下的功耗。

**容器数量分配：**依托本项目的时序预测模型实现当前时隙任务到达率的精准预测，结合M/M/N排队模型估计任务平均响应时延，优化每种任务的容器集群最优的开启数目以满足响应需求。因此，基于容器集群最优开启数量的优化问题可构建为：  
由于主机物理空间有限，容器集群数量存在上限，因此本项目拟采用一维搜索方式推断最优容器开启数量。

**容器异构资源压缩：**当计算负载超出处理能力时，云中心需要增大主机开启数量以保障服务质量，但也导致能耗的显著提升。并且由于任务的处理效率和响应之间呈现非线性关系，并且异构任务调整相同的资源量，其计算时延的变化存在较大差异。因此，**本课题拟采用深度强化学习模型对容器集群配置进行微调，灵活运用容器水平、垂直伸缩机制动态调整异构任务的计算资源。**通过适当调整异构任务容器的计算资源，使得容器处理时延上升不明显的情况下，减少开启的服务器主机数，达到节能减排的目的。具体而言，当任务的容器集群在当前时隙内的状态集合为，其中 是容器的数量， 是分配给每个容器的计算资源份额，且，是各计算资源适当的单位调节量。则对于状态，存在一组可行适应动作 ，其中 是所有动作的集合可表示为 ， 表示垂直伸缩（即表示增加资源份额，表示减少资源份额）， 表示水平伸缩容器数量（即表示增加，表示缩减），表示不做决定。而对于每个三元组 ，通过定义其成本函数为 ，即状态从 过渡到 时执行动作 的成本，并将其设置为问题P1的目标函数，即任务计算响应时延、数据传输时延以及主机功耗的加权求和，从而依托A3C强化学习算法实现容器资源的最优配置和容器规模智能调整，最终优化系统性能和能耗。

1. **研究计划**

2.1．项目期限：以实际合同签字生效日期为准，合同期建议 **1** 年。

2.2．项目计划：

|  |  |  |  |
| --- | --- | --- | --- |
| **阶段** | **研究内容** | **时间** | **提交结果** |
| 1 | 基于自注意力机制的任务时序预测模型 | 2024.1—2024.5 | 申请国内/国际发明专利1项  投稿CCF-A类论文1篇 |
| 2 | 面向异构任务的多维度资源调度优化机制 | 2024.6—2024.8 | 申请国内/国际发明专利1项  投稿CCF-A类论文1篇 |
| 3 | 面向绿色计算领域的应用优化 | 2024.9—2024.11 | 完善时序基础预测模型及其源代码 |
| 4 | 项目解题资料整理 | 2024.12—2024.12 | 发表CCF-A类论文1-2篇  完善时序基础预测模型及其调度算法 |

1. **预期成果**

经过一年的科学研究，预计在算法原型、论文、专利等方面取得如下成果：

算法原型：完成多任务跨领域的时序基础模型设计，交付一套时序基础预测模型及其调度算法源代码；

论文：发表蚂蚁认可的CCF-A类或者领域内顶级会议/期刊论文1-2篇

专利：提交国内、国际专利2项

技术指标：与现有的**XXXX baseline**相比，提升资源利用率提升5%以上；

1. **人力、设备等投入及项目预算**

**4.1．人力投入（**包括全部参与项目的人员）

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **人员姓名** | **所属机构与职务** | **人员工资**  **（元/月）** | **参与项目时间**  **(折合全时人月数)** | **人员工资总计**  **（元）** | **承担职责** | **来蚂蚁实习时间（学生）** |
| 张永敏 | 中南大学、教授 | 0 | 6 | 0 | 项目负责人 |  |
| 王威 | 中南大学 博士生 | 3000 | 8 | 24000 | 项目主要成员 |  |
| 黄鹏宇 | 中南大学 博士生 | 3000 | 6 | 18000 | 项目主要成员 |  |
| 李朝辉 | 中南大学 博士生 | 3000 | 6 | 18000 | 项目主要成员 |  |
| 王雅晴 | 中南大学 硕士生 | 1500 | 8 | 12000 | 项目主要成员 |  |
| 黄锐 | 中南大学 硕士生 | 1500 | 8 | 12000 | 项目主要成员 |  |
| 李隆德 | 中南大学 硕士生 | 1500 | 8 | 12000 | 项目主要成员 | 1个月 |
| 刘世康 | 中南大学 硕士生 | 1500 | 8 | 12000 | 项目主要成员 | 1个月 |
| 董明屹 | 中南大学 硕士生 | 1500 | 4 | 6000 | 项目成员 |  |
| 翁静雯 | 中南大学 硕士生 | 1500 | 4 | 6000 | 项目成员 |  |

上述人员为乙方专门为本项目配备的专职人员，全职投入本项目。如若本项目需由乙方研究人员以实习、访问等形式进入甲方完成本项目下任务，蚂蚁将开放渠道，该部分人员的全部费用（包括差旅、住宿、报酬等）均已包含在项目总经费中，甲方不再另行支付。

**4.2. 设备投入**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **设备名称** | **数量** | **价格（元）** | **总计（元）** | **备注** |
|  |  |  |  | 无需新购设备 |

**4.3 项目预算**

|  |  |  |
| --- | --- | --- |
| **类别** | **总计（元）** | **备注** |
| ①人员支出 | 120000 | 与4.1对应 |
| ②研发支出 |  |  |
| 出差/会议费用 | 100000 | 参照标准：国内差旅单趟5000，国际差旅单趟15000，  预计1）国内差旅10人次，合计50000元；  2）国际差旅2人次，合计30000元；  3）学生到蚂蚁实习2人/次/月，预算20000元；  总计100000元。注：以上开销以实际开销为准。 |
| 设备费 | 0 | 与4.2对应 |
| 其它费用 | 50000 | 服务器升级改造、耗材、论文发表、专利申请等费用。 |
| ③ 直接费用 | 270000 | ①+② |
| ④ 间接费用 | 29700 | 学校学院管理费8%，专票税费3% |
| 总计 | 299700 | ③+④ |

**5. 教授个人简历**

张永敏，中南大学计算机学院教授，博士生导师，计算机系副主任，**教育部青年长江学者/湖南省百人计划入选者**。2015年毕业于浙江大学控制科学与工程专业，博士期间2013.11-2014.06赴美国加州理工学院进行联合培养；2015.11-2019.06年在加拿大维多利亚大学开展博士后研究，2019年6月加入中南大学计算机学院，被聘为特聘教授，隶属于张尧学院士带领的透明计算实验室。多年来，一直从事物联网、端边云协同、网络计算相关领域的资源管理与优化调度关键问题研究，在IEEE JSAC、ToN、TMC、TSG、TITS等国际权威期刊以及IEEE INFOCOM、SECON等国际会议上发表高水平论文40余篇，授权国家发明专利10余项。其中，发表在CCF A类期刊IEEE/ACM Trans. Networking 2016的论文自2016.6/7连续六年入选ESI高被引论文，获IEEE通信协会亚太地区“Best Paper Award”。相关研究工作获2019年“教育部自然科学一等奖”、2023年湖北省自然科学三等奖。主持国家自然科学基金项目2项，参与了科技部重点研发计划、国家自然科学基金重大项目/重点专项、湖南省自然科学基金重大项目、加拿大国家自然科学基金等十余个项目，积累了丰富的项目管理与课题研究经验。