目录

1. Design and Implementation of Multi-Core Heterogeneous Edge Computing Systems（多核异构边缘计算系统设计与实现）

申请链接：https://damo.alibaba.com/air/e97a1d9dd9264d4984d0578f5df85d06

背景：

Since the 2010s, the rapid development of information technology, especially the mobile Internet and the Internet of Things, has led to a continuous growth in the total amount of computing devices and data worldwide. The increased data volume has far exceeded the total accumulated amount in the past few decades. To meet the requirements of this scenario, two modes have been formed on the cloud and the end, namely cloud computing and edge computing. In general, based on big data centres, cloud computing can flexibly provide computing requirements from small to large scales and handle the computing requirements of various endpoint devices in a centralised manner, greatly expanding the flexibility of the business. However, cloud computing has its inherent weaknesses in terms of low latency, bandwidth energy consumption, and security and privacy. As such, edge computing has developed rapidly. Edge computing is a computing model for network edge devices, and its computing data includes the downlink data of cloud services and the uplink data of devices. Edge computing and cloud computing complement each other to support the current massive data processing. Nevertheless, the current edge computing still has the following problems:

1. Insufficient edge computing chip performance. With the constant development of edge computing, an increasing number of computing applications require edge computing processing. For example, a conference terminal , which simultaneously supports 6 cameras + 32 audio as an example, these audio and video data go through considerable algorithm processing such as splicing, identification, detection, sound source localization, intelligent noise reduction, etc., which require a chip with huge computing power. Unfortunately, there is no existing edge computing chip that can perfectly satisfy this scenario.

2.Huge cost of edge chips with high computing power. Taking Nvidia's NX as an example, the single chip provides 21T HashRate at a cost of around US$400. The high price limits its wide range of applications.

3. Difficult integration of multiple types of algorithms. The hardware on which current algorithms run are generally CPU, GPU, NPU, DSP, FPGA, etc. Various kinds of hardware have their own emphasis, each adapted to different types of algorithms, and complex applications usually require each algorithm to run concurrently to produce results. As such, it is necessary to integrate various algorithm hardware. However, it is difficult to

implement the above-mentioned algorithm hardware in a single edge computing chip.

Due to the Moore effect and the gradual development of the manufacturing process, it is difficult to obtain a single-chip edge computing chip with excellent performance, low price, and good adaptability to various types of algorithms at the same time. Is there a solution that can relatively achieve the above advantages at this stage? This project aims to study the design and implementation of multi-core heterogeneous edge computing systems and, by leveraging existing technical outcomes, design and implement multi-core heterogeneous edge computing architecture to allow the breaking through of the Moore effect from a different perspective, thus enhancing the computing capability on the edge.

Based on the above problems, we hope to design a multi-core heterogeneous edge computing system as follows

1. It uses existing PCIE/RAPIDIO/ETH high-speed interconnection technology to connect the single chips through interconnection at low cost to form a HashRate pool, which not only fully considers cost and energy consumption, provides a unified design model, but is also compatible with common operating systems.

2. It can provide a unified entry of HashRate and normalised HashRate. The multi-core heterogeneous system is invisible to users. For users, a multi-core heterogeneous system is equivalent to a unified computing power pool, and what visible is the total HashRate of a single type of processor (such as CPU total HashRate, DSP total HashRate, FPGA total HashRate).

3. It can also design a HashRate distribution model, provide a multi-core heterogeneous HashRate scheduling system, and implement distributed algorithm processing. Meanwhile, it can also use HashRate conveniently, flexibly, and efficiently according to the characteristics of various heterogeneous processors, and provide sufficiently adequate scalability, such as real-time addition and deletion of HashRate, and access to cloud HashRate.

With the realisation of this multi-core heterogeneous edge computing system, we would be able to conveniently perform edge computing in a normalised manner and achieve the expansion of computing power of edge computing in a low-cost way. It will surpass similar products in terms of price and performance, and meanwhile allow sufficient room for imagination to realise the business. The team is currently engaged in audio and video business intelligent hardware, which fits perfectly with the project. Once the research is successful, it can greatly improve the competitiveness and superiority of the team's products and overwhelm competitors' products on a higher dimension.

目标：

- Design block diagram of heterogeneous edge computing, including block diagrams and reference designs for interconnecting each heterogeneous processor
- Interface design of the normalised HashRate and source code implementation
- Communication design and source code implementation of various types of heterogeneous processor
- The design model, document and source code implementation of the HashRate dispatching system
- Publication of 1-2 papers in CCF-A category or top conferences and journals in the field recognised by Alibaba

相关研究课题：

- Edge Computing

返回目录

2. Research on Coordinated Optimization of Internet Data Center Computing Power Based on Elastic Scheduling of Workload（基于业务负载灵活性调度的数据中心算力协同优化研究）

申请链接：https://damo.alibaba.com/air/3a27c13b13b54482b77d8daaea48a96e

背景：

With the continuing growth of the digital economy and the steady promotion of New Infrastructure Strategy, internet data centers (IDCs) have become fundamental utilities to support China's socio-economic development. The energy demand of IDC industry, in particular, is sustainably increasing, attracting more and more attention. In this case, energy efficiency and elastic scheduling of flexible workloads is not only prerequisites for the sustainable development of IDC industry, but also the way to assist in achieving the national commitment of "3060" carbon neutrality.

随着数字经济的持续发展和"新基建"国家战略的稳步推进，数据中心成为支撑我国社会经济发展的重要新型基础设施。作为典型高载能行业，数据中心行业的蓬勃发展带来了巨大的耗电量。因此，提升数据中心运行节能水平和灵活运行程度，降低数据中心的用电成本和碳排放，不仅是数据中心自身发展的需要，也是推动"双碳"目标实现的必由之路，对企业的可持续发展具有重大意义。

Technically, the execution of some workloads can provide scheduling flexibility potential. In specific, delay-intensive workloads can be postponed in time scale

considering their lower priority. Delay-sensitive services can be spatially shifted to another server cluster for processing. Considering the electricity production and carbon emission changing with time and space, the potential of shifting workloads on temporal and spatial dimensions can be utilized for cost reduction, energy saving and carbon-free operation for IDC industry. For achieving this goal, it is essential to explore the capability of IDCs to change the power demand by elastic scheduling of workload and the sensitivity to follow the price and emission signal released from power systems. The coordinated operation of power systems and IDCs can contribute to real-time balance in energy supply and demand and reduce carbon emission. In particular, the current released strategy “East Data Computing in West”, representing moving east data to west areas, further emphasized the necessity of this work.
一般而言，数据中心离线业务和在线业务都具有调度灵活性，即延迟不敏感、优先级较低的离线业务可以在时间维度上调节处理时间，延迟敏感、优先级较高的在线业务可以在空间维度上改变处理的服务器节点。考虑电力供给成本随时间、空间变化的特点，以及不同区域电网的边际碳排放不同的背景，研究业务负载调度带来的数据中心用电特性变化和电力系统运行成本时空分布特性的匹配能力，从而允许数据中心参与电力系统的联合优化运行，对促进区域发展、能源资源供给平衡、缓解能源资源分布不均和减少碳排放的促进作用。特别考虑到国家近期提出的“东数西算”战略，更进一步强化了基于业务负载调度的算力-能源协同运行优化的必要性。

In Alibaba Group, Alimama team, which is in overall charge of the advertising business, has successful practical experience in shifting advertising streaming workloads spatially. Moreover, the advertising jobs include both delay-intensive and delay-sensitive workloads. Therefore, in this research, we intend to cooperate with Alimama team in exploiting the feasibility of spatial and temporal scheduling of advertising jobs theoretically and practically, such as scheduling algorithm design, power consumption model establishment, and practical implementation. This research project is the first step of “East Data Computing in West” implementation. And, it also makes Alibaba the first promoter of carbon reduction by workload scheduling in China. The pilot experience of this project will be replicable in China.
考虑到集团阿里妈妈事业群的广告算法部门具有成功将部分广告业务在空间调度的实践经验，而且业务范围涉及在线业务和离线业务。因此，本部门拟与阿里妈妈广告算法团队通力合作，分别从能耗模型和调度算法、理论设计和实践落地两个方面互补互助，以阿里妈妈广告业务负载的时空调度为主要研究对象，以点带面地探索数据中心时空维度调度的可行性。本项目迈出了我国“东数西算”项目实践的第一步，同时使阿里云数据中心成为促进我国互联网行业碳减排的示范者和推动者，项目成果将在全国积累可复制可推广的试点经验。

In addition, this project will cooperate with Jibei power grid, Jiangsu power grid, Zhejiang power grid, and western Inner Mongolia power grid in fully awareness of the cost and resource difference. This project can increase the renewable integration and quantify the ability to reduce IDC industries in different regions.

此外，本项目将与冀北电网、江苏电网、浙江电网、蒙西电网合作沟通，充分考虑各地方电网间的成本差异、资源差异进行算力调度，量化其促进当地电网消纳可再生能源、降低数据中心行业碳排放的能力；同时，依靠“全国一体化算力网络”的利好政策，推动政策试点、工程试点在内蒙古、长三角和京津冀地区枢纽节点的建设实施。

目标：

This project aims to propose a quantitative model for IDC workload scheduling that considers the actual operation requirements of workloads and scheduling systems. Then, the proposed model is applied in a power market scenario to measure the carbon reduction that IDCs can achieve by workload scheduling. Finally, the proposed model will be used in some pilots to participate in electric markets, electricity ancillary services markets, or demand response markets to evaluate the effectiveness of the proposed model in cost saving and carbon reduction. And the performance of IDC participating in the cooperation with power systems is also quantified.

本项目旨在提出一种考虑工作负载和调度系统实际运行需求的数据中心工作负载调度量化模型。然后，所提的模型会在电力市场的场景中进行应用，来衡量数据中心能够通过负载调度来减少的碳排放量。最后，模型会在一些试点地区参与电力市场、电力辅助服务市场或需求响应市场，来评估模型在减少运营成本和碳排放的实际有效性，同时量化数据中心在参与电力系统协同优化方面的作用。

This project proposes to conduct an in-depth research on three technical problems:

本项目拟从三个技术问题开展深入研究。

- A mathematical model of our current scheduling system, including resource allocation and virtual machine allocation of different kinds of workloads.
  一种描述我们当前调度系统的数学模型，包括不同工作负载的资源分配和虚拟机分配模块。
- A simulation model of the cooperation of IDCs and the power system, considering workload scheduling and power system operation
  一种考虑算力调度和电力系统协同运行的模拟模型
- A scalable testing environment for testing the reliability and feasibility of the cooperation model
  建立一种能扩展的测试环境，以测试项目提出的协同运行模型的有效性

相关研究课题：

- Analysis on the scheduling boundaries based on the current requirements of scheduling systems for different kinds of workloads.
  根据业务调度系统要求，对各类实际调度约束条件进行分析评估。
- A data-driven model of the mapping relationship between workloads execution and power load of servers or IDCs.
  基于数据驱动方法的业务负载和数据中心用电外特性映射关系建模。
- A cooperation model between IDCs and power systems considering workload elastic scheduling, considering actual scheduling boundaries.
  建立考虑灵活性调度约束条件的数据中心和电力系统协同运行模型。

- Detailed implementation plans of IDC participation in the power markets, electricity ancillary service markets, or demand response markets for our different IDC bases.
  针对不同数据中心基地，设计数据中心参与电力市场、电力辅助市场、或者需求响应市场的详细实施方案
- Evaluate the performance of the proposed model in carbon and cost reduction for different IDC bases.
  针对试点实际情况，对负载调度所产生的碳减排量进行评估和量化，将研究成果进行实践。
- Sensitivity analysis of electricity price, energy resources, and carbon intensity on the performance of the proposed model in carbon and cost reduction.
  研究电力价格、电源种类、碳排放强度等指标作为业务负荷调度因子对降低云计算用能成本和碳减排的作用。

返回目录

3. Cloud Storage Oriented DPU Architecture（面向云存储的 DPU 关键问题研究）

申请链接：https://damo.alibaba.com/air/ed61e6e27de140f485f9be52d3593ab2

背景：
Except to CPU and GPU, DPU is innovative processing-unit technology in the world. With the rise of DPU, CSPs (Cloud Service Providers) and start-up companies pay great attention to it. While it's clear that DPU can accelerate virtualization-related functions like virtio-net and virtio-blk in computing scenes, it's unclear how DPU should deal with data movement and data processing in storage scenes. Thus, it's important to explore how DPU is adopted in cloud storage.

Meanwhile, CSPs are developing hardware and software co-design technology to improve competitiveness. DPU is one of most promising technology for codesign. Thus, it's quite important to study DPU in Alibaba, and it's challenging for cloud storage.

Different from virtualization-related computing and acceleration in computing scenes, storage is IO related and focuses on data movement and processing. The architecture of DPU for storage doesn't focus on functions like CPU Cache, ALU and Virtualization, and focuses on data processing functions like data compression and encryption. However, such DPU architecture of data processing is unclear now.

Especially, the CPU cores of DPU like Bluefield and IPU is not found out its role in storage scenes. Thus, it is important to study DPU architecture for cloud storage.

目标：

- A study report for DPU architecture, which should be related to IO datapath in cloud storage.
- A DPU PoC for DPU architecture, which should prove it's able to improve the capacity of data processing.
- One international paper and three innovative solutions.

相关研究课题：

- Storage accelerations like compression and EC.
- Different between DPU and multicore DPU.
- The role of ARM cores in DPU.
- The runtime system of DPU for accelerations, network and ARM cores.

返回目录

4. Next-generation video hardware encoder（下一代视频硬件编码器）

申请链接：https://damo.alibaba.com/air/4b92f39c5e1649e996dd4b35237f2370

背景：

In recent years, video technology has developed rapidly. At the same time, video application scenarios are becoming more abundant. Typical application scenarios include live video streaming / short video clips / video conferences / telecommuting / virtual reality and so on. In order to achieve high-quality video compression and transmission, related organizations have developed a series of video coding standards, such as H.264/AVC and H.265/HEVC. Based on these popular video coding standards, a lot of software and hardware solutions for video codec has been derived to solve the problems of video compression in practical applications. However, traditional video encoder in the software side relies on CPU to implement the motion estimation and other procedures, which is difficult to meet the exponentially increasing for computing power with the development of coding standards. Considering the huge occupancy of CPU resources by software encoders, it is difficult to cater to the trends for real-time and miniaturization of video encoders in actual application scenarios. Compared with software encoders, hardware encoders make full use of the high parallelism of related algorithms. The hardware encoders can increase the calculation density by more than ten times and save the energy consumption of the servers by more than ten times than the software encoders in the general computing platform. At present, Internet companies at home and aboard (for example, Google / Amazon / Huawei / Bytedance / Tencent / Kuaishou) have invested heavily in the development of video hardware

encoders. Hardware encoders are expected to save hundreds of millions of dollars each year for these companies. It can be seen that video hardware encoders are an important development direction in the future. At the same time, it is also a key chip technology that the country or enterprise needs to master.

In order to achieve a breakthrough in the hardware encoder, our team has set the goal of achieving the lowest cost and best virtual quality in the industry, and carried out a lot of research and development work. The hardware encoder that we propose integrates multiple functions of video encoding and processing, save the cost of customers and upgrade the video viewing experience. At the same time, the reduction of power consumption has also contributed to environmental protection.

We are currently advancing the research of next-generation video hardware encoders. Our goal is to further improve the video quality and be compatible with more video coding standards. By controlling the cost of the chip, our company can provide better video services with lower machine and bandwidth cost.

目标：

- Complete the algorithm development of H.266/VVC hardware coding. Compared with the common chip solutions of the H.265/HEVC encoder in the market, the compression rate is expected to increase by more than 25%. Compared with the existing H.265 hardware coding algorithm, the encoding algorithm complexity of VVC is expected to increase by no more than 50%.
- Complete C MODEL of H.266/VVC encoding algorithm. The architecture is compatible with the existing H.265 encoding architecture, and completes the cross-validation with the encoding algorithm.
- It is expected that the partner can complete 2 technical patent applications.
- Expect the partners to deliver code, algorithm design documents and quarterly progress reports.

相关研究课题：

- Video Compression and Processing
- Versatile Video Coding (VVC) standard
- Video Coding Mode Decision
- Hardware implementation and optimization for Video Coding
- To improve the coding algorithm performance of H.266/VVC for hardware design
- To reduce the cost and power consumption of the hardware implementation
- A Low Power Versatile Video Coding (VVC) Loop Filter Hardware
- Adaptive CU Split Decision for VVC encoding

返回目录

5. High Density and High Performance mmWave Networking System（基于跨模态电磁环境建模的 5G 毫米波高速连接与动态部署技术研究）

申请链接：https://damo.alibaba.com/air/46f556554426465fa3d189af4e0cda26

背景：

Immersive virtual environments such as Metaverse and extended reality (xR) demand high-performance wireless networking in the order of 1Gbps throughput and 10ms latency. The challenge is further amplified when there are a dozen or even more such devices within one room. How do we build a high-density and high-performance wireless networking infrastructure to support such applications? According to the current development of 5G as well as the future trend of 6G, exploiting higher-frequency wireless techniques is one of the mainstream ideas to alleviate this problem. Among them, mmWave with 30GHz~300GHz carrier frequency may be the potential answer due to its broader available spectrum for a higher channel capacity and it narrower RF beam for more directional connections. These exquisite features are not conveniently available in sub-6GHz systems due to the strict spectrum regulation and the large antenna size.

However, it's a non-trivial task to build a high-density and high-performance mmWave networking system with the state-of-the-art techniques. We believe the management of the mmWave resources becomes more challenging: despite conventional multiplexing methods in the time and frequency domain, there is an additional degree of freedom - space multiplexing - as the result of directional beams in mmWave systems. The additional DoF is a double-edged sword, which could on one side enable denser links when properly used, but on the other side create headaches of discovering, assigning, and tracking beams in both static and dynamic environment. Moreover, the careful design of interference management among multiple access points and end-user devices is also a necessity. All these have to be done in an efficient manner to ensure that the disadvantages do not cancel out all the advantages.

Despite confining the management on the PHY layer of each device, the design of higher layers is also indispensable to optimize the system performance. For instance, we believe that a centralized higher-layer design that is tailored to each application scenario would enable sharing information and control in a cross-layer, cross-device manner, thus making it possible for higher overall performance and directly optimized for target applications. As a starting point, one or more sample applications can be chosen to make it straightforward to design and benchmark the performance of the wireless network and the application experience. Sample application scenarios include Metaverse in an office environment or xR gaming in a store setting.

目标：

- A large-scale mmWave testbed consists of up to hundreds of devices and multiple access points with full access to low-level PHY control on both devices.
- A management system enables centralized application/scenario-aware orchestration of mmWave resources, achieving efficient, low interference scheduling of mmWave resources.
- Throughout tests with demo applications showing tens to hundreds of Gbps aggregated throughput in a room-scale while meeting stringent application requirements such as xR and metaverse.

相关研究课题：

- Indoor wireless environment sensing and modeling.
- Beam generating and steering for the phased-array device or the distributed phased arrays.
- Multipath effect exploitation for throughput enhancement.
- Distributed wireless networking resource management.
- Large-scale networking testbed construction.
- Future networking infrastructure for xR and Metaverse.

返回目录

6. Large scale real-time prediction and decision-making based on time and space（大规模基于时空的实时预测及决策）

申请链接：https://damo.alibaba.com/air/8d94fa076d2246fbbd0b762e2ba47b91

背景：

In the international import and export logistics, domestic warehousing and parcel distribution scenarios, there are a large number of problems such as space-time prediction, order fulfillment, resource allocation and capacity scheduling. It is necessary to combine large-scale machine learning and operation research optimization technology to build a dynamic space-time decision optimization engine, so as to achieve the effect of cost reduction and income generation of each business.

Technical research needs to make breakthroughs in several core issues, including spatio-temporal prediction algorithm system, dynamic online decision optimization and end-to-end prediction decision exploration; Behind this is the comprehensive application and combination of operation research optimization, deep learning and

reinforcement learning technology. At the same time, it also needs a set of algorithm engineering with the computational power required for large-scale online decision-making.

Therefore, we build a large-scale real-time prediction and decision algorithm platform based on logistics time and space to provide support for similar industrial Internet products, This is also different from the consumer Internet. What we build is a customized prediction and decision algorithm platform under the industrial Internet, which will greatly help the real needs of a large number of prediction and decision-making within CaiNiao logistics, and can also provide a basic framework platform for similar external problems.

目标：

- A large-scale online prediction and decision-making system based on spatiotemporal logistics for industrial Internet.
- Our forecasting and decision-making system has been applied in smart storage, smart transportation and smart distribution, and the business effect has been significantly improved.
- Our system is open to the industrial Internet and can be recognized by the industry awards.

相关研究课题：

- Spatiotemporal prediction technology
- Representation technology based on large-scale spatiotemporal AI
- Real time decision technology based on Prediction
- End to end decision technology
- Exploration of decision making based on graph model
- High performance decision optimization computing technology

返回目录

7. Multi-modal Learning for Emotion Recognition（用户情感识别的研究和应用）

申请链接：https://damo.alibaba.com/air/23a1a33f1cc1400fbcf8814ee4f39866

背景：

Tmall Genie aspires to become an intelligent speaker that understands the users' mind, and emotion recognition is one of the most important ways to achieve this. With the help of emotion recognition, Tmall Genie can provide specialized services for users.

For example, for the same request "Tmall Genie to play music", some cheerful music can be played when the user is in a happy mood, and some light music can be played when the user is tired. Besides, the emotion recognition results can also help the Tmall Genie to provide customized dialogue content as well as judge the quality of services, which can form a technical closed loop. The above process can help us to continuously polish and iteratively upgrade our artificial intelligence technology. Therefore, it is very necessary to perform accurate emotion recognition.

Towards the task of emotion recognition, numerous of approaches have been proposed. Discriminative feature learning is critical to the final results. For example, the linear discriminative analysis can realize this by minimizing the intra-sample distance as well as maximize the inter-samples distance. However, it relies on a large amount of labeled samples, which is very expensive to manually construct such kinds of dataset. By contrast, we can easily collect unlabeled samples from the Internet. In this case, one can rely on self or semi-supervised learning to learn discriminative feature representations.

Another important factor for emotion recognition is multi-modal fusion. For example, we can collect not only the audio data but also the video as well as images from the users. By making full use of these multi-modality data, the accuracy of emotion recognition can be further improved. It is meaningful to investigate the multi-modal fusion strategies.

目标：

- Deep learning based modality-specific emotion recognition model with limited labeled samples
- Large-scale self-supervised and contrastive learning or pretrained model for discriminative feature learning
- Effective multi-modal fusion method to improve the accuracy of emotion recognition

相关研究课题：

- Videos based facial expression recognition in the wild
- Feature-level and decision level fusion for multiple modalities
- Transfer learning for emotion recognition
- Emotion recognition under the situation of incomplete modalities
- Unsupervised large-scale pretrained model
- Cross-modal retrieval and recommendation

返回目录

8. High-performance Computing, Numerical Weather Prediction, and Machine Learning（AI 加速天气模式计算）

申请链接：https://damo.alibaba.com/air/294e7d28b0494f7dacd4926a9d097fcc

背景：

气象预报是国计民生的基础设施。准确快速地预测和跟踪极端天气和气候事件对于国民安全，农业生产等都有重要意义。基于气象预报的风能和太阳能预测对风电厂、太阳能电站的选址、功率预测等能起到最核心的作用，是新能源提升发电量和消纳率，保证其高质量运行的基石。从气候变化科学的角度来看，能够提升温室气体监测评估的能力，为国家实现碳达峰和碳中和目标提供更多的科学支撑。与此同时，气象预测方面的研究能推动基础科学的进步（超算架构、计算科学等），有非常好的科研价值和社会影响力。此外，气象预测在航空、交通、农业、物流、金融等其他行业也有着广泛应用。

天气预报包含多个物理模式（每一个对应一组微分方程），通常在超级计算机上进行计算求解。中国的机构和学校多使用气象研究和预报系统（WRF），这套模式相对陈旧，性能也比较局限，所以通常在 WRF 的预报基础上，还会从国外机构（比如欧洲中期气象预测中心 ECMWF 和美国全球预报系统 GFS）采购其数值模式的结果来改进预测。但是最精确的数据和最先进的模型一般不对外开放，因而难以获取。AI 的兴起，给科学计算提供了新的思想方法和数学工具，能帮助我们求解之前难以企及的问题，在数值气象预报模式的计算速度和精度上，也给我们提供了弯道超车的可能性。

Weather forecast is an infrastructure of national welfare and people's livelihood. Rapidly and accurately predicting and tracking extreme weather and climate events are of great significance to national security, agricultural production, etc. Wind and solar resource predictions from weather forecasts play an essential role in the site selection and power prediction for wind and solar power plants. It is the cornerstone for new energy to increase power generation and consumption rate, ensuring high-quality operations. Weather forecast capability also improves the ability to monitor and evaluate greenhouse effects from the climate change perspective, providing more scientific support for achieving the carbon peak and neutrality goals. As a research topic, it promotes the progress of fundamental sciences such as supercomputer architecture and scientific computing, with high research values and community impacts. Moreover, weather prediction also benefits many other industries, including aviation, transportation, agriculture, logistics, finance, etc.

Weather forecasting has multiple physical modes (each corresponds to a set of differential equations), usually solved on supercomputers. Chinese institutions and universities mostly use the Weather Research and Forecasting (WRF) model system.

This model, on the other hand, is rather outdated and has a restricted performance. Therefore, data from other institutions, such as European Center for Medium-Range Weather Forecasts (ECMWF) and Global Forecast System (GFS), are often purchased and deployed to improve the forecasts. Unfortunately, the most accurate data or state-of-the-art models are often under export control and are publicly inaccessible. The rise of AI has provided scientific computing with new thinking perspectives and mathematical tools to help us solve previously unattainable problems. There is a great opportunity for us to overtake in the computational speed and accuracy for numerical weather predictions, with our strengths in AI.

目标：

本研究项目的目标是更高效的气象和气候方针，针对多种单一物理模式以及多物理模式的融合，进行快速推演、数据融合、纠偏以及根据时间和地理特点进行准确的模式选择。主要的研究方向包括但不局限于：

- 针对气象方程的数值算法和底层数值代数的优化；
- 并行计算的设计和针对 HPC 的性能优化，尤其是在异构体系上的优化；
- 利用 AI 修正物理模型（包括物理参数化过程）及其预测；
- 针对传统四维变分和卡曼滤波等技术，利用 AI 加速或者简化数据同化过程；
- 利用 AI 加速数值模拟，包括加速物理参数化过程和直接加速气象微分方程的求解；

The project aims to provide more efficient weather and climate policy with the integration of many single-physics and multi-physis models, which enables rapid inference, data assimilation, model correction, and accurate model selection based on temporal and geographic characteristics. The research focuses will include but are not limited to the following topics:

- Specific optimization on the numerical methods and low-level numerical linear algebra routines;
- Parallelization and performance optimization on High-Performance Computing (HPC) systems, especially on the emerging heterogeneous architectures;
- AI corrections on the numerical model and its predictions, including physical parameterizations;
- AI accelerations or simplifications for data assimilation techniques, such as 4D-Var and Kalman filters;
- AI accelerations for numerical simulations, including physical parameterizations and directly solving the underlying PDEs;

相关研究课题：

高精度的气象模拟往往依赖于庞大的空间计算网格，对计算资源的需求极高，因此在高效的气象模拟中如何优化底层数值算法，如何优化计算并行和资源分配，以及如何利用 AI 改进和加速数值模拟等显得尤为重要。具体的需求方向包括：

- 针对气象微分方程的底层数值方法，CPU/GPU 并行技术；
- 在算法层面利用 AI 技术改进物理模型和加速数值模拟；

High-resolution numerical weather prediction often relies on massive computational grids, which poses an extremely high computational burden on the simulation. As a result, a variety of critical factors have to be considered in order to achieve better efficiency, such as how to optimize the low-level linear algebra routines, code parallelization, resources optimization, and how to speed up the solution procedure for differential equations. Specific directions in demand are:

- Numerical methods for meteorological PDEs, CPU/GPU code parallelization;
- Use AI to correct the physical models and to accelerate the numerical simulations at the algorithm level.

返回目录

9. Nationally differentiated search algorithm and technology for cross-border electronic commerce（跨境电商国家差异化搜索关键技术研究）

申请链接：https://damo.alibaba.com/air/323b3023cb814a14a76838d33ada28e5

背景：

With the rapid development of cross-border e-commerce in recent years, nationally differentiated search algorithms and technologies have become increasingly important. Take AliExpress (AE) as an example, it is one of the biggest cross-border business-to-consumer (B2C) electronic commerce platforms in China owned by the Alibaba Group. AE serves users from more than 200 countries and regions around the world. Consumers from different overseas countries behave differently due to the differences in geography, language, culture, politics and economy. Among the top-four countries with the highest traffic volumes in AE, being Russia, the United States, France and Spain, the overlapping rate for products displayed in AE is 31.8%. However, the overlapping rate is only 17.1% for clicked products, and 5.1% for the purchased products, validating that fact that the user behavior patterns may differ across countries. In Poland and France, which are geographically close in Europe, the overlap of the top 10 thousand queries in English is only about 20% a day. Taking into account the impact of historical, cultural and linguistic origins, there is an inherent close correlation between consumers in different countries, which could be well used to design nationally differentiated search algorithms and technologies.

Meanwhile, it is a big challenge to model and capture the consumers' shopping interests and preferences, since most overseas consumers have very few online shopping behavior records. Amongst AE's daily active users , about 20% have not visited the site in the past one month and about 40% have viewed fewer than six product pages in history. What is worse, about 50% of users do not have a single order

record in the past month. How to design and build a cross-border nationally differentiated algorithm model under the current situation of space and low-activity data is not only of algorithmic theoretical value, but also useful for guiding the domestic e-commerce companies in opening up the new sinking markets (such as the domestic second-tier and third-tier cities).

In the process of cross-border e-commerce development, especially the fast expansion of the local sellers selling to the local consumer business model, we are facing the following problems:

1) How to store and index both the local sellers' products and the global sellers' products in search engine?
2) How to efficiently and intelligently allocate the quota volume for the locally selling products and the globally selling products in the seeking phase within a search ranking process?
3) How to balance and control the exposure of the locally selling products and the globally selling products?

If we can well solve the above problems, it will bring in the long-term business development of the global main-website as well as the local national-website.

Mostly importantly, considering the EU's privacy protection policy, the Russia data export control restrictions and other overseas local policies and regulations, how to provide efficient personalized search and recommendation algorithm services in the absence of some user core identification information and online behavior data information while protecting the user's privacy? As we know, machine learning algorithms such as federated learning and few-shot learning (FSL) have been successful applied to deal with similar problems in industry. For example, federated learning can train a machine learning algorithm model, for instance deep neural networks, on multiple local datasets contained in local nodes without explicitly exchanging data samples. In light of the development in machine learning, we wish to adopt and design some novel nationally differentiated search algorithms to protect the user's privacy.

目标：

- An effective nationally differentiated search algorithm for modeling and predicting the users' shopping interest and preference, which could be used in AE online ranking system.
- Some novel large-scale product indexing and recall algorithms and technologies realized easy, can work steady and economize memory space.
- A user privacy protection prototype system.

相关研究课题：

- Learning to rank in E-commerce
- Multi-task learning in E-commerce
- Deep Neural Network for CTR/CVR prediction
- Document indexing and data compression technology

- Federated learning in E-commerce
- Transfer learning for CTR/CVR prediction
- Zero sample learning and few-shot learning (FSL)

返回目录