ASC Student Supercomputer Challenge 2024

Preliminary Round Notification.

Dear ASC24 Teams.

Welcome to the 2024 edition of the ASC Student Supercomputer Challenge (ASC24)!

The ASC Student Supercomputer Challenge is now in its 11th edition of continuous success. It has become the world's largest supercomputing hackathon, striving to foster the next generation of young talents, inspire exploration, innovation, and collaboration in supercomputing and AI. After the ASC24 registration kicked off at SC23 on November 16, 2023, we received enormous interest from hundreds of registered teams. Now, the competition is moving to the next phase: the Preliminary round!

Preliminary Round.

In the preliminary round, the ASC24 Committee looks forward to each team giving their best effort in accomplishing all the tasks approved for this competition. Teams are required to submit proposal documentation, including cluster design details, source code optimization approaches, and output files. The ASC24 evaluation committee will review the proposals in English.

Submission Guidelines.

All ASC24 teams are expected to upload the following items to the ASC24 official website (http://www.asc-events.net/StudentChallenge/ASC24.html) by 24:00 on January 21, 2024 (UTC/GMT +8:00).

a) The proposal document should be named by combining the university or college name and the contact person, for example: ABC University_John_Doe. The document should be in either .docx or .pdf file format.

b) All additional information should be compressed using ZIP or other tools into one file, using the same naming convention: ABC University_John_Doe. The compressed file should include at least four folders, as per the requirements detailed in Appendix A.

- Output files of HPL
- Output files of HPCG
- Required files of OpenCAEPoro
- Required files of LLM inference

A confirmation email will be sent shortly after receiving all the above information. For any further inquiries, please contact the ASC committee via:

- Technical Support: techsupport@asc-events.org
- General Information: info@asc-events.org
- Press: media@asc-events.org

Wishing you the best of luck in your ASC24 journey!

ASC24 Committee

Appendix A:

Proposal Requirements.

I. Introduction of the university department activities in supercomputing (5 points).

1. Supercomputing-related hardware and software platforms.
2. Supercomputing-related courses, training, and interest groups.
3. Supercomputing-related research and applications.
4. A brief description of the key achievements in supercomputing.

II. Team introduction (5 points).

1. Brief description of the team setup.
2. Introduction and the photo of each member, including group photos of the team.
3. Team's motto or catch-phrase.

III. Technical proposal requirements (90 points).

1. Design of HPC system (15 points).

a) The system should be designed for the best computing performance within a 3KW power consumption limitation.

b) Specify the system's software and hardware configuration and interconnection. Describe the power consumption, evaluate the performance, and analyze the advantages and disadvantages of your proposed architecture.

c) The components listed in the table below are provided for reference only. They are based on the IEIT NF5280M7 server, which supports up to 2 GPUs.

Item	Name	Configuration
Server	IEIT NF5280M7	CPU: Intel® Xeon® Gold 6430 Processor * 2 Memory: 32G x 16，DDR5，≥4400 MT/s Hard disk: 480G SSD SATA x 1
HCA card	HDR	InfiniBand Mellanox ConnectX®-7 HDR
Switch	GbE switch	10/100/1000Mb/s，24 ports Ethernet switch
	HDR-IB switch	Mellanox Quantum (TM) HDR InfiniBand Switch, 40 QSFP56 ports, 2 Power Supplies (AC), unmanaged, standard depth, P2C airflow, Rail Kit, RoHS6

Cable	Gigabit CAT6 cables	CAT6 copper cable, blue, 3m
	InfiniBand cable	InfiniBand HDR copper cable, QSFP port, compatible with the InfiniBand switch in use.

◆ The hardware configuration in the ASC24 competition finals may differ from the table above.

2. HPL and HPCG (15 points).

The proposal should include descriptions of the software environment (operating system, compiler, math library, MPI software, software version, etc.), performance optimization and testing methods, performance measurement, problem, and solution analysis, etc. In-depth analysis of HPL and HPCG algorithms, along with the respective source codes, would be a plus.

Download the HPL software at: http://www.netlib.org/benchmark/hpl/.
Download the HPCG software at: https://github.com/hpcg-benchmark/hpcg

It is recommended to run verification and optimization of HPL and HPCG benchmarks on x86 Xeon CPU and Data Center GPU platforms. If other hardware platforms are used, you are welcome to submit related analyses and results that demonstrate adequate performance.

3. Optimization for LLM inference (30 points).

Task Description.

Large Language Models (LLMs) have revolutionized the field of AI, demonstrating unprecedented capacity across a wide range of tasks. However, the inference process for LLMs poses certain challenges. For instance, the models are often too large to fit into memory, resulting in prolonged inference times and substantial computational costs. These obstacles hinder their widespread implementation in various applications.

To emphasize the significance of LLM inference and inspire enthusiasm among undergraduates in this field, LLaMA2-70B, a LLM with 70 billion parameters and a dataset with 10 thousand samples provided by the Committee are used to optimize inference performance in the preliminary of ASC24.

Despite the availability of optimized and high-performance inference frameworks such as TensorRT-LLM and vLLM, we encourage participants to begin with the baseline code and build a tailored, high-performance inference engine, considering the architectural characteristics of their HPC cluster. Also, during the preliminary stage, to discourage teams from focusing solely on low-precision optimization, participants are only permitted to use FP16 or BF16 precision based on their computing devices. The use of 8-bit or any numerical precision lower than that is strictly prohibited. Participants are required to provide a detailed description of their optimization strategies as well as the results achieved in their proposals.

Dataset.

The dataset used in the preliminary has 10k samples with prompts, prompt length and output length. This dataset has the following characteristics:

- Multi-domain Coverage: The dataset contains text data from various domains, including news, encyclopedias, novels, forums, and more, and covering different topics, styles, and viewpoints, enabling the model to have better generalization across different domain tasks.
- Multilingual Support: The dataset includes text data from multiple languages such as English, Chinese, Korean, Spanish, etc., which allows the model to understand and generate text across different languages.
- Large-scale Data: The dataset is sampled from a massive amount of text data, which helps improve the language understanding and generation capabilities of the model.
- Length Diversity: Too long and too short sequences are filtered out. The dataset contains 10k samples with length range from 4 to 1024, covering a vast majority of the length range for everyday use.

In summary, the dataset provides high-quality language samples with multi-domain coverage, multilingual support, large-scale data, and diversity characteristics. The dataset could be downloaded from the ASC repository: https://github.com/ASC-Competition.

Result Submission.

For the preliminary round, each team should submit a folder containing running log files, the source code, and any other necessary files that could be used to reproduce the running process. Please submit all the requested files in the following format:

Items	Contents
LLM_inference	Root directory.
LLM_inference/Log	Inference log file.
LLM_inference/*.py	Language model inference script or code files used in the inference process.
LLM_inference/proposal	Doc of pdf file including the results and the comprehensive optimization methods.
LLM_inference/other_items	Other necessary files used to reproduce the running process also needs to be included in this folder.

Evaluation.

During the scoring process, the ASC24 Committee will initially focus on the improvement in inference throughput performance. Furthermore, proposals with detailed optimization strategies and underlying principles will also be considered as the primary basis for scoring. Here are a few important points to be aware of:

(1) Participants must submit all the files required in the LLM_inference directory mentioned above; otherwise, the score for this part will be set to 0.

(2) FP16 or BF16 should be used during the inference processing, FP8, INT8 or INT4 are not

allowed, and otherwise the score of this part will be set to 0.

(3) Any truncation for input prompt or output length is not allowed during the inference processing. Otherwise, the score of this part will be set to 0.

(4) The submitted inference log file needs to include the following contents:

   a) All hyper parameters, which including random seed, model path, dtype and so on.

   b) Information of each iteration step, including time consumption and other necessary items during each iteration.

   c) Information about the entire process, including throughput (requests per second), total number of tokens, total number of prompt tokens, total number of output tokens, and any other necessary details.

(5) Participants are required to provide comprehensive details of the model inference process, machine specifications, and optimization methods used. These details will serve as the primary basis for scoring by the ASC24 committee.

(6) Greedy search is used for token generation. The related parameters including num_beams (set to 1), do_sample (set to false), temperature, top-p, max_new_tokens, etc. used for token generation in the baseline code are not allowed to be modified.

Baseline Code and Model Weights.

The ASC24 committees supply a baseline code for this task, which can be downloaded from the following link: https://github.com/ASC-Competition. The participants could start from it and modified it for high inference performance. Besides, the model weight of LLaMA2-70B could be downloaded from: https://huggingface.co/meta-llama/Llama-2-70b.

4. The OpenCAEPoro Challenge (30 points).

Task Description.

OpenCAEPoro is a numerical simulator designed for multi-phase and multi-component flow in porous media, with a particular emphasis on petroleum reservoir problems. It is built using C++ and incorporates MPI parallelism to enhance its performance and scalability.

To run the application, some external libraries are required. All these libraries, along with OpenCAEPoro, can be accessed through our GitHub repository (https://github.com/OpenCAEPlus/OpenCAEPoro_ASC2024). It is important to ensure that your system has these libraries installed to facilitate a smooth operation of OpenCAEPoro. The complete installation steps could be found in Readme.md.

This case is designed for three-phase(oil, gas, water) 3D black oil simulation, it has a three-layer permeability structure and contains an injection well and a production well. Gas is injected in the top layer, whereas oil is produced from the bottom layer. The gas injection results in swelling of oil and in gas sweep towards production well. The reservoir is closed, which means no mass or heat exchange between reservoir and the external region. Case 1 has the size of 10000ft × 10000ft × 100ft, meshed uniformly by 80×80×40. This case will simulate for 3600 days.

The test case could be found under the OpenCAEPoro main directory and run with following command:

```
cd OpenCAEPoro/
mpirun -np <core_num> ./testOpenCAEPoro ./data/case1.data verbose=1
```

Goal.

The target of this task is to minimize the **OBJECT TIME**, which will be printed on the screen after the simulation finished. The region of code to be modified and optimized are

1. OpenCAEPoro

2. PETSC_FIM_solver.cpp in petsc_solver (lines 112th-210th, which has been marked)

Note.

1. Any code that is related to the method parameters is not allowed to be modified.

2. All of the parameters in the input files should NOT be changed.

3. Please keep the directory structure and the resulting files unchanged.

Validation.

The output file **SUMMARY.out** will be generated after the simulation. There are many items in it and we will select some of which at specific time point for result verification. Specifically, the chosen items for validationare

```
FPR: field average pressure
FOPR: field oil production rate
FGPR: field gas production rate
FWPR: field water production rate
WBHP-INJE1: bottom pressure of INJE1
WBHP-PROD1: bottom pressure of PROD1
```

and the above variables will be checked at the set of time points Tth day.

```
50, 180, 360, 720, 1080, 1440, 1800, 2160, 2520, 2880, 3240, 3600
```

The existed SUMMARY_ref.out is for reference. For example, if

$$\frac{1}{|T|}\sum_{t\in T}\frac{|FPR(t)-FPR^{ref}(t)|^2}{FPR^{ref}(t)}<0.01$$

where $|T|$ is size of T and we have $|T|=12$ here, then item FPR passed the verification. The simulation result is regarded as correct only when all items satisfy the above formula.

Result Submission.

When you have finished your optimization and simulation, please submit all the requested files of the task case in a compressed file with the following content:

Compressed file name	Contents	
OpenCAEPoro.tar.gz	petsc_solver/src/PETSC_FIM_solver.cpp	Modified source code in petsc_solver.
	OpenCAEPoro/src	Directory of modified source code in OpenCAEPoro
	case1	Task case directory includes: 1. Input file (case1.data) 2. Screen output during simulation procedure (testOpenCAEPoro.log) 3. Output files generated during the simulation (FastReview.out, statistics.out, SUMMARY.out). 4. Validation file SUMMARY_ref.out (optional)

For example, in case1 directory, all files which are required for submitting are shown in the picture below:

```
> ls case1
case1.data  FastReview.out  statistics.out  SUMMARY.out  SUMMARY_ref.out
```

Evaluation.

The correctness of results and performance improvement will be key points in the evaluation procedure. The proposal should include detailed optimization strategies, and the underlying principles will also be considered as the main basis for scoring. Here are a few important points to be aware of:

(1) The participants must submit all the files required in the result compressed file, otherwise, the score of this part will be set to 0.

(2) The result should pass the validation test, otherwise the score of this part will be set to 0.

(3) Participants are required to provide comprehensive details of the installation process, the machine specification, and the optimization strategies used, which will be considered as the primary basis for scoring by the ASC24 committee.

(4) Any code that is related to the method parameters is not allowed to be modified. The region of code to be modified are restricted to:

   a) OpenCAEPoro

   b) PETSC_FIM_solver.cpp in petsc_solver

(5) All of the parameters in the input files should NOT be changed.

For any further questions, please contact techsupport@asc-events.org