Green AI - ESPCI 2024: Practical work guide

This document presents instructions and questions regarding the practical work sessions. All the materials (slides and codes) can be recovered from the following GitHub repository: https://github.com/Deyht/green_ai_espci

Start by cloning this repository and making a copy to work on:
`git clone https://github.com/Deyht/green_ai_espci`
`cp -R green_ai_espci work_dir_green_ai`

The participants must provide a report where they answer the questions and describe their observations. The participants can work in pairs and provide a single report with the two names. **The deadline for the report is Friday May 3rd at 23:00 Paris time**. The report for Part A and B can be provided as a single file or as individual files at your convenience.
All the produced codes and figures, as well as the report, must be uploaded as a single archive (.zip, .tar, ...) named after the participant's full name and practical work part (e.g., `surname_name_pw_report_partA.zip`) at the following link: https://share.obspm.fr/s/iiSEz5BDsKk2b5d

Part A: Optimization and HPC

This first part tackles the subject of high-performance computing for the matrix multiplication operation that is extensively used in all modern AI models. Our objective is to find the most efficient way of implementing and using this specific operation to improve numerical efficiency and reduce the amount of energy required.

We will implement a classical matrix multiply operation of an $M \times K$ matrix A with a $K \times N$ matrix B to obtain an $M \times N$ matrix C. The content of an element of C is given by:

$$C(i,j) = \sum_k A(i,k) \times B(k,j) \tag{1}$$

The indices i, j, and k go through M, N, and K, respectively. We provide Python scripts and C source codes that contain different implementations of this operation in `green_ai_espci/opt_matmul/`.

Python

1. In the `matmul.py` script, we provide the code that allocates the three matrices and initializes A and B to random values. The same code also contains two hand-written implementations of the matrix multiplication operation, `matmul_naive` and `matmul_numpy_sum`. This script can be opened and edited with any text reader (e.g., gedit), and must then be run from a command line terminal using: `python3 matmul.py`

   Working on the computers from the ESPCI classroom, you can load a Python environment that contains all the necessary libraries with the command: `conda activate simul`

   For this problem, allocation and initialization times are negligible, so the compute time can be measured using the command: `time python3 matmul.py`

   Use this command to evaluate the compute time (**real** time) of the script running the `matmul_naive` function for a small matrix size (e.g., $M = N = K = 512$). This will be your reference time for computing the speedup of subsequent versions. Ensure that only one compute operation is uncommented when doing compute time measurements.

   Note: *Compute time predictions can vary due to other loads on the system. Always run your compute time estimations a few times to average the variability.*

2. Change the script to execute the second implementation `matmul_numpy_sum` and evaluate the compute time for the same matrix size. What is the speedup relative to `matmul_naive`?

3. The two handwritten functions can be compiled with Numba by adding the following lines:

```
from numba import jit
@jit(nopython=True, cache=True, fastmath=False)
def matmul_naive(A, B, M, N, K):
    [...]
```

   In the provided script, the lines are present and should just be uncommented. Note that the compilation with Numba adds time at the first execution, so execute at least twice before measuring the time. Measure the new time of both functions compiled with Numba. What is their respective speedup compared to your reference time? What do you observe? Which one is the fastest?

4. To have a better representation of the computing efficiency of different implementations, we can estimate the number of floating-point operations per second (FLOPS). This can be done by dividing the total number of "useful" operations done by the total time of computation. For the matrix multiplication operation where $M = N = K$, the number of operations is simply N^3. Using one of the two Numba-compiled implementations, draw a curve representing the performance in GFLOPS as a function of the problem size from 256 to 2048 by steps of 256 and then up to 4096 but with steps of 512. What do you observe? Try explaining the shape of the curve.

5. Now try comparing the compiled handmade implementations with optimized matrix multiplication operations in Python using the `@` operator and the `matmul` and `dot` functions from Numpy. We note that these functions are likely parallelized by default, so the comparison with the handwritten version is unfair. To force the execution on a single core, use the following line in the same terminal on which you run your code: **export** `OMP_NUM_THREADS=1`. Running on a single core, the `real` and `user` time returned by the `time` command should be identical.

   Which of the optimized implementations is the fastest for a large matrix size of $M = N = K = 2048$, and what is the typical speedup compared to your best Numba-compiled implementation? Draw a new performance versus problem size curve for the fastest optimized operation up to $M = N = K = 4096$. How does it compare to the performance curve of the Numba-compiled hand-written implementation? What are your explanations for the shape of this new curve?

Optimized C implementation

We provide the `matmul.c` source code that contains six different implementations for the matrix multiplication with an increasing level of optimization (from v1 to v6). The `main` function allocates the three matrices and initializes them to random values. The matrices are flattened as 1D arrays, and we adopt the column-major indexing formalism.

Note: *For a matrix with M rows and N columns, i and j indexing the rows and columns, the* `C(i,j)` *element of matrix encoded in flattened column-major can be accessed with* `C[j*M+i]`*.*

We added timers to measure the elapsed time between two markers in the code. Encapsulating the call to a function between these markers allows us to measure the corresponding computation time. We also added a call to the optimized SGEMM matrix multiplication function from the OpenBLAS library. This allows us to verify that our custom implementations are correct regarding computed values for the C matrix, and it also provides an optimization performance goal. Like for optimized library in pythons, OpenBLAS is parallelized by default, so you must use the same export to force it to run on only one CPU core: **export** `OMP_NUM_THREADS=1`

The different implementations' performance gets progressively closer to the OpenBLAS implementation of the `sgemm` operation. The provided code can be compiled using:

```
gcc matmul.c -o matmul -lopenblas -lm
```

Do not copy past the command from the PDF as it seems to result in errors. The code must be recompiled before execution every time you change the source!

6. The `matmul_v1` function corresponds to a naive 3-loop implementation. Using $M = N = K = 1920$, what is the typical execution time of this v1? How does it compare to the time of your previous best naive Numba-compiled Python function? What are the GFLOPS for both versions at this given matrix size?

7. Try adding optimization flags to your compilation line. Measure the time and estimate the GFLOPS when compiling with `-O1`, `-O2`, and `-O3`. What do you observe?

8. The `matmul_v2` function adds an accumulator to compute the sum. Measure the computation time and the GFLOPS of this v2 for the four possible optimization flag (none, `-O1`, `-O2`, `-O3`). What do you observe regarding the effect of the optimization flags? Why does the use of an accumulator improve the performance?

9. The `matmul_v3` function transposes the matrix A to have index continuity in the k-loop following:

$$C(i,j) = \sum_k A^T(k,i) \times B(k,j) \tag{2}$$

   Compiling this function with `-O3`, what are the resulting compute time and GFLOPS, and how does it compare to the v1 and v2? Now try to add the following additional optimization codes in the compilation line `-O3 -march=native -ffast-math -funroll-loops`. What are the effects on the compute time and GFLOPS, and why? Try dropping off one option and identify the one that has the strongest effect on performance.

10. The `matmul_v4` function uses the GCC vector data structure to perform an explicit SIMD vectorization of the matrix multiplication operation. For this, it creates new vectorized versions of A and B and fills them with the corresponding data. The A matrix is still transposed (but on the fly with the conversion to the vector type) to ensure memory continuity. The sums of products required to fill C are then made using the FMA instruction by doing the operations on vector data structures. Compile the code calling this version with all the previous optimization flags. What are the effects on the compute time and GFLOPS, and how does it compare to the optimized v3? Explain your observations.

11. Produce scaling curves for the v1 to v4 versions representing the GFLOPS as a function of the matrix sizes from 48 to 1920 using only multiples of 48 (the step size is left to your appreciation, but you should have at least about ten points) and considering that $M = N = K$. For each version, use the set of optimization flags that resulted in the best performances (some flags are detrimental to the v1 and v2 versions). What do you observe? How can you explain the performance dependency with the problem size? What could be the limitation of this v4 version?

12. The `matmul_v5` and `matmul_v6` functions rely on the same kernel function. This kernel works at the scale of the CPU registers and relies on the fact that the same data are required for several operations in a matrix multiplication operation. By storing a chunk of data as registers and reusing them as much as possible before loading new data, this kernel should maximize memory throughput and start to reach compute limitations of the CPU. The kernel works on vector data structure registers so it can use vectorized FMA operations. The details of this kernel's inner workings are given in the course. The `matmul_v5` version simply decomposes the problem in chunks that have the size of the kernel and calls it for all the possible K at a given kernel location in C. Using the default kernel configuration and compiling the code with all the previous optimization enabled, what are the computation time and GFLOPs of this v5 for $M = N = K = 1920$?

13. The optimal size of the kernel is dictated by the available number of CPU vector registers. Try modifying the size of the kernel and search if there is a better configuration than the default one. If yes, provide your best configuration and explain why it is better than the default one. Do this search twice, first for a large problem size (e.g., $M = N = K = 1920$) and then for a small problem size ($M = N = K = 512$). What do you observe? Is the optimal configuration different?

14. Draw the performance versus problem size curve in the same way as before and compare it to the curves obtained for the previous versions. What do you observe, and what is still limiting this v5 implementation?

15. Try to invert the order of the i-loop and j-loop in `matmul_v5` and measure the compute time and GFLOPS. Do you observe an effect, and if yes, in what direction and why? Is this observation in agreement with the limiting point you identified in the previous question?

16. The `matmul_v6` function also uses the vectorized register kernel, but this time, it also tries to reuse the data stored in the different L-cache levels of the CPU. For this, it defines blocks in A and B with a size that is a multiple of the kernel size and calls the kernel for each possible sub-region. Doing so implies that a single call to the kernel is insufficient to obtain the final value of a cell in C and that the contribution must be accumulated over all the possible block positions in the K dimension. After compiling with all the optimization options, estimate the compute time and the GFLOPS of this last version for $M = N = K = 1920$. What is the remaining relative performance difference between this version on the reference OpenBLAS version? Draw the performance versus problem size curve for the v6 function and OpenBLAS. What do you observe? (be sure to do the required export so OpenBLAS runs only on one CPU core).

OpenMP parallelization

We provide an OpenMP parallelized version of the previous C code in `matmul_para.c`. This new version can be compiled with OpenMP support by simply adding `-fopenmp` to your compilation line. The number of threads on which the code will be executed is then controlled by the environmental variable `OMP_NUM_THREADS`. You can adjust this value and re-run a code without recompiling as it is queried at execution time. You can set this variable to a value of X by using the command: **export** `OMP_NUM_THREADS=X`

17. Our objective is to evaluate how well our custom implementation scales with the number of OpenMP Threads. For a fixed problem size of $M = N = K = 1920$, evaluate the compute time and GFLOPS for the non-parallel version and the parallel version but with only one thread. What do you observe and why?

18. Still using a fixed problem size, draw a curve representing the achieved speedup as a function of the number of OpenMP threads ranging from 1 to 16. The speedup for a code running with N threads is defined as the compute time for 1 thread divided by the compute time for N threads.

19. To analyze the previous curve, you must first identify the physical properties of the CPU in your system, which can be done using the `lscpu` command line. From this, identify the number of physical cores and logical threads in your system and indicate them in your report. With this additional information, describe and explain the shape of the previous speedup curve.

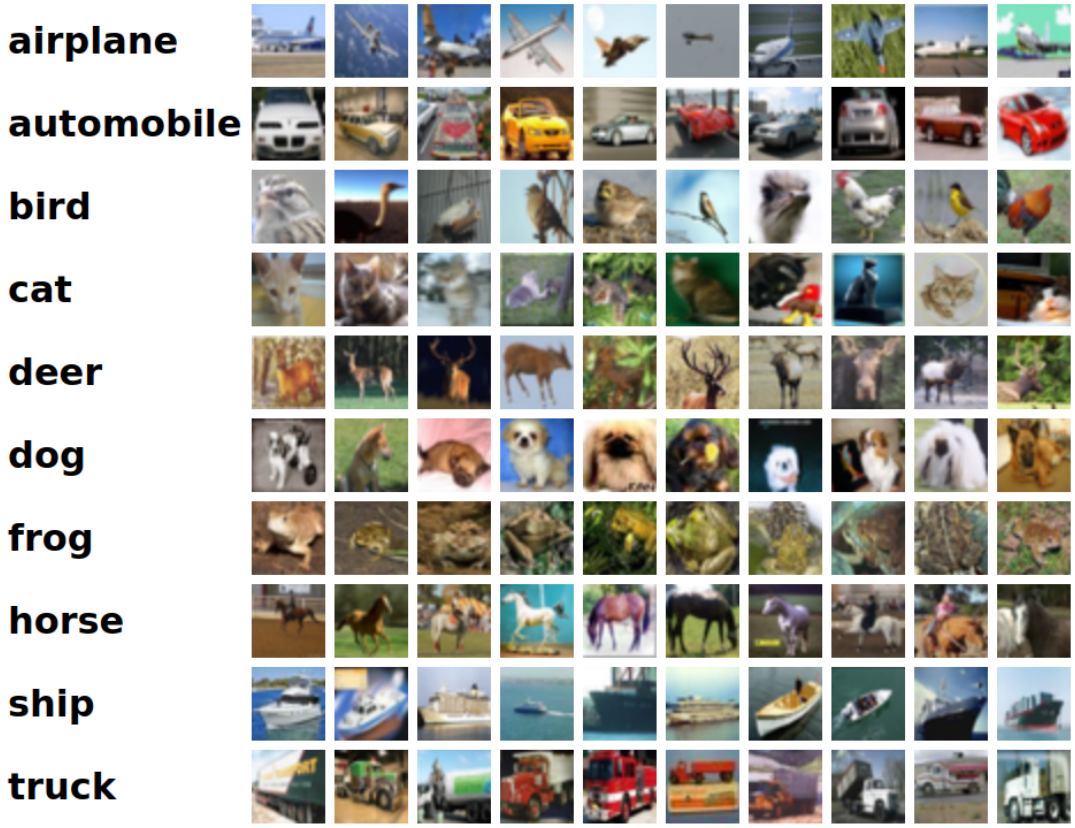
20. Using multiple CPU cores at the same time on a given problem will both lower the computation time and increase the power draw. However, due to the mutualization of several parts of the chip, the increase in power draw induced by the involvement of additional cores is usually not linear and much lower than the power draw of the first core. If we consider that using all P physical cores of the CPU induces a doubling of the power draw, what would be the ratio between the energy consumed for the computation using 1 thread and P thread for the parallelized version of the `matmul_v6`? Estimate the same ratio for the parallelized OpenBLAS SGEMM function. What do you observe?

    Note: *The total energy consumed by a given computation can be approximated by $E = \Delta P \times T$, with E the energy in Joules, ΔP the increase in power draw compared to the system baseline in Watts, and T the total time of the computation.*

Part B: CNN efficiency-based optimization on GPU

This last part tackles the subject of optimizing a Convolutional Neural Network model for a metric that combines model accuracy, numerical efficiency, and model size.

For this, we will use the CIFAR-10 dataset, which comprises 60000 images of 32x32 pixels labeled into 10 classes. 50000 images are used to train supervised learning models, with 5000 examples for each class, and 10000 images are used for testing trained models, with 1000 examples for each class.



Your objective is to explore network structures following the guidelines from the lecture to build a classification model that maximizes the following score metric

$$S = \left(\frac{E_r}{E}\right)^{w_E} \times \left(\frac{T_r}{T}\right)^{w_T} \times \left(\frac{P_r}{P}\right)^{w_P} \tag{3}$$

where E is the classification top-1 error rate (one minus the global accuracy), T is the compute time, and P is the number of trained parameters of the model. The r indexed values represent the same quantities for a simple reference model for which the score result is set to 1.0. The three terms represent the relative errors to this reference model, and the contribution of each part to the total score is weighted by the w_E, w_T, and w_P powers.

Our reference model is an architecture close to a LeNET-5 with small adjustments. The reference values for E_r, T_r, and P_r are provided with the scripts, and we set $w_E = 1$, $w_T = 0.95$, and $w_P = 0.1$. With these scaling factors, a model that improves the error rate by a factor of two while having the same compute time and number of parameters as the reference model will get a score of 2.0. It goes almost the same way for improving the compute time by a factor of two while preserving the error rate. The number of parameters in the model has a smaller effect and is only here to prevent the use of highly parametric architectures.

Fully functional training and inference Python scripts are provided in `green_ai_espci/opt_cnn/`. We recommend using the Google Colab notebook version of the scripts and running them using a T4 GPU, which was used to define the baseline compute time for our reference architecture. You can use other systems to train your model, but inference time should be measured in Colab with a T4 for a fair comparison. In the `aux_fct.py` script, we provide several utility functions to download, load, visualize, and dynamically augment the images. **You might need to edit small parts of this file in Colab** to change the input image size of the model (default 32x32) or the augmentation policy. **Note that changes made to a file in Colab are not saved by default! Work with a copy of the files in your Google Drive or keep a trace of your modifications. Also, all files and progress are lost every time the Colab session is reset, so save and download the files, results, and network models regularly.** With the free version of Colab, the daily computing time is limited, so consider using multiple Google accounts. The saved models are available in the `net_save/` repository that is automatically created when starting a network training. The default naming scheme only refers to the training iteration, so rename your saving files with comprehensive information about your model to keep a trace of your progress. A saved model can be reloaded for inference or further training, but you must upload it back to Colab first.

We provide training and prediction scripts that use the CIANNA framework to construct the network model. You should be able to adapt the model's architecture with only minor script adjustments, as described in the lecture slides. You can refer to CIANNA's WIKI page for a complete framework description. You can also look at the full API documentation to add layer types that are absent from the LeNET-5 example. You are also free to replace these parts with any other Deep Learning development framework you are familiar with (TensorFlow, PyTorch, etc.) as long as you measure the inference time using a T4 GPU. **Note that, for this exercise, it is forbidden to use external data for training or to use a pre-trained model on other data as a starting point.** Once you have a working model, you can enter its inference properties in the shared Google Sheet used as a dynamical leaderboard. You can tackle this part in pairs or by forming a team (6 people per team maximum). Still, each pair must provide an individual report for this part. **The top three scoring teams on the leaderboard at the report submission deadline will grant their team members an extra bonus point in their grade. The achieved result has to be reproducible to be eligible (in both inference and training).**

We expect you to produce results for at least 3 different architectures or image augmentation strategies. Each model must differ significantly from the others (Note that adding 1 single layer to the network does not count as a significantly different model). In the report, you will summarize your architecture research strategy, summarize your observations regarding the accuracy of the tested models, and provide possible explanations for these behaviors. Do not hesitate to provide detailed and technical information. You must provide all the modified codes (notebook and `aux_fct.py` files) and at least one saving file for each model you describe in your report so your inference results can be reproduced. These files must be provided following the procedure explained at the beginning of the document before the report deadline.