BENCHMARKING THE PERFORMANCE AND ENERGY CONSUMPTION OF THE AVX512 AND VNNI INSTRUCTION SETS

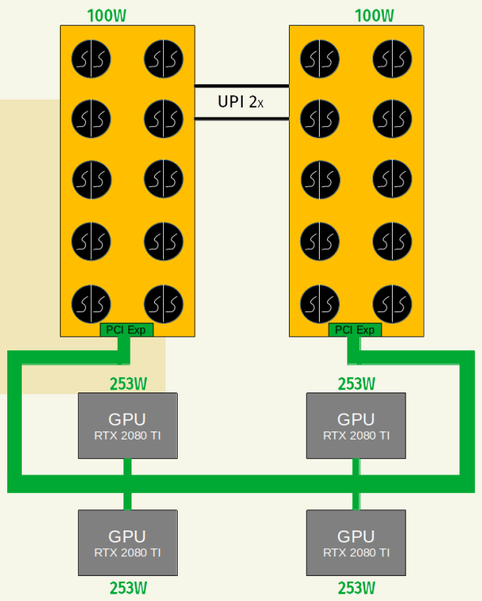
End of Degree Project

Jon Arriaran Cancho
Jose Antonio Pascual
Computer Science - Degree in Computer Science

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

The birth of this project was inspired by the most recent Intel Xeon Cascade Lake series processors, which were released with the possibility of executing VNNI instructions applying the already available AVX-512 instruction set. VNNI instruction set could be executed only on GPUs until nowadays, so the performance and efficiency these instructions could reach on a processor, it is, at least, something unknown and worth studying.

The main goal of the project is to execute a self-created specifically aimed program. Using these VNNI instructions with AVX-512 set on many ways, it could be possible the evaluation and comparison of the performance and energetic efficiency they could obtain. Once the base evaluations were done, the idea is to continue evaluating their performance with another third parties programs such as RAPL and Singularity, for instance, complementing with those programs the previously made evaluations. Finally, better and more complete conclusions of the power consumption, execution time and frequency performance of the VNNI instruction set will be drawn.



Physical Description of a Node from the Priscilla Server (Where evaluations will be done)

1 - ZAGREUS

This benchmark executes the next VNNI instructions:

- _mm512_dpbusd_epi32(src, a, b)
- _mm512_dpbusds_epi32(src, a, b)
- _mm512_dpwssd_epi32(src, a, b)
- _mm512_dpwssds_epi32(src, a, b)

VNNI

2 - GET_RESULTS

The program based on Python that get results and generate charts from them.

.OUT → .RES → MATPLOTLIB → CHARTS

3 - RAPL

This technology measures the power and energy consumption of the benchmark execution.

POWER

RAPL

4 - SINGULARITY

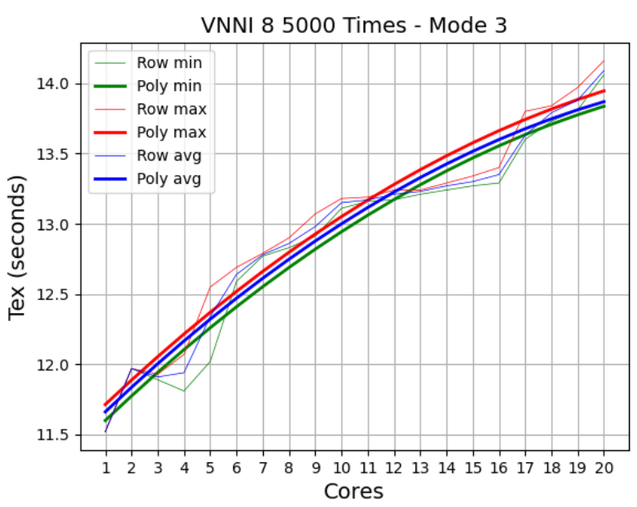
Execute Zagreus benchmark inside Singularity container and measure the performance.

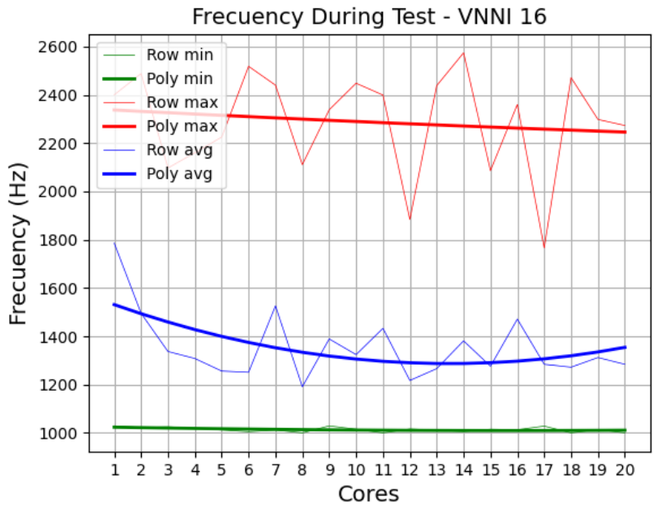
.OUT

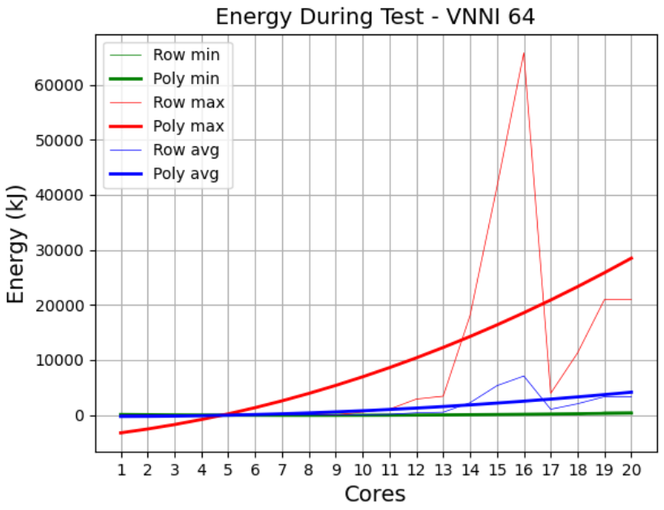
ZAGREUS

AI

RESULTS







CONCLUSIONS

It was discovered that the different amount of cores using on the execution really impact in their performance. These configurations being between the 15 and 17 cores are the worst performing ones.

Furthermore, it is also observed that using Singularity containers to execute the benchmark alters its efficiency. Yet, in this case, it is worth using this Singularity technology due to the advantages it offers.