08-07-2022

Jon Arriaran Cancho

END OF DEGREE PROJECT

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

INDEX

INTRODUCTION

INTRODUCTION

Intel Xeon Cascade Lake series

INTRODUCTION

Intel Xeon Cascade Lake series

VNNI

INTRODUCTION

Intel Xeon Cascade Lake series

VNNI

- _mm512_dpbusd_epi32()
- _mm512_dpbusds_epi32()
- _mm512_dpwssd_epi32()
- _mm512_dpwssds_epi32()

INTRODUCTION

Intel Xeon Cascade Lake series

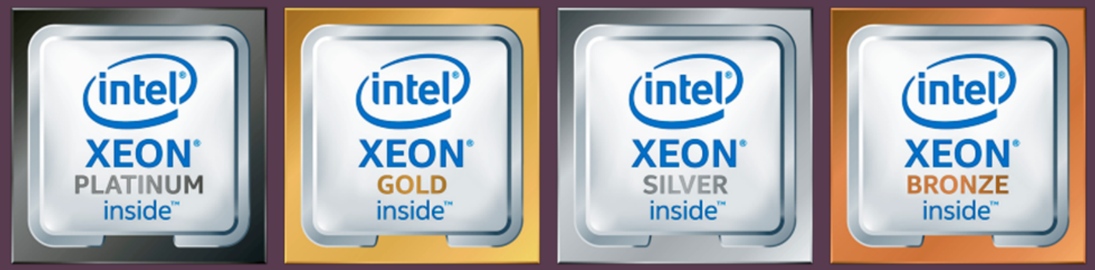
VNNI

- _mm512_dpbusd_epi32()
- _mm512_dpbusds_epi32()
- _mm512_dpwssd_epi32()
- _mm512_dpwssds_epi32()

```
FOR j := 0 to 15
        tmp1.word := Signed(ZeroExtend16(a.byte[4*j]) * SignExtend16(b.byte[4*j]))
        tmp2.word := Signed(ZeroExtend16(a.byte[4*j+1]) * SignExtend16(b.byte[4*j+1]))
        tmp3.word := Signed(ZeroExtend16(a.byte[4*j+2]) * SignExtend16(b.byte[4*j+2]))
        tmp4.word := Signed(ZeroExtend16(a.byte[4*j+3]) * SignExtend16(b.byte[4*j+3]))
        dst.dword[j] := src.dword[j] + tmp1 + tmp2 + tmp3 + tmp4
ENDFOR
dst[MAX:512] := 0
```

INTRODUCTION

Intel Xeon Cascade Lake series



INTRODUCTION

Intel Xeon Cascade Lake series

- Execution Time
- Energy
- Power
- Frequency

THE AIMS OF THE PROJECT

AIM NUM.1

Create the benchmark for executing VNNI and AVX512 instructions

AIM NUM.2

Add to the benchmark the energy measure and analysis of the executions

AIM NUM.3

Analysis of the performance of executing the benchmark inside Singularity container

TASKS TO BE PERFORMED

1

Study the new set of instructions

2

Create the set of benchmark for executing VNNI and AVX512 instructions

3

Execute the benchmarks in an HPC cluster with Cascade Lake processors

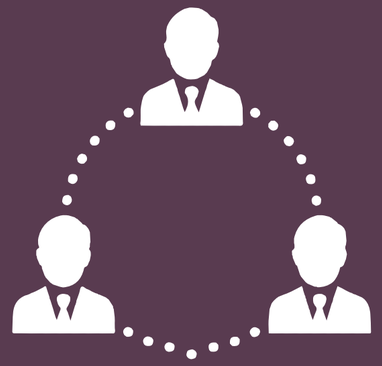
4

Evaluate the architecture using the benchmarks inside Singularity

5

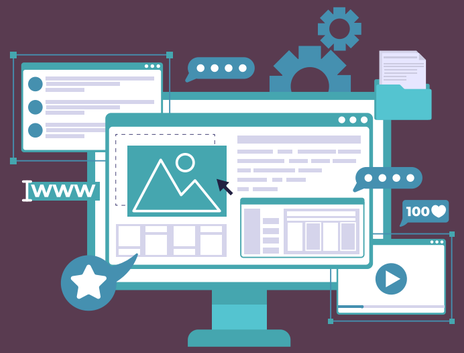
Analyse the results (performance, energy consumption)

PLANNING



Management phase

- Time estimations
- Evolution analysis
- Risks
- Deviations
- Little milestones



Development phase

- Create Programs
- Execution of the programs
- Issues
- Get results

Documentation

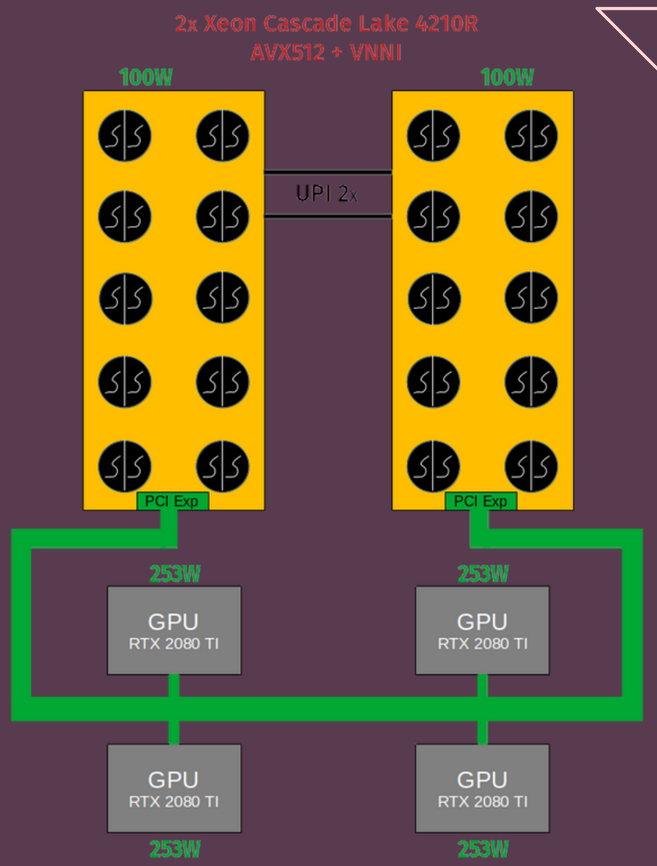
- Reports
- Final conclusions
- Analysis of the results
- Explanations

PRELIMINARIES

PRELIMINARIES

Project Deployment Cluster:

- Nodes 50-53
- Node 150



PRELIMINARIES

SLURM WORKLOAD MANAGER

Open source cluster management and job scheduling system for large and small Linux clusters.

srun, scancel, sinfo, squeue...

RAPL

Energy measurement technology aimed to read and measure the energy consumption of any execution on Linux. Read results from Linux kernel.

SINGULARITY

A container solution created for scientific and application driven workloads.

ZAGREUS BENCHMARK

ZAGREUS BENCHMARK

PARAMETERS

- mode
- command_num
- m512_size → _m512 number type
- execution_mode

FUNCTIONALITY

Program is launched on the cluster by Slurm and by the behaviour selected with the parameters and it executes randomly initializated instructions taking their execution time and frequency, saving them into a output file.

ISSUES

- More people on the cluster
- Queue waiting
- Resources

ENERGY ANALYSIS

ENERGY ANALYSIS

Adding RAPL to Zagreus

- More parameters
  - Power (kW)
  - Energy (J)
- Slurm execution way change
- More complete output results

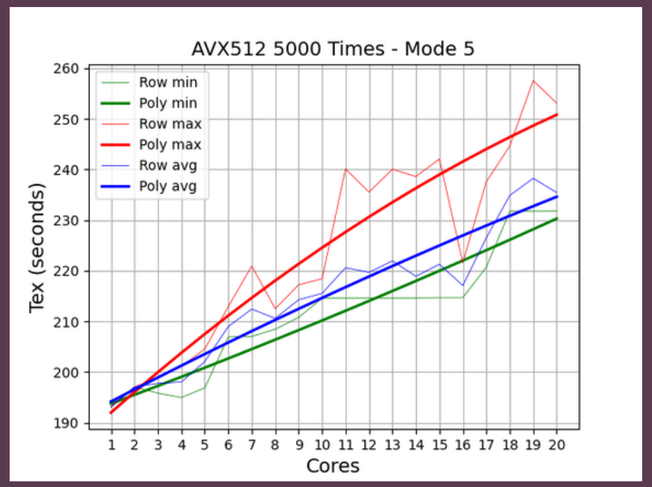
MIN

MAX

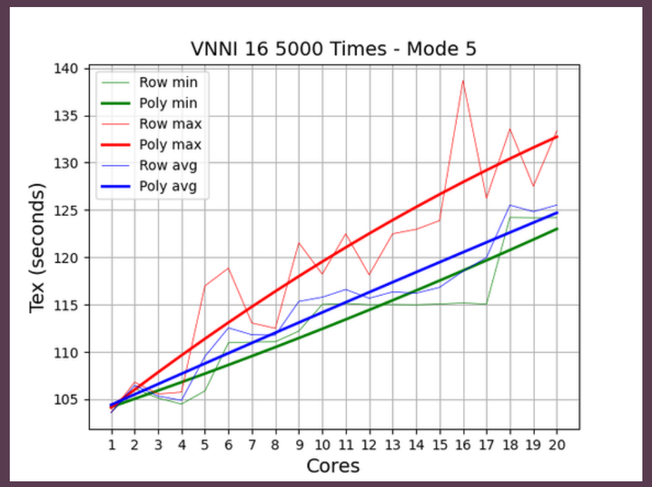
AVG

ANALYSIS OF THE RESULTS

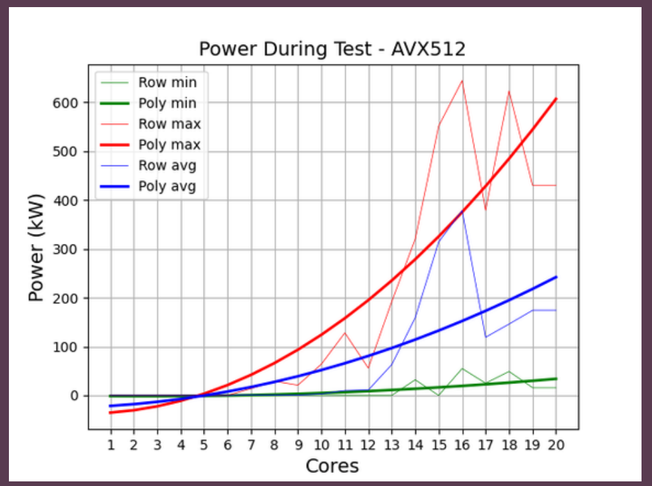
ANALYSIS OF THE RESULTS



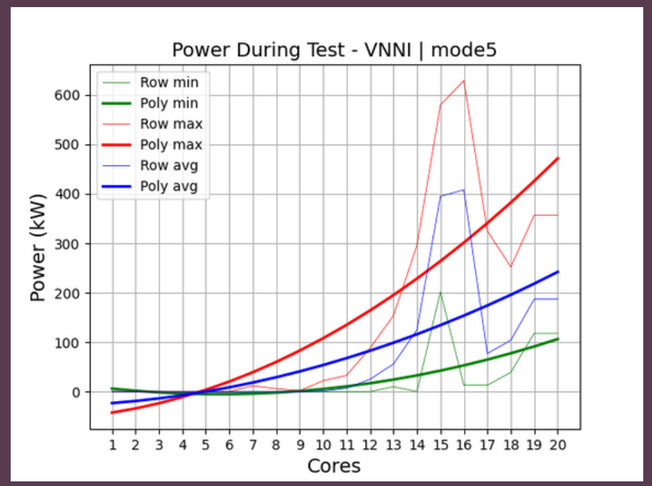
AVX512 VS VNNI



AVX512



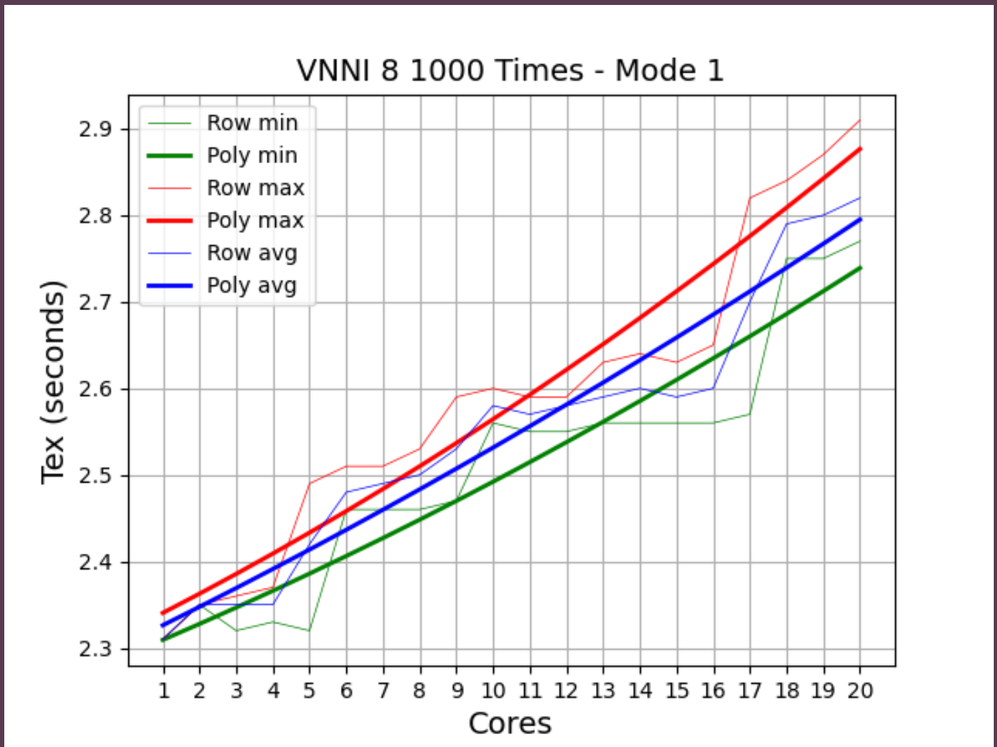
VNNI



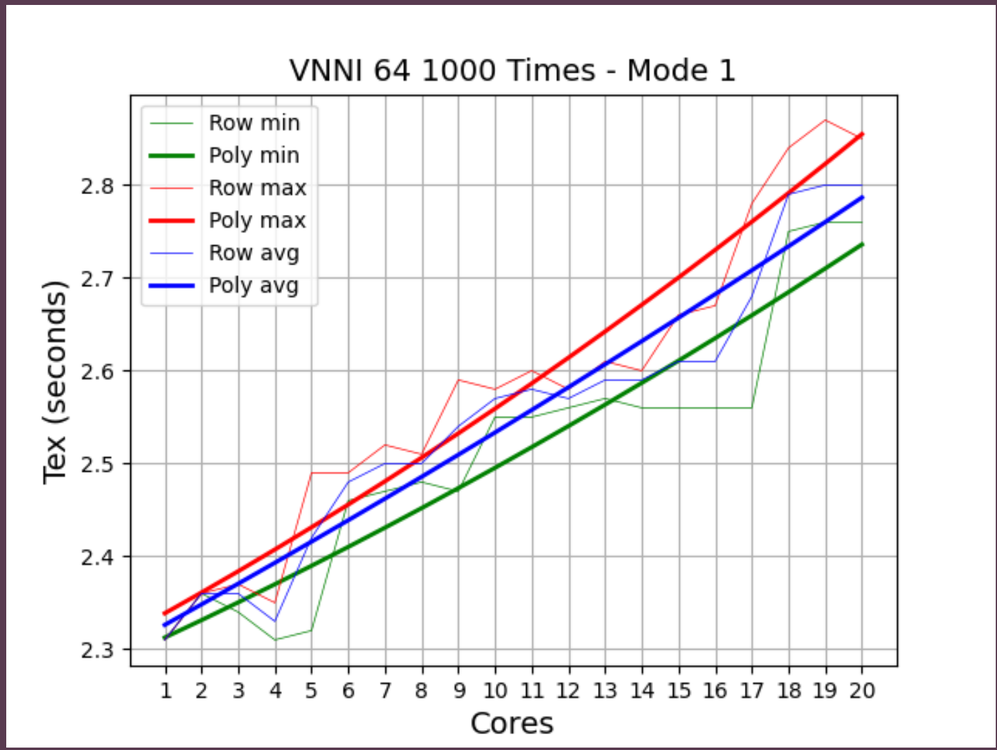
ANALYSIS OF THE RESULTS

VNNI SIZE CONFIGURATIONS

SIZE 8



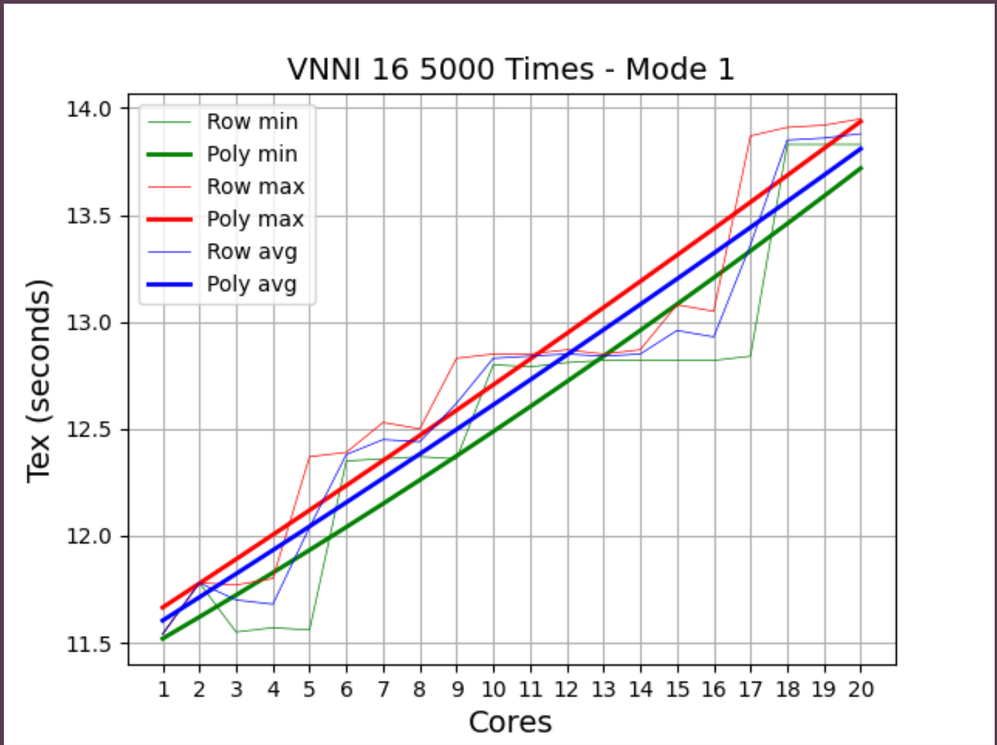
SIZE64



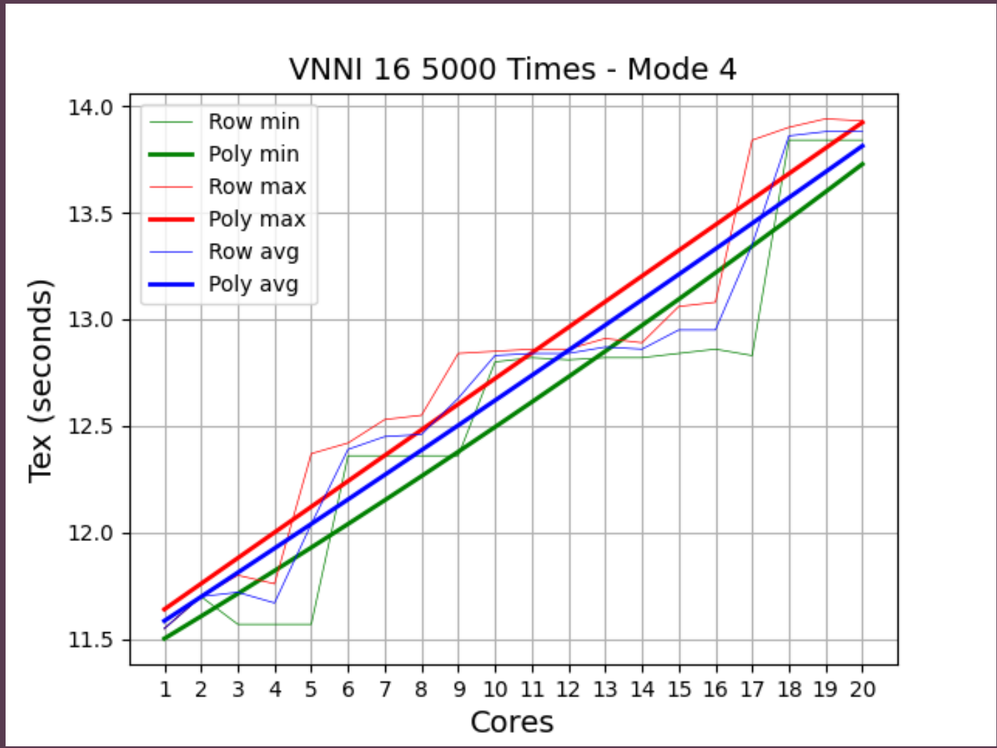
ANALYSIS OF THE RESULTS

VNNI EXECUTION MODES ANALYSIS

MODE 1



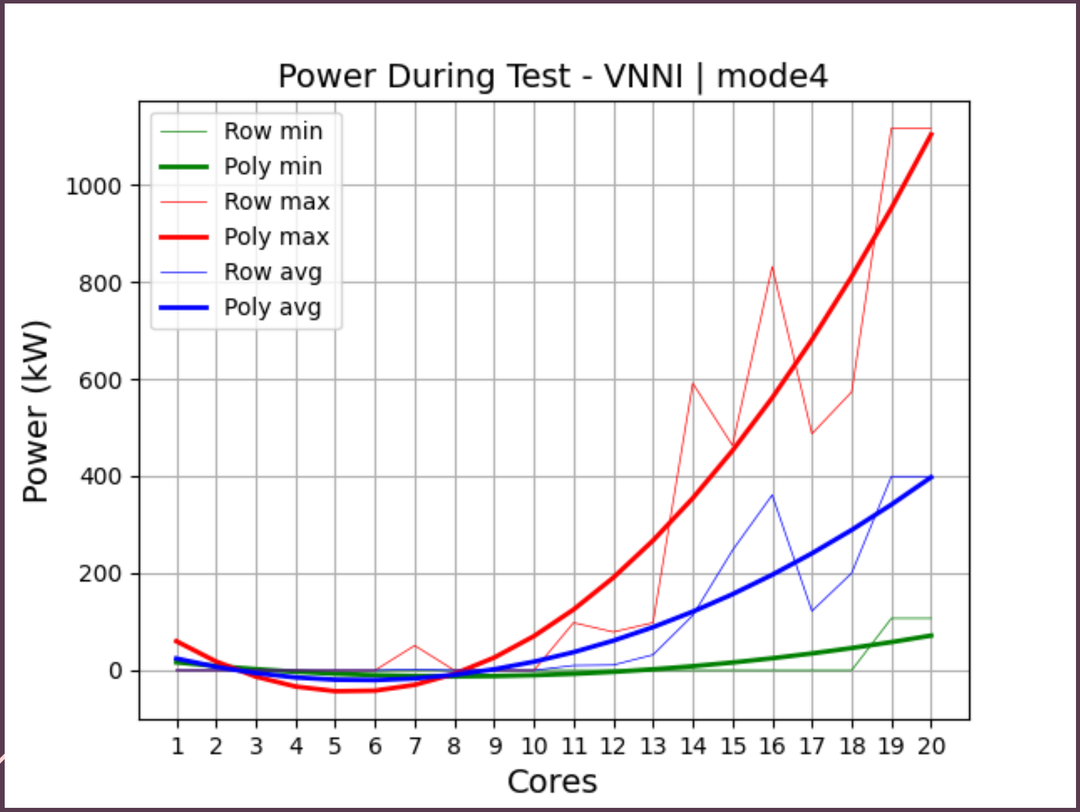
MODE 4



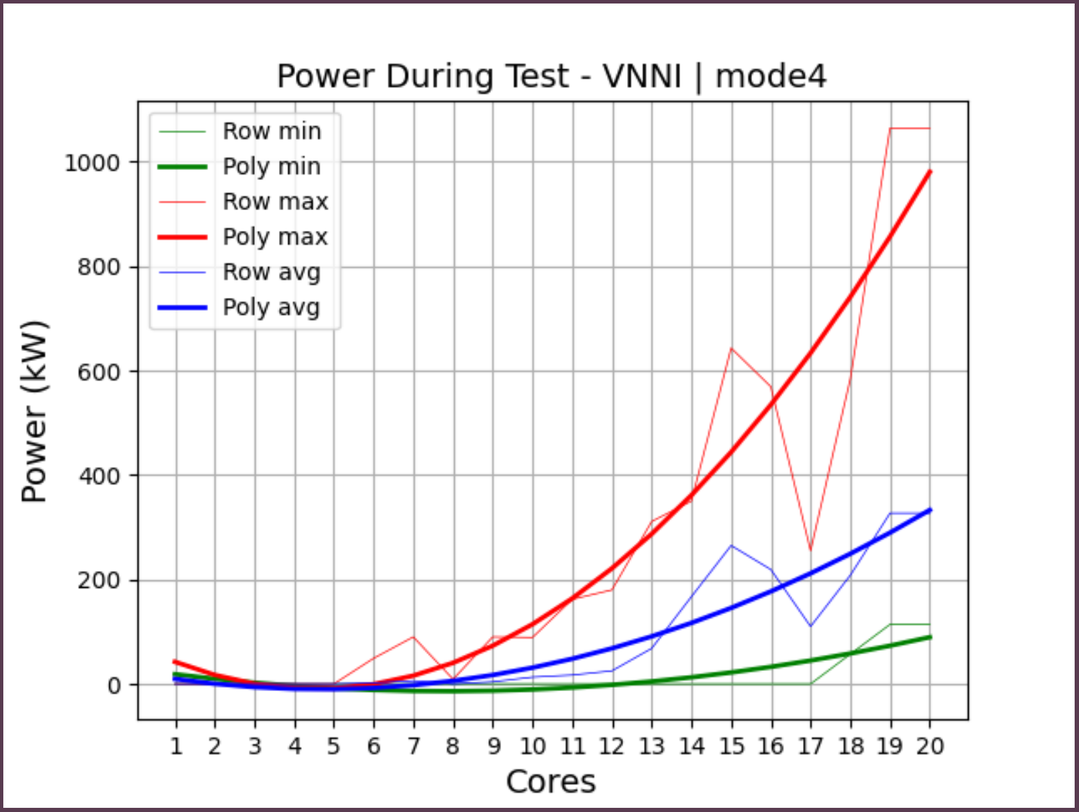
ANALYSIS OF THE RESULTS

ZAGREUS INSIDE SINGULARITY

NO SINGULARITY



SINGULARITY

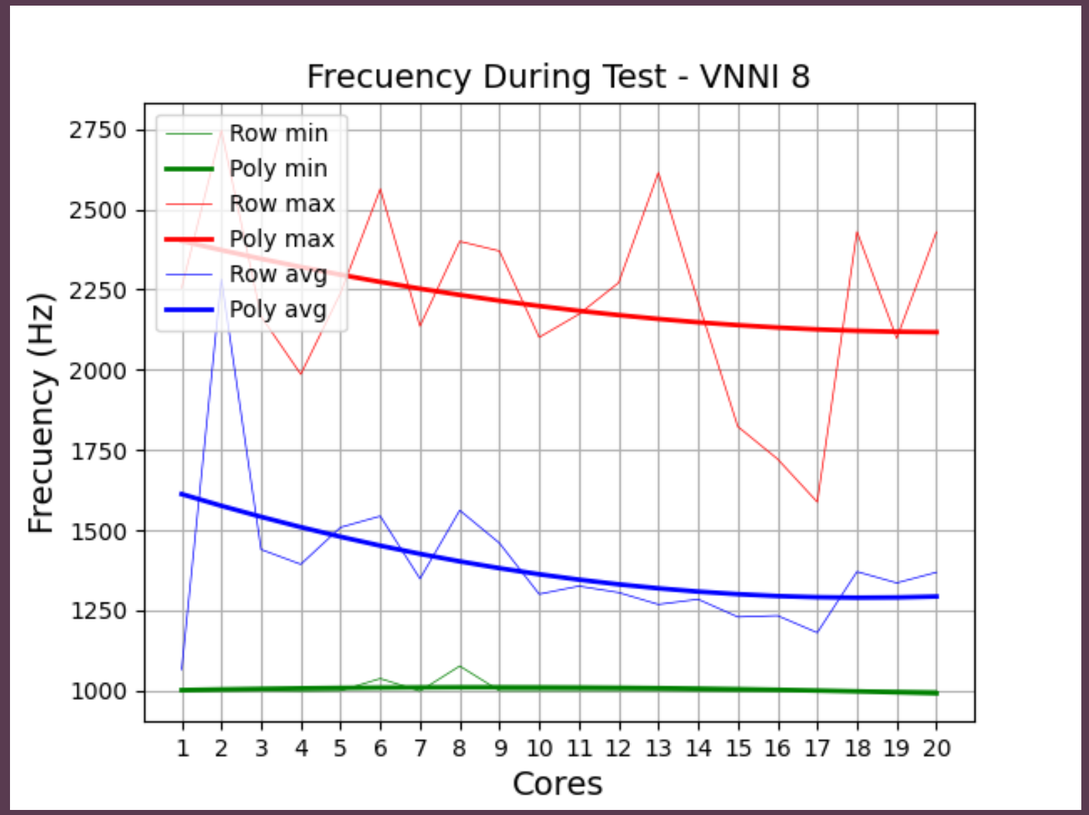


FINAL CONCLUSION

FINAL CONCLUSION

- VNNI instructions have better performance than AVX512 ones.
- The size of filling _m512 numbers is not relevant
- VNNI instructions are equal in performance

- The configurations that use 14 to 17 cores are the worst performing ones
- Singularity usage alters by a 8% the efficency of the benchmark



FUTURE WORK

FUTURE WORK

CPU VNNI performance

vs.

GPU same behaviour program

The deep analysis of the CPU manages communications done between all the cores

Thank You!

Any questions?

Jon Arriaran Cancho