LLMMMM: Large Language Models Matrix-Matrix Multiplications Characterization on Open Silicon

Louis Ledoux*†, Marc Casas*†
*Barcelona Supercomputing Center, Barcelona, Spain
†Universitat Politècnica de Catalunya, Barcelona, Spain
E-mail: {louis.ledoux,marc.casas}@bsc.es

Keywords—Large Language Models (LLM), Transformers, Generative Pre-Trained (GPT), Matrix-Matrix Multiplications, Floating-Points, arithmetic, ASIC, Open-Source Silicon (OSS)

I. Extended Abstract

A. Introduction

GPT transformers are useful for various applications, offering significant advancements in natural language processing tasks.

However, their operational costs are substantial. Prior work has highlighted the financial implications of deploying these models. [1]

Essentially, matrix-matrix multiplications (MMM), with their intensive data movement and manipulation of arithmetic weights, underscore the computational demands of these architectures. This involves handling vast matrices and managing numerous elements within these computational structures.

Recent efforts in research have concentrated on devising specialized formats and algorithms aimed at mitigating these costs. These innovations include reducing bit-width, optimizing matrix multiplication techniques (MLX), and leveraging specialized hardware such as TPUs and BitNets, along with adopting ternary networks and eliminating up to 40% of layers without significant loss in performance. []

We introduce an innovative approach by presenting a generator of ASIC kernels that remains agnostic to the Process Design Kit (PDK) specifically for MMM units. This approach caters to emerging and unconventional small floating point formats. We further undertake a comprehensive evaluation of these units, demonstrating their efficacy and potential. [] Explicetely say the list of the contributions (we createall the arithmetic, the matmult unit, high level descrption of all of that to put in tsuf framework, evaluation on placed androtue asix (gds))

B. MMM kernels generation

It is clear that the importance budgets implied in training LLMs is no afordable in a age where known computer science laws are being rstes such moore law, dennard, dark silicon era.

Which explains the heavy need for specialization, specialized cirtuits to explore the emerging float format in the cotext of mmm units. which is possible with the fast evolution of open source EDA tools [**?**], [**?**].

Figure **??** puts together the distribution of the weights dynamic range alongside with the computer format we evaluate. The evaluations comprise the computer formats of the figure with different Systolic Array sizes with five PDKs (GF180, SKy130hd, Sky130hs, nangate45, ASAP7) which allow to verify scalability of designs without the use of manual scaling.

2 variations of accum per arith, very important as computing is less expensive than communictaion, we can add some complexity in the chip to leverage data movemenrt by taking int oaccount he sequential nature of mmm and dot products maximizing entropy.

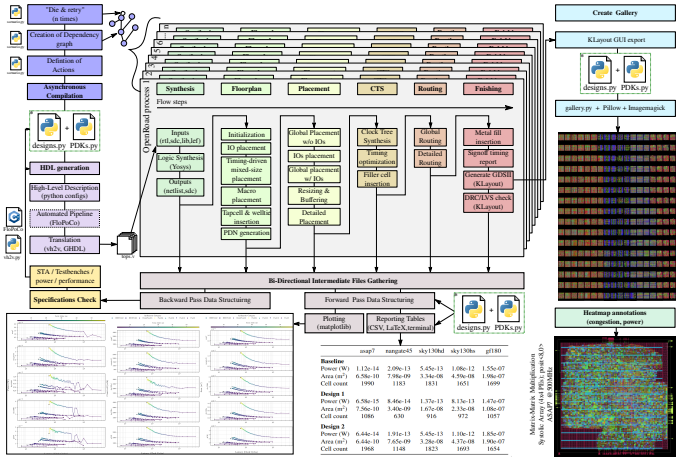

Fig. 1. Schematic Overview SUF: Centralized Management of Asynchronous OpenROAD Forks, Derived from Dependency Task Graphs. This illustration also encapsulates the extended capabilities ranging from Code Generation without manual RTL Writing to Advanced Plotting and Visualization Features.

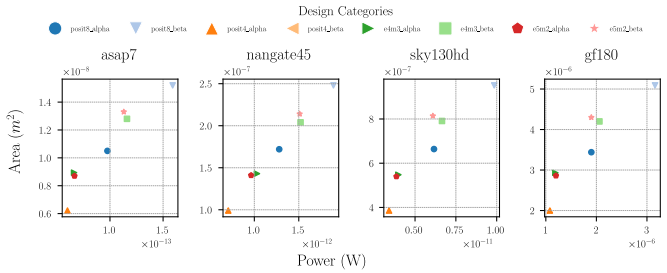

Fig. 4. Area vs. Power

C. Chips

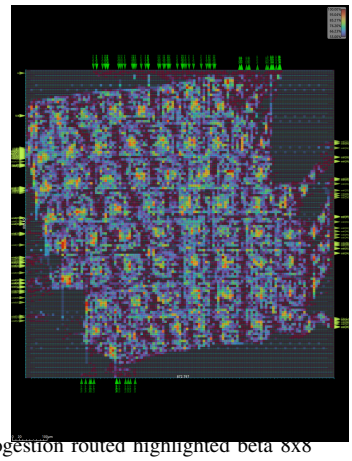

Fig. 2. e4m3 cogestion routed highlighted beta 8x8

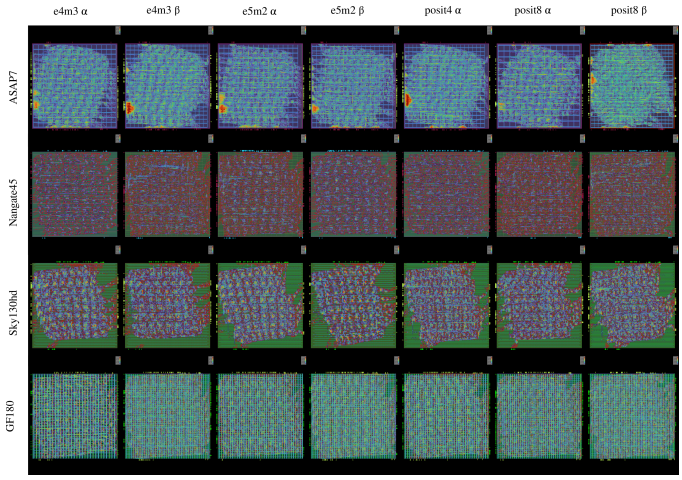

Fig. 3. All generated arrays

D. Early Results

The evaluation of 14 design entries, with posit 4 beta not working finished in two hours yielding a total of xx chips. Figure 4 shows depicts power and area accross 5 process nodes.

E. Conclusions

F. Biography

Louis Ledoux, originating from a comprehensive computer science background in Rennes (Bretagne, France), has transitioned towards a hardware focus. His journey began with a Bachelor's degree, followed by a Master's in Computer Science, culminating in a one-year internship in 2017, where he explored FPGA virtualization in the cloud. Since 2018, Louis has been engaged in a PhD in computer arithmetic at the Universitat Politècnica de Catalunya and the Barcelona Supercomputing Center, in Barcelona, Spain. His main focus are hardware implementations to address numerical requirements sparsity in HPC workloads. Beyond academia, Louis contributes to the open hardware community, participating in efabless/skywater/Google tapeouts.

II. Acknowledgment

Marc Casas has been partially supported by the Grant RYC-2017-23269 funded by MCIN/AEI/10.13039/501100011033 and by ESF Investing in your future. This research was supported by grant PID2019-107255GB-C21 funded by MCIN/AEI/ 10.13039/501100011033. Els autors agraeixen el suport del Departament de Recerca i Universitats de la Generalitat de Catalunya al Grup de Recerca "Performance understanding, analysis, and simulation/emulation of novel architectures" (Codi: 2021 SGR 00865).

References

[1] A. S. Luccioni, S. Viguier, and A.-L. Ligozat, "Estimating the carbon footprint of bloom, a 176b parameter language model," 2022.