15 Nov 2024

# Semiconductor Yield Improvement by Enabling Use of Faulty Chips for LLM Inference

Roberto Lupi

# Semiconductor Yield Improvement by Enabling Use of Faulty Chips for LLM Inference

## ABSTRACT

This disclosure describes techniques that enable the use of faulty semiconductor chips to achieve reliable large language model (LLM) performance. The output of a faulty chip is subjected to a chip-specific transformation that corrects errors. Error-correcting transformations are applied at a relatively large scale, e.g., for an entire LLM layer rather than at the level of a single matrix multiplication block. The use of unreliable components to achieve reliable computing reduces the rejection rate, and correspondingly increases the yield of semiconductor manufacturing. With the scale at which error correction is applied, costly die-size and energy consumption redundancies are reduced or eliminated.

## KEYWORDS

- Large language model (LLM)
- Matrix multiplication
- Matmul block
- Matmul unit
- Error correction
- Silent data corruption (SDC)
- Chip die density
- Semiconductor yield
- Fault-tolerant computing
- 3-level logic
- Fault characterization

## BACKGROUND

Large language models (LLMs) require substantial computational resources. The energy consumption and cost of operation of LLMs escalate with their size and complexity. Although simpler LLMs, based, for example, on 3-level logic, elementary operations, pruning low-value interconnections, etc. [1, 2], have been introduced to reduce energy demands, ultimately, even such LLMs are constrained by chip die size and energy consumption.

Silent data corruption (SDC) refers to computing errors that originate from hardware defects. For example, an impacted central processing unit (CPU) may miscalculate 1+1 as 3. SDC is mitigated using a software layer that systematically checks for errors. As transistor densities increase, SDC becomes a significant and expensive issue to resolve.

## DESCRIPTION

This disclosure describes techniques that enable the use of faulty semiconductor chips to achieve reliable LLM performance by subjecting the output of the faulty chip to a chip-specific transformation that corrects or balances errors. The error-correcting transformations are applied at a relatively large scale, e.g., for an entire LLM layer rather than at the level of a single matrix multiplication (matmul) block. The use of unreliable components to achieve reliable computing reduces the rejection rate, and correspondingly, increases the yield of semiconductor manufacturing. With the scale at which error correction is applied, costly die-size and energy consumption redundancies are reduced or eliminated.

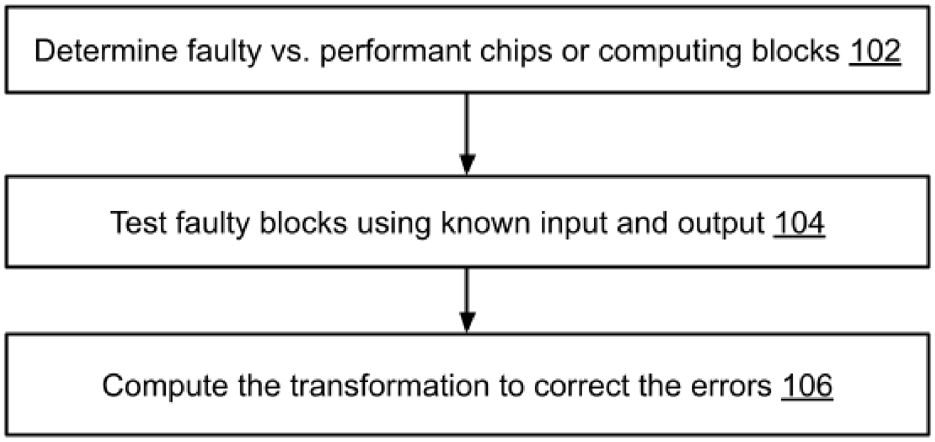
Determine faulty vs. performant chips or computing blocks 102
Test faulty blocks using known input and output 104
Compute the transformation to correct the errors 106

**Fig. 1: Reliable LLM performance using faulty chips**

Fig. 1 illustrates a technique to achieve reliable LLM performance using faulty chips. A given batch of computing chips or componentry is subjected to tests to determine faulty and performant chips, blocks, or components (102). Rather than rejecting faulty chips outright (as is conventional), chips that cross a relatively low threshold of reliability are accepted. For example, a chip with a few high-quality matmul blocks and several error-prone blocks can be accepted. By accepting chips with a greater fraction of unreliable components, semiconductor manufacturing yield improves, and redundancies that cost die size and consume energy are reduced or eliminated.

The faulty blocks are tested (104) using specific patterns of inputs with known outputs. (e.g., Hadamard matrices). Variation from the expected output is considered an error, and a transformation is computed (106) in the software layer to correct the error, e.g., to route around the hardware fault. Error correction can be achieved by using forward-forward techniques [3]. Although forward-forward can be vulnerable to skewed inputs, the input patterns can be controlled to be balanced. With an appropriate encoding, e.g., three-state error correction, positive and negative samples are passed through the same pathways on the chip, such that they statistically experience the same faults on large volumes of data.

Fault-specific error correction using software-based transformations, as described herein, can be done at a relatively large scale. For example, errors can be corrected for an entire LLM layer instead of a single matmul block. Corrections can be applied at smaller-sized blocks, following which corrections are trained at larger-sized blocks.

Training minimizes a composite multi-objective loss function based on reducing corrections at the smaller-sized blocks to generate the correct output. The training procedure recursively proceeds to larger-sized blocks. In a cloud environment, recursion can proceed beyond a single chip to rack-level components, such that a data center can reliably operate an LLM with a combination of performant and faulty accelerators. The techniques can be applied to enable a higher component density that is capable of operating under diverse operating conditions, e.g., power, thermal, etc.

Error correction can be done for a specific LLM in a warm-up phase. In a cloud environment, the initial warm-up cost can be insubstantial compared to the savings. Alternatively, results can be cached once obtained. In a device with a static LLM, error correction can be done once at the factory. Alternatively, a fingerprint can be computed at the factory, and, at the cost of a minor on-device computation, an adjustment can be pre-compiled for the LLM model.

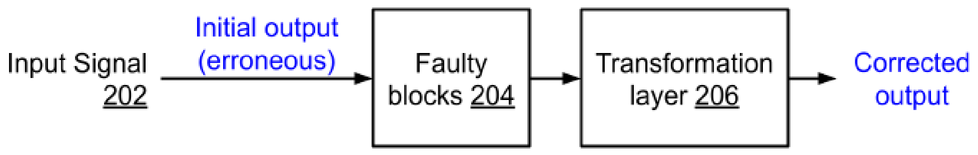

**Fig. 2: Error correction during LLM operation**

Fig. 2 illustrates an example of error correction during LLM operation. When a faulty block (204) receives an input signal (202), its initial output is erroneous. A fault-specific, software-based transformation (206) is utilized to correct the output.

## CONCLUSION

This disclosure describes techniques that enable the use of faulty semiconductor chips to achieve reliable large language model (LLM) performance. The output of a faulty chip is subjected to a chip-specific transformation that corrects errors. Error-correcting transformations are applied at a relatively large scale, e.g., for an entire LLM layer rather than at the level of a single matrix multiplication block. The use of unreliable components to achieve reliable computing reduces the rejection rate, and correspondingly increases the yield of semiconductor manufacturing. With the scale at which error correction is applied, costly die-size and energy consumption redundancies are reduced or eliminated.

REFERENCES

1. Wang, Hongyu, Shuming Ma, Li Dong, Shaohan Huang, Huaijie Wang, Lingxiao Ma, Fan Yang, Ruiping Wang, Yi Wu, and Furu Wei. "Bitnet: scaling 1-bit transformers for large language models." arXiv preprint arXiv:2310.11453 (2023).
2. Ma, Shuming, Hongyu Wang, Lingxiao Ma, Lei Wang, Wenhui Wang, Shaohan Huang, Li Dong, Ruiping Wang, Jilong Xue, and Furu Wei. "The era of 1-bit LLMs: all large language models are in 1.58 bits." arXiv preprint arXiv:2402.17764 (2024).
3. Hinton, Geoffrey. "The forward-forward algorithm: some preliminary investigations." arXiv preprint arXiv:2212.13345 (2022).
4. "Silent Data Corruption," available online at https://support.google.com/cloud/answer/10759085?hl=en#:~:text=Silent%20Data%20Corruption%20(SDC)%2C,to%20data%20loss%20and%20corruption. accessed on Sep. 07, 2024.