A Survey of Low Power Design Techniques for Last Level Caches

Emmanuel Ofori-Attah, Xiaohang Xang, Michael Opoku Agyeman
University of Northampton
Faculty of Art, Science & Technology

Abstract

- The end of Dennard scaling has shifted the focus of performance enhancement in technology to power budgeting techniques, specifically in the nano-meter domain because, leakage power depletes the total chip budget.
- Therefore, to meet the power budget, the number of resources per die could be limited. With this emerging factor, power consumption of on-chip components is detrimental to the future of transistor scaling. Fortunately, earlier research has identified the Last Level Cache (LLC) as one of the major power consuming element. Consequently, there have been several efforts towards reducing power consumption in LLCs. This posters presents a survey of recent contribution towards reducing power consumption in the LLC.

Aims & Objectives

- Conduct a survey o recent contributions towards reducing the power consumption in LLCs
- Introduce the techniques and highlight the pros and cons of each technique.

Monitoring Cache Behaviour

- One of the most effective ways to reduce power consumption in on-chip caches is by monitoring cache blocks. Along these lines, two different approaches can be implemented (bypass predictions and dead blocks).
- Dead Blocks: Majority of cache blocks in the LLCs are never reference thus , useless dead blocks dominate the LLC. Removing these dead blocks can reduce the power.
- For an LLC architecture which incorporates STT-RAM technology, bypassing writes to its bank could reduce the high write energy.

Discussion

- The most efficient way of saving power in LLCs is hybrid architectures because the power savings of leakage power is very high.
- Power consumption is however reduced at the expense of a slight degrade in performance.
- Power consumption is however reduced at the expense of a slight degrade in performance.

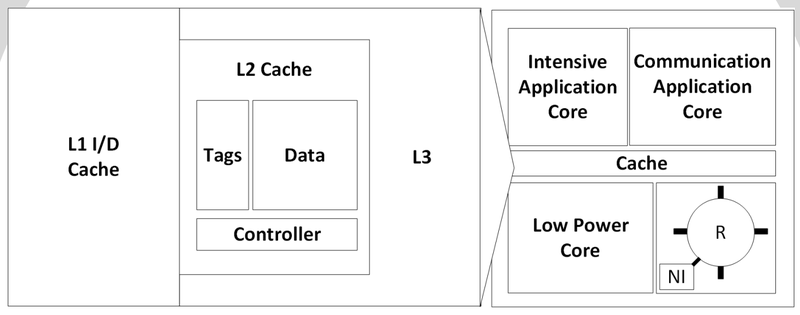
Conclusion

Our findings demonstrate that, integrating caches with STT-RAM and SRAM provides an effective solution to the leakage power consumption dominating modern technology. By creating a hybrid memory, STT-RAM banks can be used for read-intensive data while SRAM is used for write-intensive workloads. In addition to this, LLC power consumption can also be reduced by data compression schemes, eliminating dead blocks, and resizing cache size.

References

1. D. Wendel, R. Kalla, R. Cargoni, J. Clables, J. Friedrich, R. Frech, J. Kahle,B. Sinharoy, W. Starke, S. Taylor, S. Weitzel, S. G. Chu, S. Islam, and V. Zyuban, \The implementation of power7tm: A highly parallel and scalable multi-core high- end server processor," in SSCC, 2010.
2. J. Li, C. J. Xue, and Y. Xu, \Stt-ram based energy-eciency hybrid cache for cmps," in IEEE/IFIP 19th International Conference on VLSI and System-on-Chip, 2011.

Figure 1 - a multi-level cache architecture



(a) Multi-Level Cache Architecture

(b) Node: Core x3, Cache and Router

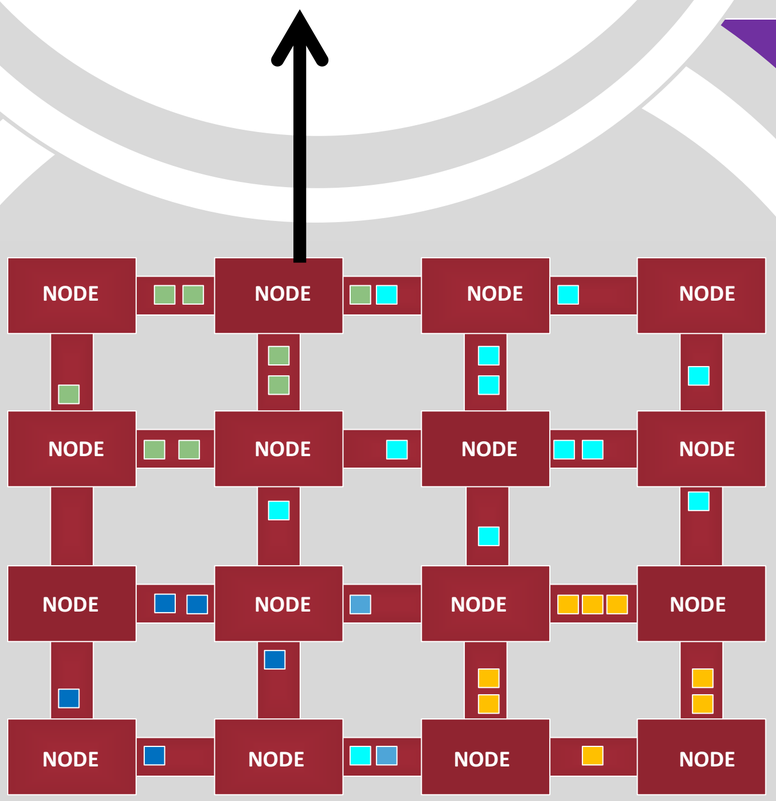


Figure 2 - 4x4 Multi-core Architecture

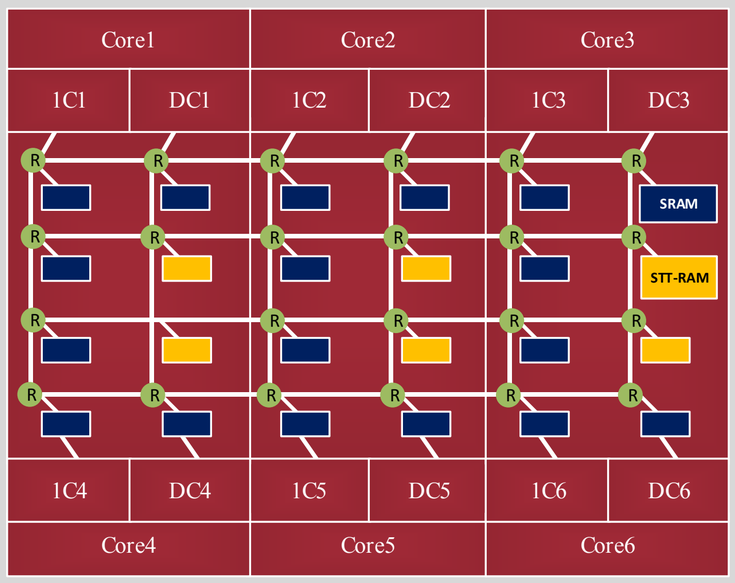


Figure 3 - Hybrid Architecture

Introduction

- Multi-level Cache Architectures (MCA) have become increasingly popular for mitigating the disparity between memory and processors trading-off power consumption.
- With leakage power set to dominate power consumption in the near future, a reduction in LLC power and area can increase the number of components which can be activated through the Dark-Silicon solution.
- This poster presents LLC power consumption techniques.

Reducing Cache Size

- Another widely used technique to reduce power is to shutdown parts of the cache which are idle.
  - However, shutting down idle parts of the cache affects performance. Particularly, when there is a sudden overshoot in the workloads.
    - Particularly, when there is a sudden overshoot in the workloads. Therefore, power-gating techniques should consider performance degradation when resizing the cache (cache banks and ways are powered-off)

Hybrid Architectures

- SRAM consumes a significant amount of power and due to its write energy, using STT-RAM as a standalone technology in LLC makes it difficult to implement.
- Hybrid memories have been proposed to exploit the low leakage power of STT-RAM and high performance of SRAM.
- In such an architecture, SRAM is used for write-intensive workloads while the STT-RAM is used for read-intensive workloads.

Figure 4 – Comparison of Technologies

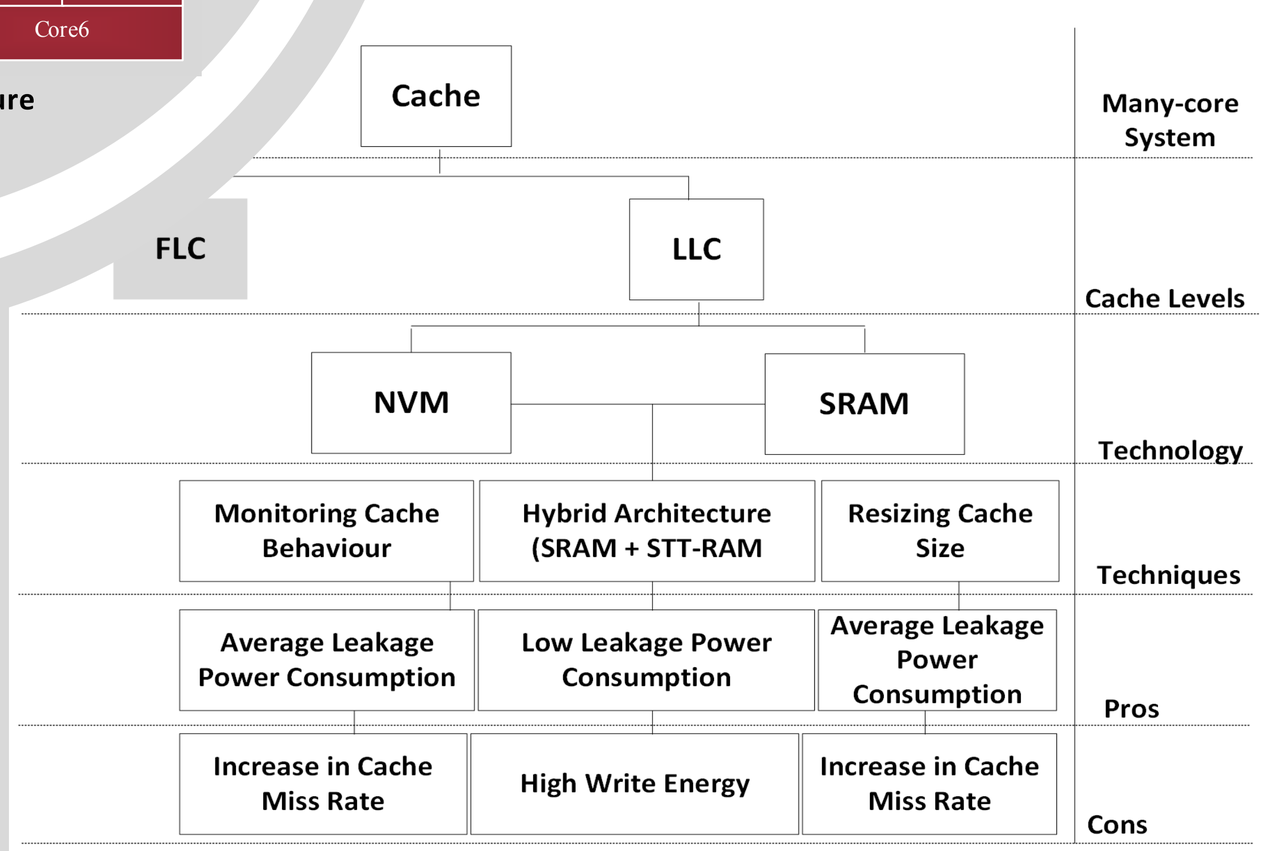


Figure 5 – Summary of the techniques presented along with their pros and cons

Summary of Table