Technology scaling impact on the interconnect of clustered scratchpad memory architectures

Iason Filippopoulos Francky Catthoor
Per Gunnar Kjeldsberg

August 5, 2015

Abstract

This work

1 Introduction

Power will be the key limiter to as interconnection networks take up an increasingly significant portion of system power.

2 Related Work

3 Current technology

3.1 Generic Work-flow

The current technology is studied as a first step towards the development of an interconnection cost model. CAD tools allow the design of the described clustered memory architectures. The synthesis and the simulation provide reliable data for the area and the power consumption of the different parts of the memory architecture. The goal is to synthesize a clustered memory architecture and extract power data for the memory banks and the interconnection logic separately. The work-flow is divided in several sub-steps:

- A number of memory banks is chosen from a library, which contains several SotA designs.
- An RTL description for connecting the memory banks into a full memory architecture design is written.

- A simulation is performed to verify the correct functionality of the memory architecture.

- A target technology is chosen and the logic synthesis of the memory architecture is performed.

- The floor-planning and the placing & routing of the memory design is performed.

- The dynamic timing and the power simulation are performed and the results are provided.

Regarding the first step, the memory models include existing state-of-the-art memories, available from industry and academia, presented in [1]. For the memory banks based on the commercial SRAM memories, the energy numbers are derived from a commercial memory compiler. For the memory banks based on the experimental standard cell-based memories (SCMEM) [3], the energy numbers are result of synthesis.

For the second step, the RTL description connects the memories using MUX, signals and other components into a functional clustered memory architecture. On the third step, the simulation and the verification performs a flash-write followed by a read on the whole memory architecture. The target technology depends on the available libraries. In this work, the TSMC library on 45nm is used. The place and route can be either performed automatically through the CAD tool or manually by the designer.

The final step, includes the extraction of parasitic, the static timing analysis and the annotation of the timing to the netlist. Afterwords, power simulations on the synthesized design are carried out using Synopsys PrimeTime, in order to obtain energy numbers.

3.2 Example design: synthesis and simulation

A group of clustered memory architectures is designed and synthesized following the presented work-flow. The simulation provides results for the current technology and the contribution of the interconnection to the overall energy consumption. The study includes clustered memories with an increasing number of memory banks, beginning with only one memory bank and continue up to five memory banks. The breakdown of energy consumption is split into two parts of the memory. The first part is the energy consumption inside each memory bank, which includes the memory cells, the necessary logic to connect the cells and the addressing circuit. The second part is the interconnection cost between the different memory banks, which includes the

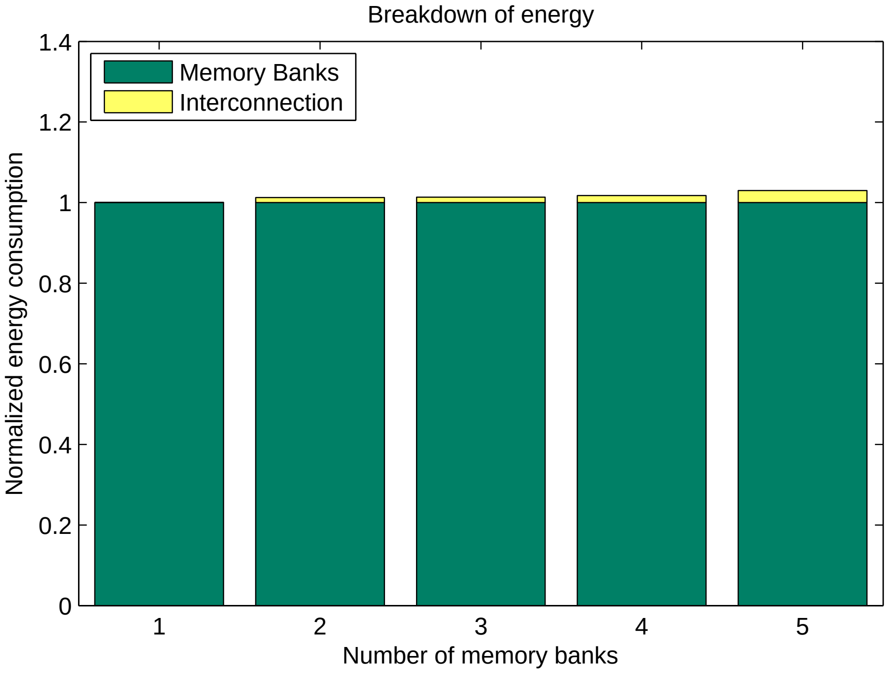

Figure 1: Energy breakdown between the memory banks and the interconnection

Table 1: Percentage of energy overhead on interconnection

Banks	1	2	3	4	5
Overhead(%)	0	1.2627	1.3687	1.7683	3.0067

necessary logic to locate and transfer the data outside of the banks. In other words, the interconnection outside the memory banks is given separately. The interconnection between the different memory cells inside one bank is included on the energy of the memory bank.

The energy breakdown between the first part of the memory banks and the second part of the interconnection (part2) is presented in Fig.1. The energy cost of the interconnection logic is small compared to the energy cost for the write and read operations on the memory banks. Tab.1 contains the exact percentages of the energy overhead on the interconnection.

The changes in the area of the design is presented in Fig.2. The area used for the placement of the memory banks is separated from the area occupied by the interconnection.

In addition, the synthesis and simulation of the memory designs provide useful information about the dimensions of the memory banks, the length

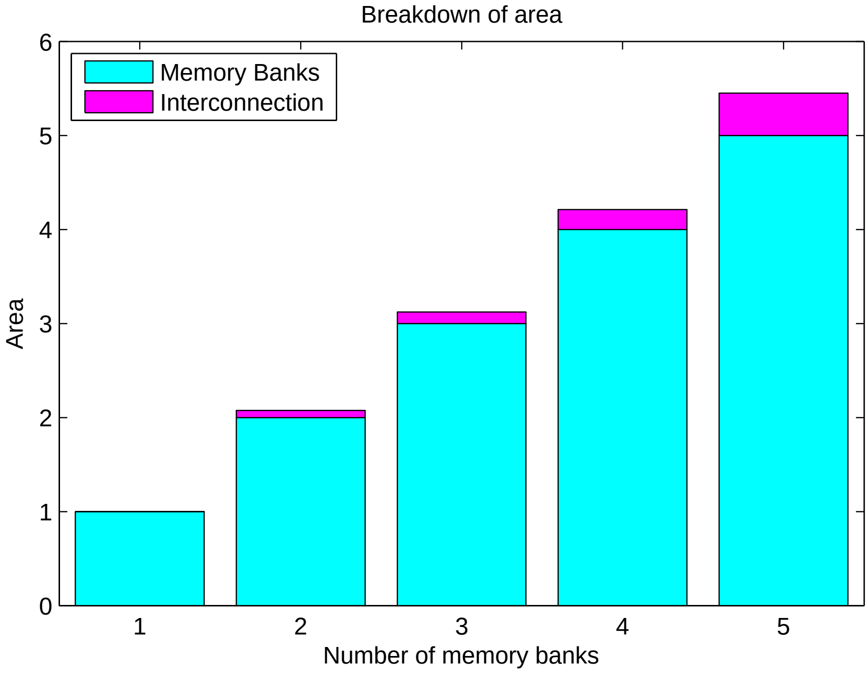

Figure 2: Area breakdown between the memory banks and the interconnection

and the capacitance of the wires. Several interesting observations are possible based on the study of the current technology. The energy overhead on the interconnection of the banks grows when there are more memory banks connected, as expected. However, the overhead is just over 3% even for a clustered memory with five banks. The area overhead is significantly higher and grows exponentially for an increasing number of memory banks. The maximum area overhead is less than 10%. This is explained by the need for extra wiring to connect the different memory banks.

4 Technology Scaling

The technology scaling projections are based on the reports released by the International Technology Roadmap for Semiconductors (ITRS) [2]. The clustered memory architecture can be divided into two parts, i.e. the memory banks and the interconnection.

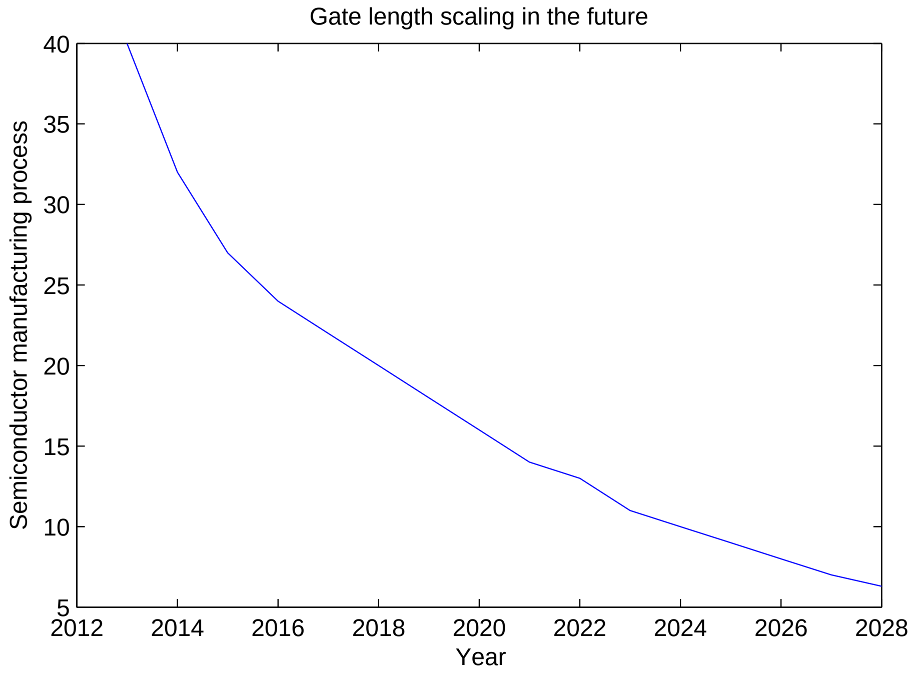

Figure 3: Impact of technology scaling into gate length

4.1 Memory Banks

The memory banks consist of the memory cells, which are built using gates. Thus, the predictions about the future behavior of the memory banks is based on the ITRS reports for logic. In the following years, the gate of the length is expected to be reduced as shown in 3. The values are approaching 5nm around 2028, which is potentially the limit using the current manufacturing process.

The reduction on the gate length leads to a reduction of the memory cell and naturally smaller memory banks. The smaller memory banks affect the area of the design and the power consumption. The projections provided by ITRS regarding the power are presented in 4. There is a significant reduction on the power consumption in the short term and slighter reduction at the end of the projections.

4.2 Interconnection

The interconnection cost is based on the projections about wiring, which differentiate compared to the projections regarding the logic gates. However, the change on the size of the memory banks affects the length of the needed

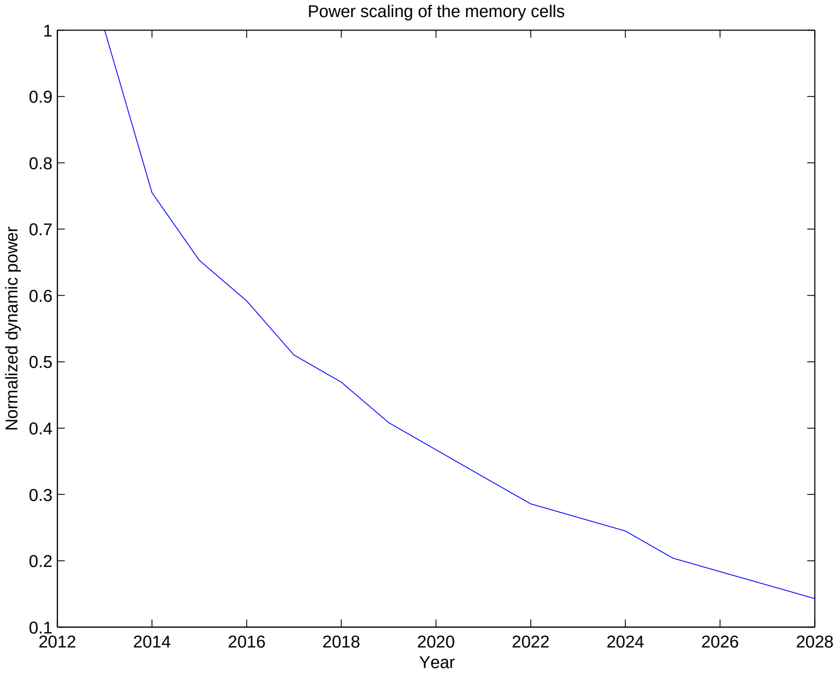

Figure 4: Normalized dynamic power and technology scaling

wires. The most important parameters for the current study is the capacitance and the power of the interconnection. Based on the data provided by

ITRS, the curves for the two parameters are presented in 5. The capacitance is expected to be reduced in the following year. However, the rate of reduction is lower compared to the expected reduction for the memory banks. The power is even worse

5 Model Construction and Projection Results

Model:

- Given the dimensions of the memories an estimation about length of the needed wires can be made.
- Capacitance for the wires can be approximated by the results given for the studied example.
- Power can be calculated as $Power = \frac{1}{2} \times f \times C \times V_{dd}^2$

6 Conclusion

References

[1] Iason Filippopoulos, Francky Catthoor, and Per Gunnar Kjeldsberg. Exploration of energy efficient memory organisations for dynamic multimedia applications using system scenarios. *Design Automation for Embedded Systems*, pages 1–24, 2013.

[2] International Technology Roadmap for Semiconductors. Annual summary, 2013.

[3] Pascal Meinerzhagen, SM Yasser Sherazi, Andreas Burg, and Joachim Neves Rodrigues. Benchmarking of standard-cell based memories in the sub-vt domain in 65-nm cmos technology. *IEEE Transactions on Emerging and Selected Topics in Circuits and Systems*, 1(2), 2011.

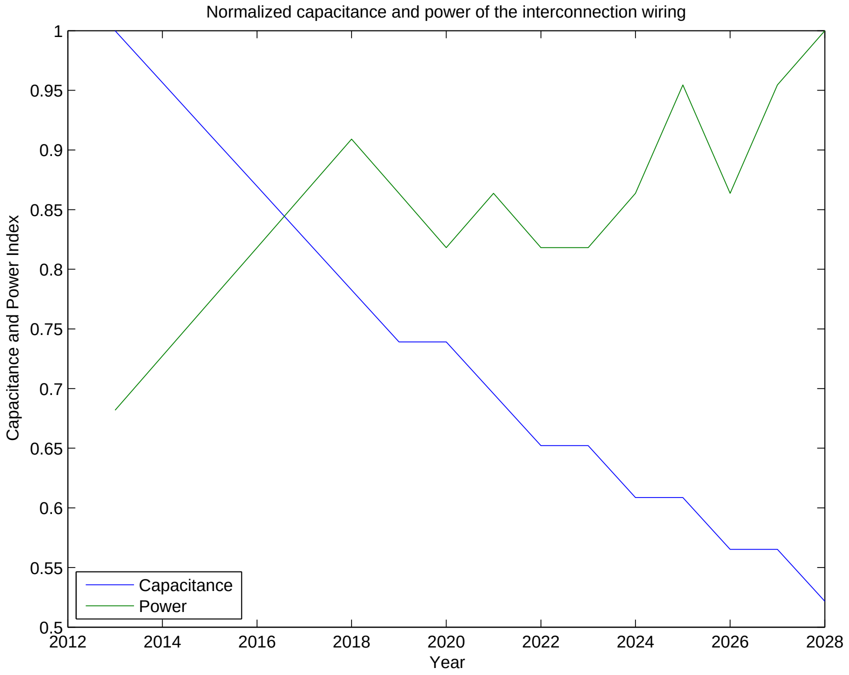

Figure 5: Impact of technology scaling into the capacitance and the power of the interconnection part