Low Power Register Files and Rob for Super Scalar Out-of-order Processor

[1]G. Dhanalakshmi and [2]M. Sundarambal

[1]Sri Ramakrishna Engineering College, Coimbatore, India
[2]Coimbatore Institute of Technology, Coimbatore, India

Abstract: In order to achieve a high performance during a single process, superscalar processors now relay on a very complex out of order execution. Increasing in on chip hardware for high performance will also increase power and energy dissipation. Register Files and Reorder Buffer are the multiported SRAM array, which are growing rapidly with larger Instruction Level parallelism. Due to the Frequent multiple access of RF and Rob, become a power consuming critical data path component in the dynamic scheduling mechanism. The proposed design of RF, aims at reducing the average power consumption using divided bitline line approach along with SFLD technique by adaptive resource resizing. Simulation results shows that the new design of RF reduces power dissipation along less hardware modifiable with least amount of redesign and verification efforts and also with or without minimal impact on the significant performance compared to convention design of RF.A similar methodology can be applied another critical data path unit Reorder Buffer also.

Key words: Register Files · ReOrder Buffer · Issue Width · Adaptive resource resizing

INTRODUCTION

With the arrival of digital convergence era, embedded applications with high performance such as multimedia, Networking and mobile computing are increasingly using complex out-of-order super scalar processors to meet the performance goals. NEC's VR5500 and VR77100 Star Sapphire [1] and the IBM PowerPC 750FX [2] processors are the examples of such processors. Due to the Technology scaling in the Ultra Deep Sub Micron (UDSM) region has allowed hundreds of millions of gates to be integrated and fabricated onto a single chip. The ever-increasing levels of on chip integration in the recent decade, increases in the computer system most performance effective. Unfortunately improvement in the performance also increases the chip's power and energy dissipation. This higher energy and power dissipation needs expensive packaging and cooling technology, which increases the cost, decreases product reliability in all segments of computing technology market place and also significantly reduces life time of the battery in the portable systems.

Designers thus have silicon budget to add more processor resources (e.g., increasing register file size, reorder buffer size, etc) in order to utilize application level parallelism and improve performance. Due to the restrictions with the low power budget and practically achievable high operating clock frequencies act as limiting factors which prevent unbounded increase in processor resource sizes.

In modern superscalar micro architectures, the access to Register File lies on the significant schedule-to-execute data path. As a new physical Register needs to be allocated for every instruction with a destination Register and many Register Files are required to implement Instruction-Level Parallelism (ILP) out of sequential programs. Moreover, to reduce the amount of data transfers, a single RAM memory structure is normally used to sustain both committed and speculative Register File values [3-5].

Register File are the heart of a superscalar microprocessor core and it providing communication of register values between producer and consumer instructions. As issue width increases, both the number of ports and the number of register required are also increases. It causing the area of a conventional multi-ported Register File to grow more quadratically with issue width [6]. The physical Register Files are located in critical data path and limit the processor clock frequency. For instance increasing Register File (RF) size increases its access time, which reduces processor speed. One of the most critical timing factors is access time which

Corresponding Author: G. Dhanalakshmi, Sri Ramakrishna Engineering College, Coimbatore, India.

determines the achievable processor operating clock frequency in the multi ported Register File. Dynamically resizing the RF widthby using Dynamic Frequency Scaling (DFS) technique which significantly improves the processor performance and also reduces energy-delay product for high speed embedded processors [7-9].

Out-of-orderprocessor pipelining process is capable of fetching, decoding and also renaming several instructions per processor clock cycle. Based on Issue width of the processor can execute and then commit up to as many instructions in each every cycle. Like this type of out-of-order multiple-issue super scalar processor accesses the Register File very frequently. Up to Issue Width reads and Issue Width writes can be able to issue in the register file per clock cycle. The physical Register File is designed as a 6T SRAM structure with as many write and read ports as the maximum number of instructions the processor can be issued in each and active components in a processor design. Due to such a frequent access make the Register File one of the power consuming units in a high performance super scalar processor. Large amount of power is dissipated in a small size SRAM structure [10, 11],which makes the Register File into "heat stroke". When the temperature of any of the processor units exceeds beyond a critical limit heat stroke will occur. To reduce the chances of a heat stroke it is crucial to reduce the power of the Register File [12].

The main Functionalities associated with the ROB and RF [13]:

- Setting an entries For the Issue width of the instructions in each and every cycle
- During commit stage, Releasing up to the Issue width of the entries in a cycle
- During the branch and branch prediction, Flushing the entries

To preserve power and energyby dynamically and also simultaneously adaptive resizing the *multiple* data pathresources based on the demands of the consumer applications. Then unused resources are deactivated temporarily andreactivated when the resource demands goes up. The net result produces a drastic savings in the overall power/energy with minimal significant impact on the processorperformance [14, 15].

Related Work and Back Ground: Many of researches that has been proposed for changing the architecture and also a structure of the Register File (RF) for improving and getting better performance results. Some of the techniques utilizing localities of communication to divide the microarchitecture into various distributed clusters and each and every containing a subset of the RF [16, 17]. A significant issue in the proposed design of such systems is used to map instructions to various clusters. These type of technical schemes have the prospective to scale down the larger issue widths of the RF. But it requires complex a robust intercluster control logic for mapping instructions into clusters and also to handle inter-cluster independencies and dependencies. On the other hand, one set of approaches maintain a centralized microarchitecture which partition the processor units such as the *RF* [18, 19] for the reduction accessing time, energy and power dissipation. Partitioning of the *RF* into multiple Register banks for example reduces the ports on the partitions. It will reduce the charging, pre-charging and sensing times and the related energy and power dissipation. But these type of reductions coming under the cost of value multiplexing and also port conflict troubles. The main disadvantage of all banking techniques is in the complexity which will adds speculation and particularly the complexity they introduce in Register Caches and banking conflict issues for handling the coherency. In the resource-constrained environments, this added complexity becomes more critical issue for embedded processors. Early Register File allocation and de-allocation based on L2 misses to better performance [20]. Stillsuch a type of scheme increases complexity. It also requires other resources to identify the sources and the destinations of the load independent and load-dependent instructions such as bit vectors, ROB additional bits and to store de-allocated values in backup Register File.

Mostly many of past work is based on the design of the Register File either attemptlimit its size or to limit the number of ports used [21-24]. In [23], bypassing of the Register File data which will be used to the decode stage from the fetch stage. Hence putting registers which are in unused stateinto the low power mode in the early stage of the pipeline. In order to avoid considerable performance penalty, before being accessing the registers had to be set back to high power mode in one cycle. There are proposals at the cost of added arbitration hardware for the reduction of the number of ports. But these techniques requiresignificant modifications to thePipeline stage. Borch *et al.* have representedproblems related withreducing the number of Register File ports [1]. Caching Registers, banked Register Files and two-level Register Files units are analyzed to reduce the number of registers for power required power reduction [1, 22, 19, 25,26]. In [9], Alastruey *et al.* projected to
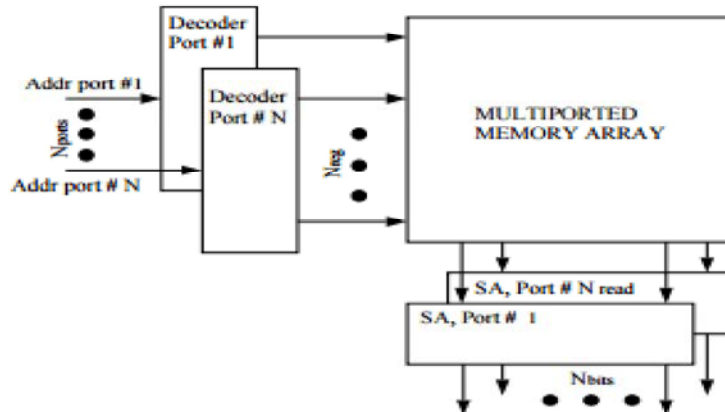
Fig. 1: Generic register file memory cell with Nread read ports and Nwrite write ports

support speculative register renaming by adding an auxiliary off core Register File. Releasing the physical and logical Register File before its coming to the commit stage andits consumers have read it before [2, 27]. The Majordrawback of Speculation based technique is in the complexity that speculationadds in register caches and banking conflicts.

Conventional Design of Register Files: The general structure Register File memory with N_{read} read ports and N_{write} write ports is shown in Fig. 1.

All the bit lines should be precharged high to read or write the data in each and every cycle in memory. During the write operation on the bit lines set to be logic high into a cell when the word line a fired. To read the data content of an entry, one of bit line or will be temporarily discharged. The sense amplifier is used to detect such type of differences. But the bit lines must run across the entire ROB or Register File height. The write bit lines are the most important sources of power dissipation in ROB and RF structures during multiple accesses to the same cell in the each cycle [4]. Most of the leakage current dissipation is occurred due to the leakage currents of the memory cells, to the bit lines which flow through the two off pass transistors. In order to eliminating the leakage current dissipations in the memory cells, power consuming units would be to turn off the unused entries and their associated wordline drivers using such a gated-V_{dd} or gated V_{ss} power gating technique (Sleep mode). The transition from sleep mode to active mode adds a one-cycle delay to the ROB or Register File access. It has significant performance impact of frequently activate and deactivate the entire memory unit. Resizing the ROB could be achieved by partioning it into several dependent and independent units with separate sense amplifiers, input drivers and output drivers as explained in [4].

The complexity of ROB and Register Files can be avoided by the divided bit lines technique for SRAMs to reduce the bit line capacitance and hence its dynamic power dissipation. Figure 2 shows the circuit diagram of Register File.

To divide the bit line into several sub-bit lines two or more SRAM cells are combined together. Dynamic power dissipation is achieved by reducing the effective capacitance. To downsize the RoB or RF, the select signal of the low partition is being AND with other higher partitions together with the down signal. No read and write operationcan be completedto/from the partition. So that the entire partition will be turned off safely. The power gating of gatedV_{dd} technique is used to turn off the entire partitions in the RF (RoB) [25], to reduce the voltage in all of the partition memory cells and eliminating its leakage current completely. Same technique is used to eliminate leakage current in the disabled partition of word-line driver. When a cache miss period occurs, it triggers upsizing the unit and turning on the disabled partition and triggers downsizing the RoB at the end of the cache miss period. When the segment is empty, then the downsize signal is asserted. The benefits of such resizing RoB or RF is in reducing both dynamic power and leakage power dissipation. While turning off the entire segment of memory cells and word line driver, Leakage current power dissipation is suppressed. Hence Dynamic power consumption is suppressed due to a smaller equivalent capacitance on the bit lines and tag lines. Duringthe cache miss period ends, the size of RF and RoB is reduced back to half of their size. It is necessary to detect when the lower partitions of RoB or RF become empty after the end of cache miss period. It can be achieved by implementing an additional bit in each row of the lower partition. While using lower partition, additional bit is set and then reset when the entry is released during commit operation.
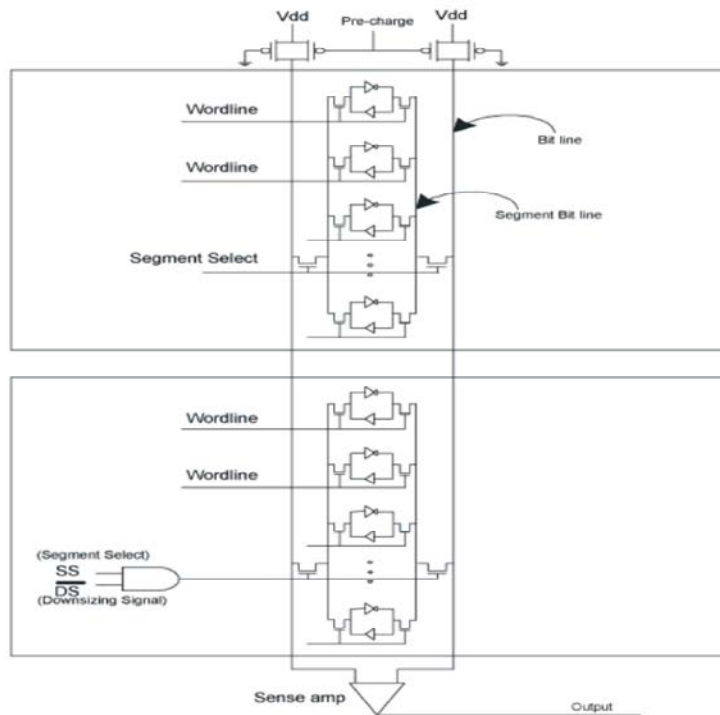
Fig. 2: Circuit diagram of conventional RF [1]

By Orin these bits we can detect whether the partition is empty or not. The logic to ORing all bits in the lower partition of RFis not on the critical path. Then the downsizing decision is done by parallel to accessing ROB and Register file. All the bit lines should be pre-charged high (fired) to read or write an entry cycle. To read the data content of an entry, one of bit line or will be conditionally discharged. To detect such a multiple access, the sense amplifier is used.

Proposed Design of Register Files: In order to improve the performance of the register file, dynamic logic is used in deep submicron technology based microprocessor design.

However the main problem associated with dynamic logic design is more sensitive to second order effects such as sub threshold leakage current. Due to rapid change in technology scale down, supply voltage and threshold voltage reduction for improving the performance of the processor units. Reducing leakage current with high performance is major concern in the processor unit design. Figure 3 shows the circuit level implementation of Register file. The new design domino based circuit design in the register file is implemented to reduce the power consumption. The proposed design works in such a way that it contributes towards reduced power dissipation by minimize -parasitic capacitance,

leakage current and current contention**.** Speed is decreased dramatically when large fan-in gates and the capacitance of the node is large,. In addition to this, due to many parallel leaky paths in wide fan in gates, noise immunity of the gate is also reduced. Although upsizing of the keeper transistor can improvenoise robustness. But power consumption and delay are increased due to large contention. In order to avoid these types of problems, pull down network implements the logic function and it is separated from the keeper transistor by implementing comparison stage. The current of pullup network is compared with worst case leakage current.

To reduce the leakage current the footer nMOS transistor MN2 is connected to the source of evaluation nMOS transistor to obtain the FDL [21] in the register file design. The speed the SFDL is lower than the footless one for the reason that of the stacking effect. But the noise immunity is higher. When clock is low, the dynamic node is precharged to VDD.

In this phase the footed transistor MN2 is turned OFF, which is used to reduce the leakage current. Footer transistor MN2 is turned ON, when clock is set to high. So, the state of output node is obtained to get a desired result depending on incoming data to pull-down network. SFLD circuit is combined with the Register File in the super scalar processors for reducing the average power consumption and also minimal impact on the performance.
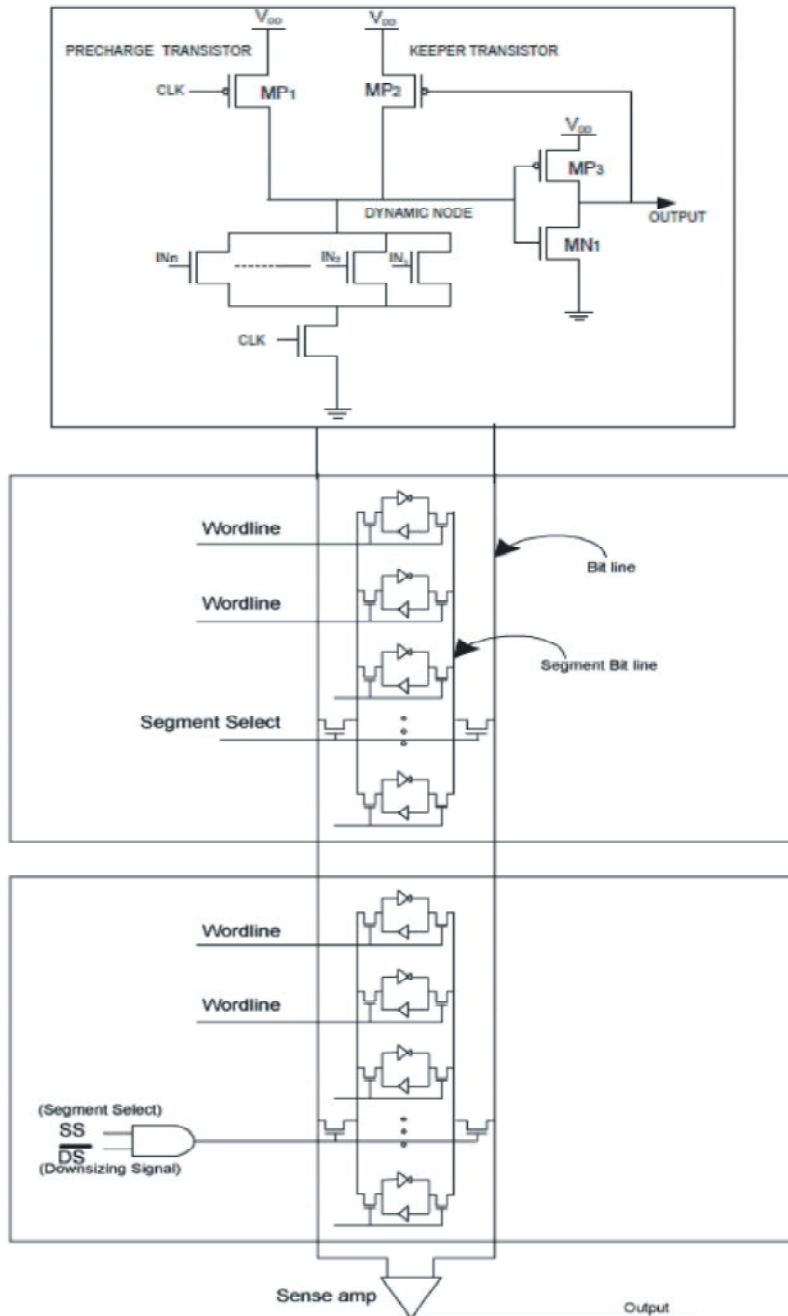
Fig. 3: Circuit diagram for the proposed RF

SIMULATION AND RESULTS

Simulation Environment: Both conventional RF and proposed RF have been simulated using BSIM 3V3 45nm technology on Tanner EDA tool. The testing environment is created and tested with the same input patterns on room temperaturewith supply voltage ranging from 0.6V to 1.0V.

Simulation Analysis: The tested results of conventional RF and proposed RF functional waveforms are obtained during waveform simulations which are shown in Fig 4 and Fig 5 respectively.

The performance comparison of the proposed RF with the conventional RF is shown in Table-1. The Power reduction is achieved for the proposed circuit design at 45nm fabrication technology compared with conventional circuit design of RF.
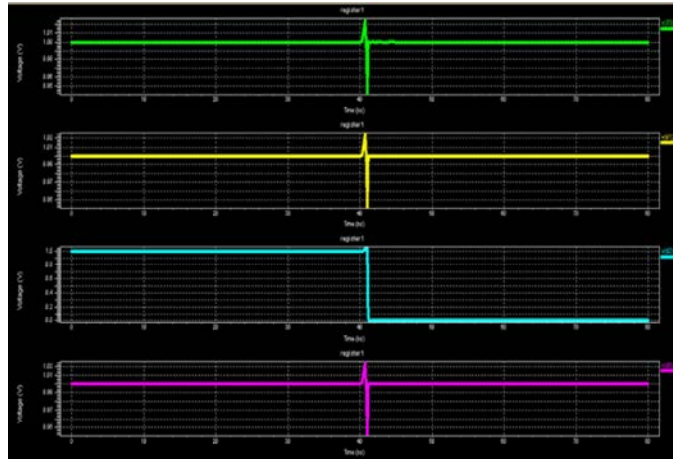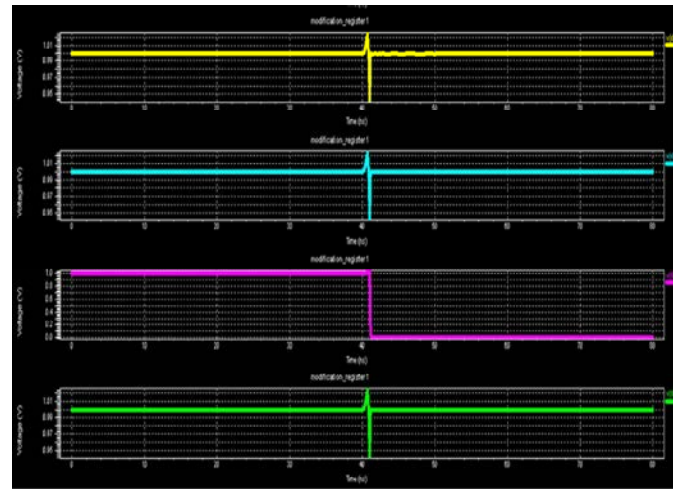
Fig. 4: Waveform of the Conventional RF

Fig. 5: Waveform for the Proposed RF

Table 1: Performance Comparison of Conventional RF with Proposed RF

	Average Power consumption	
Supply voltage (VDD)	Conventional RF (uW)	Proposed RF (uW)
0.6V	173.68	98.52
0.7V	202.18	123.89
0.8V	297.59	153.27
0.9V	305.24	198.48
1.0V	408.33	205.38

CONCLUSION

The design of Register Files is a one of the critical part of a superscalar Out of Order execution processor. Due to its Frequent multiple access in dynamic scheduling, the Register Files is responsible for a significant amount of overall processor power and energy dissipation in the high performance processor. The proposed design register files aims to both dynamic and static power dissipation. It is designed using Tanner EDA tool at 45 nm technology. In the high performance Out of Order superscalar processor, the test results of new design of RF using divided bit line approach with SFLD technique outperforms in decrease in both static and dynamic power consumption with or without minimal impact on the performance and stability improvement of the memory cells and when compared to Conventional RF. In addition to this, the power saving technique used is technology independent one and also it can be combined with some other orthogonal power saving techniques to save power.

REFERENCES

1. StijnEyerman, 2006. Efficient Design Space Exploration of High Performance Embedded Out-of-Order Processors, in DATE 2006.

2. IBM Corporation. PowerPC 750 RISC Microprocessor Technical.

3. Tune, E., R. Kumar, D.M. Tullsen and B. Calder, 2004. Balanced multithreading: Increasing throughput via a low cost multithreading hierarchy, in Proc. 37th Annu. Int. Symp. Microarch. (MICRO-37), Portland, OR, pp: 183-194.

4. Ponomarev, D., G. Kucuk and K. Ghose, 2002. Energy-efficient design of the reorder buffer, presented at the Int. Workshop Power Tim. Model, Opt. Simulation (PATMOS), Seville, Spain.

5. Abella, J., R. Canal and A. González, 2003. Power- and complexity-aware issue queue designs, IEEE Micro, 23(5): 50–58.

6. Canal, R. and A. Gonzalez, 2001. Reducing the complexity of the issue logic, presented at the Int. Conf. Supercomput., Naples, Italy.

7. Brooks, D., V. Tiwari and M. Martonosi, 2000. Wattch: A framework for architectural-level power analysis and optimizations, in ISCA 2000.

8. Mesa-Martinez, F.J., J. Nayfach-Battilana and J. Renau, 2007. Power model validation through thermal measurements, presented at the Int. Symp. Comput. Arch., San Diego, CA.

9. Alastruey, J., T. Monreal, V. Vinals and M. Valero, 2007. Microarchitectural support for speculative register renaming, presented at the Proc. 21st IEEE Int. Parallel Distrib. Process. Symp., Long Beach, CA.

10. Han, Y., I. Koren and C.A. Moritz, 2005. Temperature aware floorplanning, presented at the Workshop Temp. Aware Comput. Syst., Madison, WI.

11. Skadron, K., M.R. Stan, W. Huang, S. Velusamy, K. Sankaranarayanan and D. Tarjan, 2003. Temperature-aware microarchitecture, presented at the ISCA, San Diego, CA.

12. Hasan, J., A. Jalote, T. Vijaykumar and C. Brodley, 2005. Heat stroke: Power-density-based denial of service in SMT, presented at the Int. Symp. High-Perform. Comput. Arch., San Francisco, CA.

13. Homayoun Houman, AvestaSasan, Jean-Luc Gaudiot and Alex Veidenbaum, 2011. Reducing power in all major CAM and SRAM based processor units via centralized,dynamic resource size management, IEEE transactions on Very Large Scale Integration System, 19(11).

14. Li, H., 2003. VSV: L2-miss-driven variable supply-voltage scaling for low power, in MICRO.

15. Homayoun, H., S. Pasricha, M. Makhzan and A. Veidenbaum, 2008. Improving performance and reducing energy-delay with adaptive resource resizing for out-of-order embedded processors, presented at the ACM SIGPLAN/SIGBED Conf. Lang., Compilers, Tools for Embed. Syst. (LCTES), Tucson, AZ.

16. Terechko, A., M. Garg and H. Corporaal, 2005. Evaluation.of speed and area of clustered VLIW processors, VLSI Design.

17. Ergin, O., 2004. Increasing Processor Performance through Early Register Release, in ICCD2004.

18. Tseng, J.H., 2003. Banked Multiported Register Files for High- Frequency Superscalar Microprocessors, ISCA 2003.

19. Balasubramonian, R., 2001. Reducing the complexity of the register file in dynamic superscalar processors, in MICRO-34.

20. Sharkey, J. and D. Ponomarev, 2007. An L2-Miss-Driven Early Register Deallocation for SMT Processors, in ICS 2007.

21. Geissler, S., 2002. A low-power RISC microprocessor using dual PLLs in a 0.13/spl mu/m SOI technology with copper interconnect and low-k BEOL dielectric, in ISSCC 2002.

22. Tseng, J.H. and K. Asanovic, 2003. Banked multiported register files for high-frequency superscalar microprocessors, in Proc. 30th Int. Symp. Comput. Arch, pp: 62-71.

23. Ayala, J.L., M. Lopez-Vallejo, A. Veidenbaum and C.A. Lopez, 2003. Energy aware register file implementation through instruction predecode, in Proc. IEEE Int. Conf. Appl.-Specific Syst, Arch., Processors (ASIP), pp: 86-96.

24. Homayoun, H., S. Pasricha, M. Makhzan and A. Veidenbaum, 2008. Dynamic register file resizing and frequency scaling to improve embedded processor performance and energy-delay efficiency, presented at the 45th Des. Autom. Conf., Anaheim, CA.

25. Park, I., C.L. Ooi and T.N. Vijaykumar, 2003. Reducing design complexity of the load/store queue, presented at the Int. Symp. Microarch., San Diego, CA.

26. Abella, J. and A. González, 2006. SAMIE-LSQ: Set-associative multipleinstruction entry load/store queue, presented at the IEEE Int. Parall. Distrib. Process. Symp. (IPDPS), Rhodes Island, Greece.

27. Balkan, D., J. Sharkey, D. Ponomarev and K. Ghose, 2006. SPARTAN, speculative avoidance of register allocation to transient values for performance and energy efficiency, in Proc. 15th Int. Conf. Parallel Arch. Compilation Techn. (PACT), pp: 265-274.