Extended Abstracts of the 2000 International Conference on Solid State Devices and Materials, Sendai, 2000, pp. 360-361

E-3-5

Fast and Compact Central Arbiter for High Access-Bit-Rate Multi-Port Caches

Nobuhiko Omori, Koji Kishi, Takayuki Gyohten, Jongshik Kim and Hans Jürgen Mattausch

Research Center for Nanodevices and Systems, Hiroshima University, 1-4-2 Kagamiyama, Higashi-Hiroshima 739-8527, Japan
Phone: +81-824-24-3046 Fax: +81-824-22-7185 e-mail: nomori@sxsys.hiroshima-u.ac.jp

1. Introduction

Recent 1-chip processors implement super-scalar architectures for parallel execution of multiple issues and operate at GHz clocks. Future development will go to simultaneous multithreading and eventually complete 1-chip multiprocessors [1], so that data- and instructions streams of enormous bit-rates must be exchanged with processor cores. A 32bit, 12 issue, 1GHz core will e.g. require about 1.2 Tbit/sec to support its peak-performance. To meet these requirements a 1-port, 32bit cache would have to operate at 37.5GHz in a random access mode, i.e. at much higher clock frequencies than the core. This calls for application of multi-ported caches. However, multiple ports lead to the possibility of access conflicts (at least in write accesses) and an arbiter for conflict detection and regulation becomes necessary. In this paper we propose a central arbiter to keep area-overhead and access-time penalty negligible up to 32 ports.

2. Cache-Miss Rate, Cache-Organization and Arbiter

Minimization of cache-miss rate is important, because access to main memory leads to wait-times of several cache-access cycles, before the required data becomes available. To reduce cache-miss rate, large storage-capacity is desirable.

However, conventional architectures can't realize large capacity and port-number (N) simultaneously, because storage-cell area increases with N^2 [2]. Therefore, application of a recently proposed hierarchical architecture, which trades-off increased access-conflict probability for reduced area, is attractive [3,4]. Fortunately, access conflicts cause only a wait-time of one cache-access cycle, which is much less severe than a miss. Conflict probability can thus be even moderately larger than miss rate, to achieve substantial reduction of multi-port cache area [4], while cache-performance loss is small. The trade-off is demonstrated in Fig. 1 for a directly mapped 8-port cache with 512Kbit data-capacity, organized in 16K words of 32bit length. Large area-reduction to 1/3-1/4 is expected. A possible architecture for a directly mapped hierarchical multi-port cache is shown in Fig. 2. Increased wordlength for storing tags along with data-words plus circuitry for identifying misses are necessary. For associative multi-port caches, conventional concepts can be transferred in a similar straightforward way.

The interrelation between cache organization and arbiter complexity also favors hierarchy. If ports are implemented in the memory cells, the arbiter has to compare all address bits to detect conflicts. With a hierarchical organization, conflict detection involves only the address part of the upper hierarchy level. This substantially simplifies conflict-detection, which consumes most of the arbiter area. Thus the trade-off in Fig. 1 does not only substantially reduce cache-area, but additionally simplifies the arbiter for conflict handling.

A central arbiter as in Fig. 2 is attractive, because it provides an elegant way to remove arbitration delay-time from the critical access-path. Basic idea is to carry out arbitration and address decoding in parallel. Arbitration results are fed into the critical path after the decoder, to stop access from all but one conflicting port. Thus only one gate in the critical path needs a fan-in increase by an additional input. All other arbitration concepts and especially a decentralized concept would result in larger delay penalties.

3. Proposed Circuitry for the Central Arbiter

Our main objective is simplicity and minimum sequential gate number, for compactness and short delay.

Figure 3 shows the simplest arbiter concept with a port-importance hierarchy (PIH) algorithm for conflict regulation, where the port with smallest number n gets priority. The conflict-detection part consists of multi-input EXOR gates [5] for simultaneously comparing port addresses in pairs. Consequently, delay-time depends not on port number N, but only on address-bit number m_n. Since hierarchical cache organization minimizes m_n, minimized delay and area for conflict-detection result. Altogether N(N-1)/2 multi-input EXOR gates are required. The PIH conflict-regulation uses simply a single stage of NAND-gates with a maximum fan-in of N-1. Thus delay increases only moderately with N.

The PIH-algorithm is a worst-case possibility for conflict regulation, because access-rejection probability from each port is different and port N is always rejected when involved in a conflict. Figure 4 shows a simple best-case conflict-regulation algorithm, which alternately chooses between two inverted importance hierarchies. Approximately fair conflict regulation with equal access-rejection probability results for practical block numbers $M_n=2^{m_n}$.

4. Design-Study for the Central Arbiter up to 32 Ports

The design study is carried out for approximately equal access-rejection probabilities below 3% in all cases. This is larger than minimum miss rate, but leads to a cache-design with smallest area. Figure 5 shows the linear increase of m_n and the quadratic increase of multi-input EXOR gates with N under this boundary condition. The layouts for the PIH-algorithm in a 0.5μm, 2 metal CMOS technology are compared in Fig. 6. As expected, arbiter area increases strongly with N^2 and is dominated by the conflict-detection part. However, in comparison to the complete multi-port cache, the arbiter consumes only a negligible area fraction of less than 1% up to 32 ports as shown in Fig. 7. Reasons are: (a) Relative area-increase in comparison to the complete multi-port cache is only linear, because M_n increases linearly with N for constant access-rejection probability. (b) The 4-port starting point is below 0.1% of total 4-port cache area.

Simulated delay times and sequential gate-number (fan-in $\leqq$3) in the critical path are shown in Fig. 8. As expected, delay increases moderately, mainly due to larger capacitive loads and a small increase in sequential gate number.

Figure 9 compares design examples for 8 ports with PIH and fair algorithm. Such central arbiters are needed for the 8-port cache in the trade-off example of Fig. 1. Since the fair algorithm [6] allows a 2 times smaller M_n (i.e. a 1 bit shorter m_n), conflict-detection decreases in area, while conflict-regulation increases. Approximately equal very small areas of $0.27mm^2$ and $0.26mm^2$ result for these arbiter possibilities. This is only about 0.12% of the area estimated for the corresponding 8-port cache in the same technology. Delay times of about 2.5ns are small enough, to be largely hidden by parallel operation of central arbiter and address decoder.

5. Conclusion

High access-bit-rate, multi-ported caches will be needed in near future to maximize performance of single-chip, multi-issue processors. For such caches a fast, super-compact central arbiter is proposed and analyzed in a design study. The arbiter is applicable for any multi-port cache, but especially suited for a compact hierarchical architecture. Negligible area below 1% of total cache size is verified up to 32 ports. Delay

time increases only slightly with port number N and is short enough to be largely hidden from the critical access-path by parallel operation with address decoders. In summary, the proposed central arbiter is the most efficient solution for high bit-rate, multi-port caches.

References

[1] L. Hammond et al., IEEE Computer **30**, 79 (1997)
[2] Y. Tatsumi et al., Electronics Letters **35**, 2185 (1999)
[3] H.J. Mattausch, Electronics Letters **35**, 1441 (1999)
[4] H.J. Mattausch et al., Proc. 25th European Solid-State Circuits Conference (ESSCIRC'99), (1999) p. 126
[5] N.H.E. West and K. Eshraghian, Principles of CMOS VLSI Design (Addison-Wesley, 1993), p. 540
[6] H.J. Mattausch et al., Electronics Letters **34**, 861 (1998)

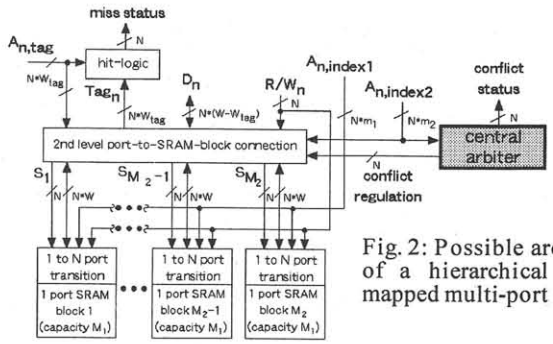

Fig. 2: Possible architecture of a hierarchical directly mapped multi-port cache.

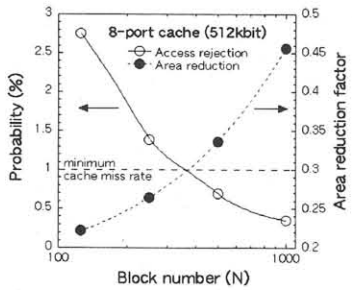

Fig. 1: Access-conflict probability and area-reduction for a directly-mapped hierarchical 8-port cache with 512Kbit data-capacity as a function of block number. Minimum cache-miss rate is shown for comparison.

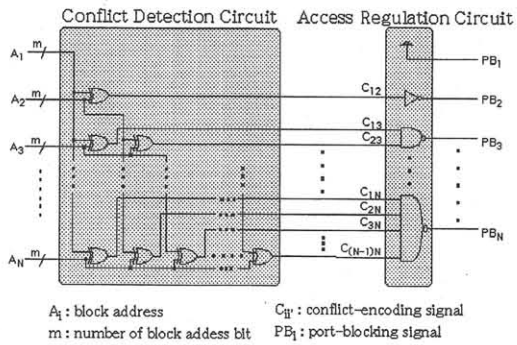

Fig. 3: Central arbiter for the port-importance-hierarchy (PIH) conflict-regulation algorithm.

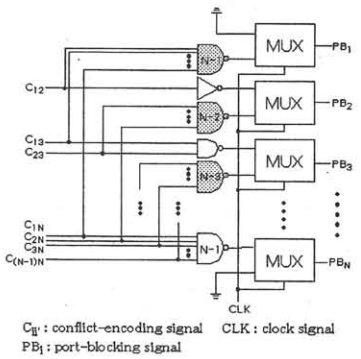

Fig. 4: Circuit for best-case fair conflict regulation.

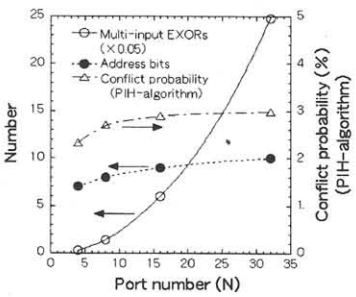

Fig.5: Block number M_n, address bits m_n and resulting conflict probability for the PIH algorithm as a function of port number N.

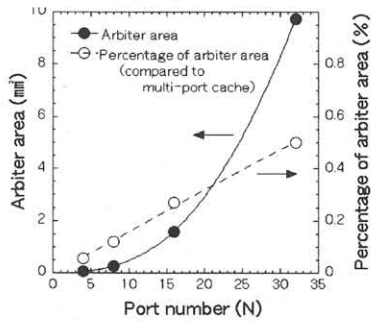

Fig. 7: Central-arbiter area and percentage of complete multi-port cache area.

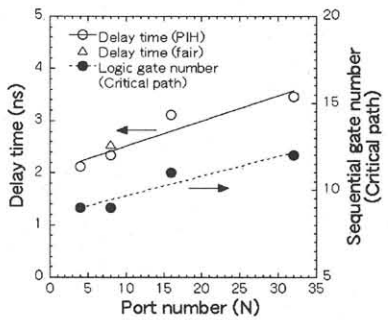

Fig. 8: Simulated delay times and number of sequential gates in the critical path (fan-in $\leqq 3$) for the layouts of Fig. 6 and Fig. 9.

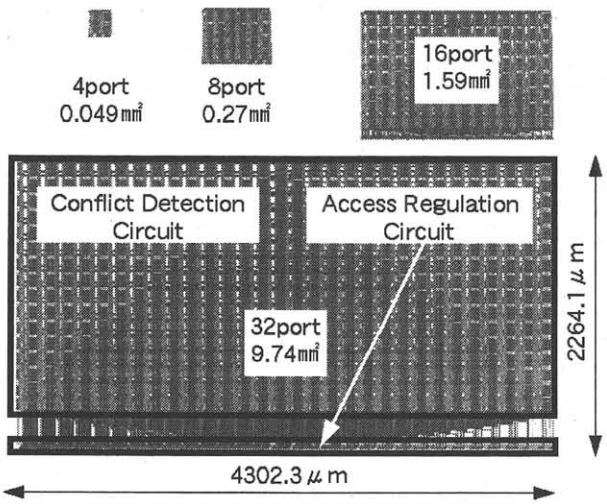

Fig. 6: Central-arbiter layouts with the PIH algorithm in a 0.5μm, 2 metal CMOS technology for 4, 8, 16 and 32 ports.

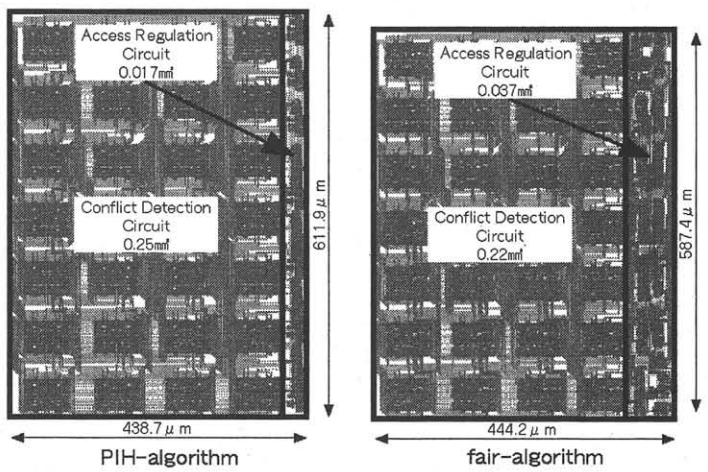

Fig. 9: Comparison of central-arbiter layouts with PIH-algorithm and fair algorithm in 0.5μm, 2 metal CMOS for a hierarchical 8-port cache.