A Low-cost Memory Architecture with NAND XIP for Mobile Embedded Systems

Chanik Park, Jaeyu Seo, Sunghwan Bae, Hyojun Kim, Shinhan Kim and Bumsoo Kim

Software Center, SAMSUNG Electronics, Co., Ltd.
Seoul 135-893, KOREA
{ci.park, pelican, sunghwan.bae, zartoven, shinhank, bumsoo}@samsung.com

ABSTRACT

NAND flash memory has become an indispensable component in mobile embedded systems because of its versatile features such as non-volatility, solid-state reliability, low cost and high density. Even though NAND flash memory is gaining popularity as data storage, it can be also exploited as code memory for XIP (execute-in-place). In this paper, we present a new memory architecture which incorporates NAND flash memory into an existing memory hierarchy for code execution. The usefulness of the proposed approach is demonstrated with real embedded workloads on a real prototyping board.

Categories and Subject Descriptors

B.3.2 [**Memory Structures**]: Design Styles – Cache memories; D.4.2 [**Operating Systems**]: Storage Management – Secondary storage; B.6.1 [**Logic Design**]: Design Styles – Memory control and access.

General Terms

Algorithms, Measurement, Performance, Design

Keywords

Memory architecture, NAND XIP, Priority-based caching

1. INTRODUCTION

A memory architecture design is a main concern to embedded system engineers since it dominates the cost, power and performance of mobile embedded systems. The typical memory architecture of mobile embedded systems consists of ROM for initial bootstrapping and code execution, RAM for working memory, and flash memory for permanent data storage. In particular, emerging memory technology, the flash memory, is becoming an indispensable component in mobile embedded systems due to its versatile features: non-volatility, solid-state reliability, low power consumption, etc.

CODES+ISSS'03, October 1-3, 2003, Newport Beach, California, USA

The most popular flash memory types are NOR and NAND. NOR flash is particularly well suited for code storage and execute-in-place (XIP) [1] applications, which require high-speed random access. While NAND flash provides high density and low-cost data storage, it is not applicable to XIP applications due to the sequential access architecture and long access latency

Table 1 shows different characteristics of various memory devices. Mobile SDRAM has advantages in performance but requires high power consumption over other memories. There is a tradeoff between low power SRAM and fast SRAM in terms of power consumption, performance and cost. In non-volatile memories, NOR flash provides fast random access speed and low power consumption, but the cost of NOR flash is high compared to NAND flash. Even though NAND flash shows long random read latency, it has advantages in power consumption, storage capacity, and erase/write performance in contrast to NOR flash.

Table 1. Characteristics of various memory devices. The values in the table were calculated based on SAMSUNG 2003 memory data sheets [1-2].

		Current (mA)		Random Access (16bit)		
Memory	$/Gb	idle	active	read	write	erase
Mobile SDRAM	48	0.5	75	90ns	90ns	N.A
Low power SRAM	320	0.005	3	55ns	55ns	N.A
Fast SRAM	614	5	65	10ns	10ns	N.A
NOR	96	0.03	32	200ns	210.5us	1.2 sec
NAND	21	0.01	10	10.1us	200.5us	2 ms

Even though NAND flash memory is widely used for data storage in mobile embedded systems, NAND flash memory researches for code storage are hardly found in industry or academia.

In this paper, we present a new memory architecture for NAND flash memory to provide XIP functionality. With XIP functionality in NAND flash, the cost of the memory system can be reduced since the NAND flash can be used not only as data storage but also as code storage for execution. As a result, we can obtain cost-efficient memory systems with reasonable performance and power consumption.

The basic idea of our approach is to exploit the locality of code access pattern and implement a cache controller for repeatedly accessed codes. The prefetching cache is used to hide memory latency resulting from NAND memory access. In this paper we concentrate on code execution even though data memory is also an important aspect of memory architecture. There are two major

[1] XIP is the execution of an application directly from the flash memory instead of having to download the code into the systems' RAM before executing it.

contributions in this paper. First, we demonstrate that the NAND XIP is feasible in real-life systems through a real hardware and commercial OS environment. Second, we apply highly optimized caching techniques geared toward the specific features of NAND Flash.

The rest of this paper is organized as follows. In the next section, we describe the trend of memory architecture in mobile embedded systems. Section 3 reviews related work in academia and industry. In Sections 4 and 5, we present our new memory architecture based on NAND XIP. In Section 6, we demonstrate the proposed architecture with real workloads on a hardware prototyping board and evaluate cost, performance and power consumption over existing memory architectures. Finally, our conclusions and future work are drawn in Section 7.

2. Motivational Systems: Mobile Systems

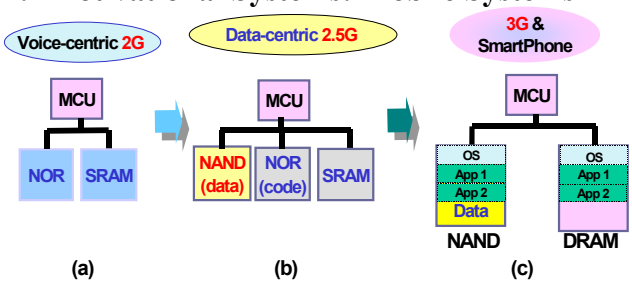


Figure 1. The memory architectures for mobile systems

Figure 1 shows the trend in memory architecture of mobile systems. The first approach is to use NOR for code storage and SRAM for working memory as shown in Figure 1(a). It is appropriate for low-end phones, which require medium performance and cost. However, as mobile systems evolve into data centric and multimedia-oriented applications, storage with high performance and huge capacity has become necessary. The second architecture (Figure 1(b)) seems to meet the requirements in terms of storage capacity through NAND flash memory, however its performance is not enough to accommodate 3G applications which consist of real-time multimedia applications. In addition, the increased number of components increases system cost. The third architecture (Figure 1(c)) eliminates NOR flash memory and uses only NAND flash memory by using a shadowing[2] technique. Copying all code into RAM offers the best performance possible, but contributes to the slow boot process. A large amount of DRAM is necessary to hold both the OS and the applications. The higher power consumption from power hungry DRAM memory is another problem for battery-operated systems. Thus, it is important to investigate an efficient memory system in terms of cost, performance and power consumption.

3. RELATED WORK

In the past, researchers have exploited NOR flash memory as caches for magnetic disks due to its low power consumption and high-speed characteristics. eNvy focused on developing a persistent storage architecture without magnetic disks [7]. Fred et al showed that NOR flash memory can reduce energy consumption by an order of magnitude, compared to magnetic disk, while providing good read performance and acceptable write performance [9]. B. Marsh et al examined the impact of using NOR flash memory as a second-level file system buffer cache to reduce power consumption and file access latency on a mobile computer [8].

Li-Pin et al investigated the performance issue of NAND flash memory storage subsystems with a striping architecture, which uses I/O parallelism [10].

In industry [5], NAND XIP is implemented using small size of buffer and I/O interface conversion, but the XIP area is limited to boot code. The OS and application codes should be copied to system memory at booting time.

In summary, even though several researches have been made to obtain the maximum performance and low power consumption from data storage, few efforts to support XIP in NAND flash are found in academia or industry.

4. NAND XIP ARCHITECTURE

In this section, we describe NAND XIP architecture. First, we look into the structure of NAND flash memory and illustrate the basic implementation of NAND XIP based on caching mechanism.

4.1 BACKGROUND

A NAND flash memory consists of a fixed number of blocks, where each block has 32 pages and each page consists of 512 bytes main data and 16 bytes spare data as shown in Figure 2. Read/write operation is performed on page basis. In order to read a page, command and address are inputted to NAND flash memory through I/O pins. After a fixed delay time, selected page is loaded into data and spare registers. Spare data can be used to store auxiliary information such as bad block identification and error correction code (ECC) for associated main data. NAND flash memories are subject to a condition called “bad block”, in which a block cannot be completely erased or cannot be written due to partial or 2-bit errors. Bad blocks may exist in NAND flash memory when shipped or may occur during operation. The system design must be able to mask out the invalid blocks via address mapping. For example, as shown in Figure 3, if block 2 and 5 are initially bad, the bad block numbers and the replacement block numbers are recorded in a bad block map table. The bad block map table is managed by hardware or software.

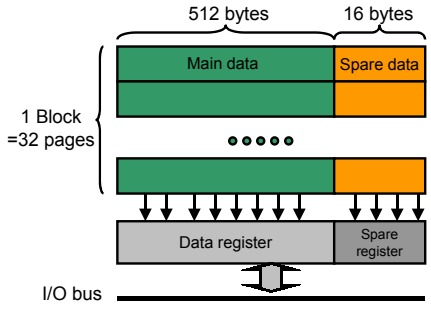


Figure 2. Structure of NAND flash memory

[2] During system booting time, entire code image is copied from permanent storage into systems’ RAM for execution.

In order to implement the NAND XIP, we should consider the following points.

- Average memory access time
- Worst case handling
- Bad block management

The performance of memory system is measured by average access time [3]. In order to implement XIP functionality, the average access time of NAND flash memory should be comparable to that of other memories such as NOR, SRAM and SDRAM. Though average memory access time is a good metric for performance evaluation, worst-case handling, or cache miss handling is another problem in practical view since most mobile systems such as cellular phones include time-critical interrupt handling for call processing. For instance, if the time-critical interrupt occurs during cache miss handling, the system may not satisfy the given deadline and to make it worse, it may lose data or connection. Another consideration in NAND XIP is to manage bad blocks because bad blocks cause discontinuous memory space, which is intolerable for code execution.

4.2 BASIC IMPLEMENTATION

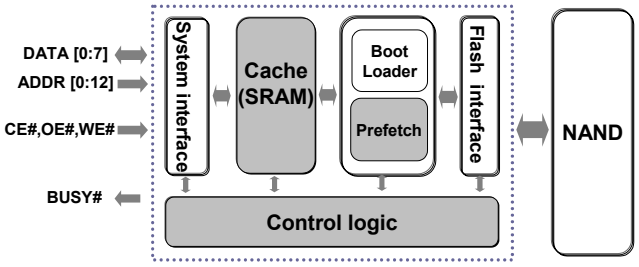


Figure 3. NAND XIP controller

The proposed architecture consists of a small amount of SRAM for cache, interface conversion logic, the control logic and NAND flash memory as shown in Figure 3. Interface conversion is necessary to connect the I/O interface of NAND flash memory to memory bus. For cache mechanism, direct map cache with victim cache is adopted based on Jouppi's work in [4] with optimization for NAND flash. In [4], the victim cache is accessed on a main cache miss; if the address hits the victim cache, the data returned to the CPU and at the same time it is promoted to the main cache; the replaced block in the main cache is moved to the victim cache, therefore performing a "swap". If the victim cache miss also occurs, NAND flash memory access is performed; the incoming data fills the main cache, and the replaced block will be moved to the victim cache. In next section, we modify the above "swap" algorithm using system memory and *page address translation table (PAT)*. The prefetching cache is used to hide memory latency resulting from NAND memory access. Several hardware prefetching techniques can be found in literature [13]. In our case, prefetching information is analyzed through profiling process and the prefetching information is stored in spare data at code image building time.

5. INTELLIGENT CACHING: PRIORITY-BASED CACHING

Though the basic implementation is suitable for application code which shows its spatial and temporal localities, it may be less effective in systems code which has a complex functionality, a large size, and interrupt-driven control transfers among its procedures [14]. Torrellas et al. presented that the operating system has the characteristics which large sections of its code are rarely accessed and suffers considerable interference within popular execution paths [14]. For example, periodic timer interrupt, rarely-executed special-case code and plenty of loop-less code disrupt the localities. On the other hand, real-time applications should be retained as long as possible to satisfy the timing constraints[3]. In this paper, we distinguish the different cache behavior between system and application codes, and adapt it to the page-based NAND architecture. We apply profile-guided static analysis of code access pattern.

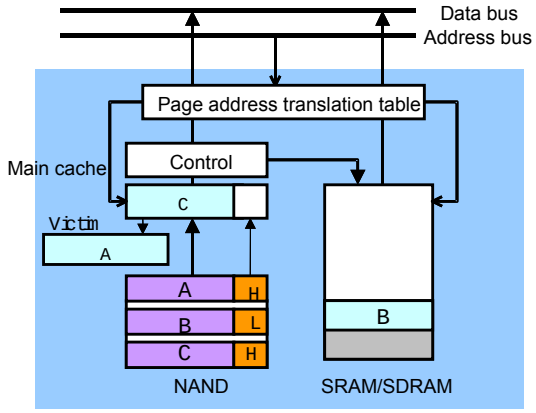


Figure 4. Priority-based cache architecture

We can divide code pages into three categories depending on their access cost: high priority, mid priority and low priority pages. Even though the priority can be determined by various objectives, we set the priority to pages based on the number of references to pages and their criticality. For example, if a specific page is referenced more frequently or has time critical codes, it is classified as a high-priority page and should be cached or retained in cache to reduce the later access cost in case that the page is in NAND flash memory. OS-related code, system libraries and real-time applications have high-priority pages. On the other hand, mid-priority page is defined to be normal application code which is handled by normal caching policy. Finally, low-priority page corresponds to sequential code such as initialization code, which is rarely executed. PAT is introduced to remap pages in bad blocks to pages in good blocks and to remap requested pages to swapped pages in system memory. We illustrate the caching mechanism in detail in Figure 4. First, when page *A* with high-priority is requested, it is cached from NAND flash to main cache. Next, when page B is requested from the CPU, it should be moved to main cache or system memory. Here assuming that page B is in conflict with page A, page B is moved to system memory (SRAM/SDRAM) since page B is low priority page ("*L"* in spare area of NAND flash memory means *low-priority*). At the same time, PAT is updated so that later access to page B is referred to system memory. Again, when page C is requested and in conflict with page A, page C replaces page A and page A is discarded from main cache since C's priority is high. The evicted page A is

[3] In this paper, real-time applications indicate multimedia applications with soft real-time constraints.

moved to victim cache. In summary, on NAND flash's page demand, the controller discards or swaps existing cache page according to the priority information stored in spare area data. The detail algorithm is explained in Figure 5. Another usage of spare area data is to store prefetching information based on profiling information gathered during code execution. This static prefetching technique improves memory latency hiding without miss penalty from miss-prediction at run-time.

```
Priority Caching (address)
{
   page = convert(address);
   if (isInPAT(page))
      main memory hit;
   else if(isInMainCache(page))
      Cache hit;
   else if(isInVictimCache(page))
      Victim hit;
   else {  // miss, fetch a page from NAND flash memory

       page_priority = NAND[page].priority;
      if (cache[page%CACHE_SIZE].priority ==HIGH)
        NAND[page] → main memory;
      else if (cache[page%CACHE_SIZE].priority == MID)
      {
         cache[page%CACHE_SIZE] → victim;
         NAND[page] → cache[page%CACHE_SIZE];
      }
      else
          NAND[page] → cache[page%CACHE_SIZE];
   }
}
```

Figure 5. Priority-based cache algorithm

6. EXPERIMENTAL SETUP

This section presents our experimental environment. Our environment consists of a prototyping board, our in-house cache simulator with pattern analysis and a real workload, or PocketPC[6]. The prototyping board has ARM9-based micro-controller, SRAM, SDRAM, NOR flash and NAND memories as shown in Figure 6. It is used not only to implement a real cache configuration on FPGA but also to gather memory address trace from running applications for cache simulator. The specification of main processor and NAND flash memory is shown in Table 2. The cache simulator explores various parameters such as miss rate, replacement policy, associativity and cache size based on memory traces from the prototyping board. The real embedded workload, PocketPC supports XIP-enabled image based on the existing ROMFS file systems in which each application can be directly executed without being loaded into RAM.

Table 2. The specification of the prototyping board

Parameter	Configuration
CPU clock	200 MHz
L1 Icache	64 way, 32byte line, 8KB
Bus width	16 bit = 1 word
NAND read initial latency	10 us/512byte
NAND serial read time	50 ns/word
SRAM read time	10 ns/word
SDRAM read time	120 ns/4word
NOR read time	210 ns/4word

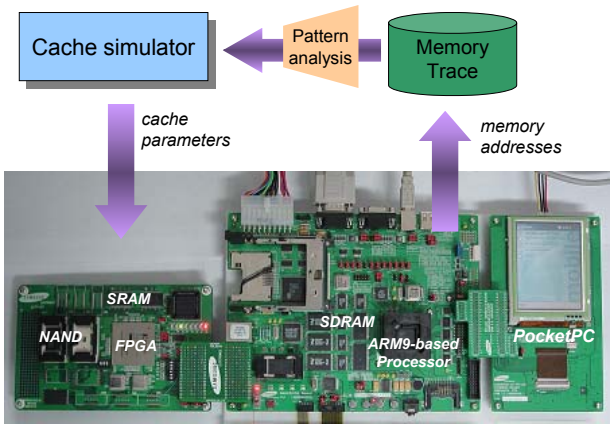


Figure 6. Prototyping board for NAND XIP

6.1 PERFORMANCE MODELING

In general, the memory system performance is evaluated by the average access time:

*Average memory access time = hit time + miss rate * miss penalty* . (1)

Hit time is the time to fill the L1 cache (32byte line) and the miss penalty is the time to fill the L2 cache from NAND flash memory.

The total energy consumed by the system is the sum of energies consumed by system components such as processor, bus, memory and the DC-DC converter. In this paper, we concentrate on the energy consumption of the processor and memory system:

$$E_{system} = E_{CPU} + E_{memory} . \quad (2)$$

The energy consumption of CPU is divided into two parts:

$$E_{CPU} = P_{CPU\_active} \times T_{CPU\_active} + P_{CPU\_idle} \times T_{CPU\_idle} . \quad (3)$$

T_{CPU_active} is assumed to be constant regardless of memory system since the number and execution time of instructions are fixed. Thus, we are concerned about T_{CPU_idle} which is proportional to off-chip memory access time.

In memory case, the number of accesses to the memory is directly proportional to energy consumption. The unit access energy consumption can be estimated from the active power specified in the data sheet [1-2]:

$$E_{memory} = E_{access} \times N_{access} . \quad (4)$$

The cost of memory system is computed from parameters shown on Table 1:

$$Cost_{memory\_system} = \sum_{m}^{N} density_m \times unit\ cost_m + Cost_{controller} . \quad (5)$$

The $Cost_{controller}$ is computed based on the synthesis result in Xilinx FPGA: totally 270,000 gates including 4KB and using off-chip SRAM for TAG and cache memory.

Finally, booting times are measured using real-time measurement tools.

6.2 EXPERIMENTAL RESULTS

In Figure 7, we compare the miss ratio over various configuration parameters such as associativity, replacement policy and cache size. We collected address traces from PocketPC while we were executing various applications such as "Media Player" and "MP3 player" since they are popular embedded multimedia applications which involve real-time requirements.

Note that the cache size is the most important factor to affect miss ratio as shown in Figure 7.

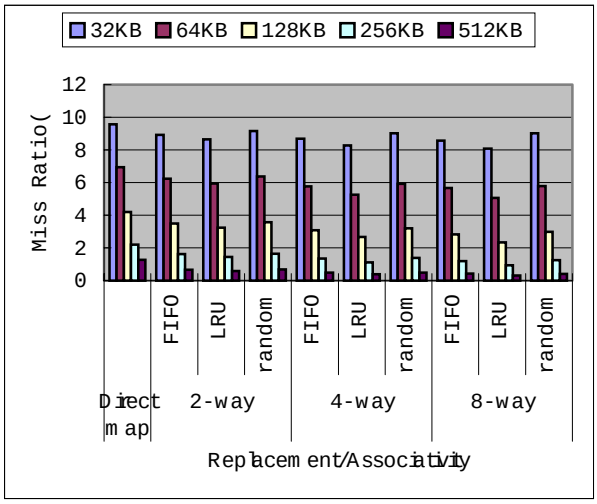
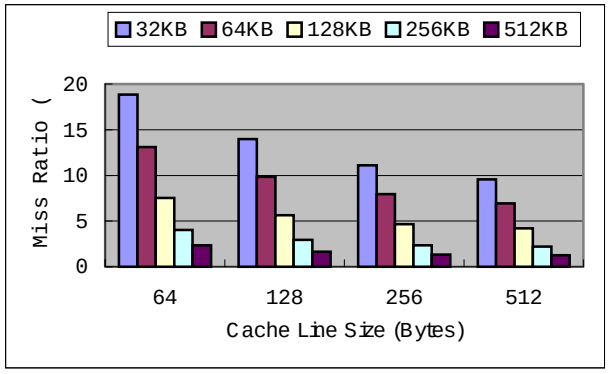
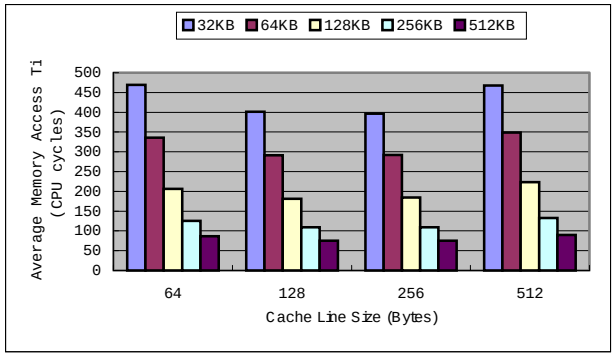


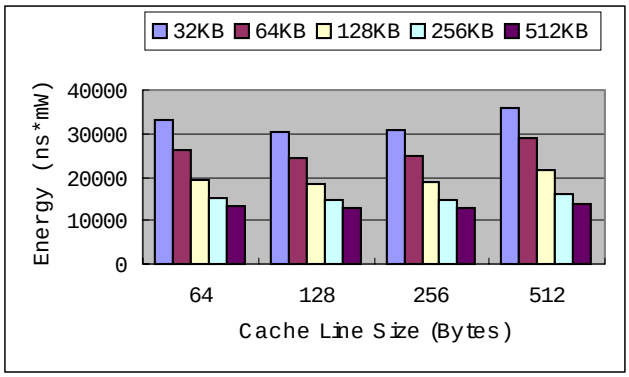
Figure 7. Associativity with different replacement policies and cache sizes versus miss ratio



(a)



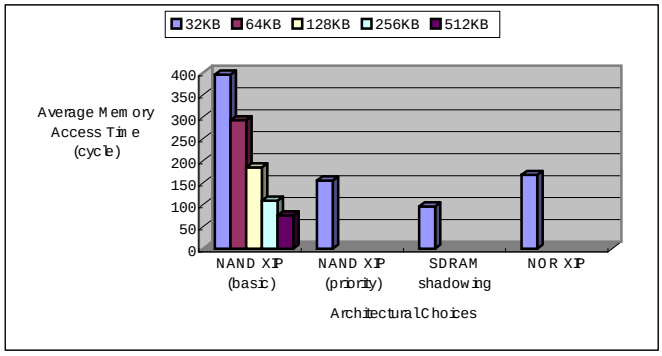
(b)



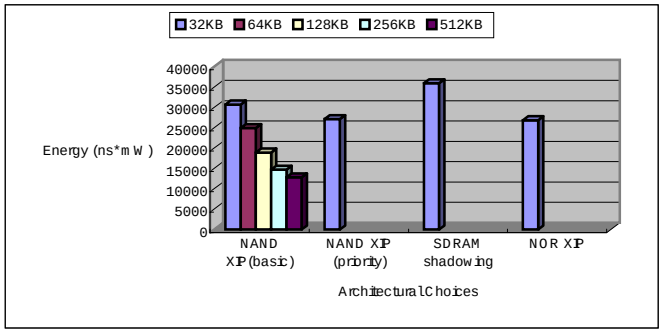
(c)

Figure 8. Cache line size versus (a) miss ratio, (b) access time per 32-byte and (c) energy consumption

To analyze the optimal cache line size in NAND XIP cache, simulation has been done with the memory trace which is gathered from the prototyping board. The line size of 256-byte shows better numbers in average memory access time and energy consumption over all other cache sizes as shown in Figure 8. Therefore, the line size of NAND XIP controller is determined to be 256-byte hereafter.



(a)



(b)

Figure 9. Overall performance comparison of different memory architectures: (a) average memory access time and (b) energy consumption.

Architectural Choices	Components	Booting Time (sec)	Cost ($)
NOR XIP	NOR(64MB)+SDRAM(32MB)	13	60
SDRAM Shadowing	NAND(64MB)+SDRAM(64MB)	15	35
NAND XIP (basic)	NAND XIP(64MB)+SRAM(256KB) Controller+SDRAM(32MB)	14	26
NAND XIP (priority)	NAND XIP(64MB)+SRAM(64KB) Controller+SDRAM(32MB)	14	24

Figure 10. Overall booting time and cost comparison of different memory architectures.

Figure 9 and 10 show overall performance comparison among different memory architectures based on our prototyping results. NOR XIP architecture (NOR+SDRAM) shows fast boot time and low power consumption at high cost. Even though SDRAM shadowing architecture (NAND + SDRAM) achieves high performance, it suffers from a relatively long booting time and inefficient memory utilization. Finally, our approach NAND XIP shows reasonable booting time, performance and power consumption with outstanding cost efficiency. Two implementations of NAND XIP (basic and priority) are also compared. The NAND XIP (basic) is composed of 256KB cache and 4KB victim cache. On the other hand, the NAND XIP (priority) has 64KB cache, 4KB victim and 2KB PAT. The NAND XIP (priority) has advantage of cache size reduction with the assistance of system memory and priority based caching.

6.3 Worst Case Handling

Even though NAND XIP offers an efficient memory system, it may suffer from worst-case handling, namely cache miss handling. A straightforward solution is to hold the CPU until the requested memory page arrives. This can be implemented using wait/ready signals' handshaking method. However, the miss penalty, 35us[4] is not trivial time especially in case that a fast processor is used. Besides CPU utilization problem, if the time-critical interrupt request occurs to the system, the interrupt may be lost because the processor is waiting for memory's response. In order to solve this problem, a system-wide approach is needed. First, OS should handle the cache miss as a page fault as in virtual memory management. The CPU also should supply "abort" function to restart the requested instruction after cache miss handling. In analytical approach view, task timing analysis and schedulability test will be helpful in the domain of real-time[12].

7. CONCLUSIONS

We presented the new low cost memory architecture with NAND XIP functionality which enables NAND flash memory to be used as code storage as well as data storage. In order to hide a long read access latency, we applied the priority-based caching mechanism geared for NAND specific features. As a result, we can accomplish cost and power efficient memory systems with reasonable performance and booting time. As future work, compiler's assistance will be helpful to improve NAND XIP in embedded memory architectures. Code packing and positioning will be achieved in near future.

[4] It is calculated as sum of NAND initial delay (10us) + page read (512 x 50ns).

REFERENCES

[1] Samsung Electronics Co., "NAND Flash Memory & SmartMedia Data Book", 2002.

[2] http://www.samsung.com/Products/Semiconductor/index.htm

[3] J.L. Hennessy and D.A. Patterson, "Computer Architecture: A Quantitative Approach", 2nd edition, Morgan Kaufman Publishers, 1996.

[4] N. Jouppi, "Improving Direct-Mapped Cache Performance by the Addition of a Small Fully-Associative Cache and Prefetch Buffers," ISCA-17:ACM/IEEE International Symposium on Computer Architecture, pp. 364-373, May 1990.

[5] http://www.m-sys.com/files/documentation/doc/Mobile_DOC_G3_DS_Rev1.0.pdf

[6] Microsoft Co., "Pocket PC 2002", http://www.microsoft.com/mobile/pocketpc

[7] M. Wu, Willy Zwaenepoel, "eNVy: A Non-Volatile, Main Memory Storage System", Proc. Sixth Int'l Conf. On Architectural Support for Programming Languages and Operating Sys. (ASPLOS VI), San Jose, CA, Oct. 1994, pp. 86-97.

[8] B. Marsh, F. Douglis, and P. Krishnan, "Flash memory file caching for mobile computers," Proc. 27th Hawaii Conf. On Sys. Sciences, Wailea, HI, Jan. 1993, pp. 451-460.

[9] F. Douglis et al., "Storage Alternatives for Mobile Computers", Proc. First USENIX Symp. On Operating Sys. Design and Implementation, Monterey, CA, pp 25-37, Nov. 1994.

[10] Li-Pin Chang, Tei-Wei Kuo, "An Adaptive Striping Architecture for Flash Memory Storage Systems of Embedded Systems," The 8th IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS 2002) September, 2002. San Jose, California.

[11] J. R. Lorch and A. J. Smith, "Software Strategies for Portable Computer Energy Management", IEEE Personal Communications, June, 1998.

[12] C.-G. Lee, K.-P. Lee, J. Hahn, Y.-M. Seo, S.L. Min, R. Ha, S. Hong, C.Y. Park, M. Lee, and C.S. Kim, Bounding Cache-related Preemption Delay for Real-time Systems, IEEE Transaction on Software Engineering, vol 27, no 9, pp805-826, 2001.

[13] C. Young, E. Shekita, "An Intelligent I-Cache Prefetch Mechanism," in Proceedings of the International Conferences on Computer Design, pp. 44-49, Oct 1993.

[14] Josep Torrellas, Chun Xia, and Russell Daigle, "Optimizing Instruction Cache Performance for Operating System Intensive Workload", Proc. HPCA-1: 1st Intl. Symposium on High-Performance Computer Architecture, p.360, January 1995.

[15] Zhang Yi, Steve Haga and Rajeev Barua, "Execution History Guided Instruction Prefetching", Proceedings of the Sixteenth ACM Int'l Conference on Supercomputing (ICS), New York City, NY, June, 2002.