Improved memory throughput using serial NOR flash - part 1

Cliff Zitlaw, Spansion Inc. - June 11, 2012

Editor’s note: This work was first presented at the Embedded Systems Conference (ESC) 2012. You can register to view other papers from the proceedings here. For more information about ESC 2013 (San Jose, CA; April 22-25), click here.

NOR flash memory remains the preferred non-volatile technology for discrete memories in embedded systems. For some applications, NOR usage is migrating away from the parallel NOR bus to products based on the lower-pin-count serial peripheral interface (SPI) to optimize the memory subsystem. Many systems using NOR-based SPI memories have reached 108 MHz bus clock rates using a QuadIO (x4) interface to achieve a 54 MB/s sustained read throughput while retaining compatibility with the original interface specified more than 25 years ago. The latest SPI-NOR product releases are incorporating a DDR-read operation and reaching throughputs of up to 80 MB/s. As system-level read throughput requirements continue to increase, a number of strategies have been deployed to optimize the SPI interface for higher performance. This article describes both system-level and memory-device strategies that have been deployed to allow higher SPI read throughputs.

History
SPI was introduced in the early 1980s by Motorola and continues to be popular throughout the embedded market. The original interface (see figure 1) has an efficient four-pin (CS#, SCK, SI, SO) bus structure that rapidly achieved market dominance over the proprietary Inter-Integrated Circuit (I²C) bus from Philips Semiconductor and Microwire interface from National Semiconductor introduced during the same period. Initial EEPROM products for all three of the interfaces used low-cost eight-pin DIP and SOIC packages.

![Figure 1: The pinout for the serial peripheral interface (SPI, left) acquired more market share than the Inter-Integrated Circuit (I²C) bus (center) or the Microwire interface (right).](Click image to enlarge)
The early EEPROM offerings had limited operating frequencies by today’s standards: 1 MHz for SPI and Microwire, and 100 KHz for I²C. While the initial offerings had modest clock rates, by the time NOR flash had emerged as the dominant non-volatile memory (NVM) technology in the mid 1990s, it was clear that read throughputs would become a first order value proposition.

The SPI bus has evolved over time, improving read throughputs by increasing both clock rates and bus width. Bus width increased from the original x1 to x2, and finally a x4 interface (see figure 2). The evolution of the SPI bus width has been accomplished by retasking the original six SPI bus signals to support the x2 and the x4 bus widths without increasing the overall pin count. Today’s NOR-based QuadIO SPI memories commonly reach 108 MHz clock rates to achieve a 54 MB/s sustained read throughput. Clock rates of 133 MHz are starting to appear on some SPI memories. The SPI bus interface is used on virtually all higher-density NOR flash offerings.

Comparison with alternative NOR flash memory interfaces

Two of the more significant criteria used when considering NOR flash devices are the sustained read throughput and the number of pins used for the bus interface. Figure 3 compares different NOR flash bus interfaces and their respective read throughputs. The legacy parallel interfaces (async, page, and ADP burst) all require more than 40 pins with an address/data multiplexed version of the parallel bus (ADP) having around 30 pins. Historically, the parallel interfaces have provided higher read throughputs, but recent advances in SPI bus performance have significantly closed the throughput gap.

The read throughput comparison in figure 3 shows that SPI NOR exceeds the abilities of the asynchronous NOR bus and approaches the level of the page mode interface. This comparable performance to async and page products is especially relevant for the many embedded chipsets that do not support either of the burst NOR interfaces. Recent product announcements of 133 MHz QuadIO (66 MB/s) and 80 MHz DDR-QuadIO (80 MB/s) tilt the performance in favor of the SPI
interface wherever the parallel burst and page mode interfaces are not available. **System-level improvements**

**Direct CPU read accesses**

Host systems have historically interfaced with SPI flash memories through an integrated peripheral SPI controller. To accomplish a read operation, the host software will:

1. Load the target address into the controller address register
2. Load the read command into the command register
   a. This loading would automatically initiate the read transaction on the SPI bus
3. Poll the SPI controller status register until the target data has been output by the SPI memory and captured by the host SPI controller
4. Extract the target data from the data buffer in the SPI controller

This process is dramatically slower than what occurs during a read from a parallel flash memory device where the external memory is mapped directly into the CPU address space. In the parallel-usage case, the CPU simply reads from the target address, and the data is returned without the need for additional software intervention. This process is a requirement for XiP operation where op-code fetches must be performed directly for efficient code execution.

Several recently released system-on-chip (SoC) products have included a bimodal SPI controller strategy that maintains the legacy peripheral access infrastructure while also allowing read operations to be performed directly from the CPU address map. This invention has led to the rapid adoption of direct execution of code out of SPI memories. Direct mapping into the CPU memory map eliminates throughput bottlenecks that exist during transfers through the SPI peripheral controller.

**Dual SPI interfaces**

In a few high-performance applications, the need for increased read throughput and high-density support has caused SoC manufacturers to develop dual-channel QuadIO SPI interfaces (see figure 5). The two channels are operated simultaneously in QuadIO mode to double the read throughput while only increasing the interface by six signals. The increased pin count (from six to 12 pins) is still much lower than the 40+ signals that would be required for a parallel NOR interface.
Device-level improvements
Device level behavior has also been pressured by the need for higher read throughputs. The first order requirement of increasing the clock rate has been implemented but a number of other parameters and characteristics have also received attention to make the higher data rates viable in a real world system.

Bus-timing enhancements
Two of the more significant timing parameters that have received scrutiny are the clock to data valid time (tV) and the data hold after clock time (tHO). These two parameters describe how long data is valid on the bus. tV describes when data becomes valid after a clock edge, and tHO describes how long data will remain valid after a clock edge. Much work has been done to minimize tV and maximize tHO to extend the data valid period for operation at higher clock frequencies (shorter clock periods).
**Shift from 3 V to 1.8 V operating voltages**

Device operating voltages are starting to transition from the legacy 3 V supply to the 1.8 V voltage level. This migration has been largely driven by the use of SPI devices in cell phones, where the lower operating voltage is attractive from a power-consumption perspective. A peripheral advantage to the lower operating voltage is that signal swings on the bus interface are reduced when operating with lower voltages. This smaller voltage swing between logic states means shorter transition times, which are necessary at higher operating frequencies. Current 1.8 V offerings are starting to appear with operating frequencies of 133 MHz with the possibility of even higher frequencies.

**Output drive strength control**

One emerging trend to maximize signal integrity is to allow the output drive strength to be optimized in the target environment. This capability has long been part of the high-speed DRAM world and is essential to maximizing signal integrity at higher data rates. Typical implementations provide four settings that are configurable in-system by the host processor. Environmental problems related to capacitive loading, trace impedance, and trace length can be mitigated by adjustment of the output drive strength. Figure 7 shows the output drive capabilities from one of the devices that support drive strength control.

---

Did you find this article of interest? Then visit the Memory Designline, where we update daily with design, technology, product, and news articles tailored to fit your world. Too busy to go every day? Sign up for our newsletter to get the week's best items delivered to your inbox. Just click here and choose the "Manage Newsletters" tab.