Unleashing the Power of Embedded DRAM

by Peter Gillingham, MOSAID Technologies Incorporated
Ottawa, Canada

Abstract

Embedded DRAM technology offers many advantages in System On Chip products. Computing applications demand memory with low latency and zero soft error rate. Graphics and networking need high bandwidth. Mobile applications require extremely low power. All applications benefit from the high density afforded by embedded DRAM technology. A single DRAM architecture cannot provide an optimal solution across the full range of applications. This paper describes in detail the advantages of embedded DRAM technology over external memory and embedded SRAM, and presents three 90nm embedded DRAM architectures optimized for specific applications.

Embedded DRAM Value Proposition

DRAM technology offers the highest density random access memory due to a simple 1T1C structure consisting of a single access transistor and a single storage capacitor. Typical 90nm embedded DRAM processes offer cell sizes in the range of 0.2?m2. In contrast, 90nm SRAM cells typically occupy around 1.0?m2 due to a more complex 6T structure comprised of cross-coupled latch and dual access devices. Taking into account the more extensive peripheral circuitry associated with DRAM, the overall density of a completed embedded DRAM marcocell is about four times that of the SRAM alternative.

In DRAM a read operation is destructive since the charge on the storage capacitor is shared with the larger bitline capacitance. This data must be sensed and fully restored to the cell capacitors during each memory cycle. In a typical DRAM row operation, many thousands of bits are sensed and are available for use on-chip. In commodity DRAM, only a small fraction of the bits which are sensed in each row cycle are made available off-chip. In a 256M DDR2 SDRAM with a x16 400Mb/s/pin interface for example, 8192

	High Bandwidth	High Speed	Low Power
Capacity	16Mb	2Mb	8Mb
Interface	2048bit synchronous	128bit synchronous	64bit asynchronous
Array Size	256r x 1024c	128r x 256c	256r x 512c
Page Size	8192bit	512bit	512bit
# of Banks	2 banks	16 banks	1 bank
Macro Size	6.0mm2	1.4mm2	3.3mm2
Density	2.69Mb/mm2	1.32Mb/mm2	2.45Mb/mm2
Data Rate	500MHz	1GHz	166MHz
Cycle Time	16ns	4ns	6ns
Bandwidth	128GB/s	16GB/s	1.33GB/s
Supply	1.2v	1.2v	1.0v (0.7v in standby)
Active Power	1.2W	220mW	25mW
Standby Power	1mW	66mW	35μW

Table 1. Summary of Optimized 90nm Embedded DRAM Macrocells

bits are sensed and made available internally. However, only 384 bits, less than 5% of the total number available, can be transferred through the interface within a 60ns row cycle. Even the state-of-the-art DDR2 interface represents a bottleneck to the massive internal bandwidth of the DRAM core.

Combining the memory with the logic that processes the data on a single die eliminates the I/O bottleneck and provides dramatically improved bandwidth. Applications such as graphics or networking, where large contiguous blocks of information must be transferred, can take full advantage of wide internal datapaths. In these applications latency is not a primary concern. There are also significant power savings in eliminating first the high speed terminated I/O lines connecting memory and logic chips, and second the DLL and internal mux/demux datapaths within the chips themselves.

Embedded DRAM can be optimized for low latency applications such as program, data, or cache memory in embedded microprocessor or DSP chips. With appropriate memory architecture and circuit design, GHz speeds are possible with on-chip DRAM. For many embedded applications the entire memory requirement can be implemented on chip, eliminating the latency added by the interface to off-chip memory. Although by nature DRAM is fundamentally slower than SRAM, the overall performance benefit of having four times the memory in the same area on-chip next to the processor more than compensates for the disadvantage of slightly slower internal memory operation. For an equivalent capacity SRAM, the longer interconnections required create additional wire delay that offsets any internal speed advantage.

Embedded SRAM is facing two serious problems in deep submicron geometries. The first is standby power dissipation from subthreshold leakage and gate tunneling. With 1.0v supply operation the difference between "on" and "off" states in the transistor has now become quite small. A fast process with high current drive capability will have significant subthreshold current in the nominally "off" state with zero gate-source voltage. With gate oxide thicknesses now reduced to just a few atom's thickness, direct gate tunnelling from gate to channel is increasing exponentially. Each 6T SRAM cell has several subthreshold and gate tunnelling leakage paths. A 1Mbit SRAM implemented in standard 90nm process will consume more than 1mA in standby as a result of these two leakage mechanisms. In contrast, a DRAM cell has virtually no leakage. The current required for periodic refresh of a 90nm DRAM array combined with the leakage currents in the peripheral circuits is several orders of magnitude lower than the equivalent size SRAM leakage. In low power mobile applications where standby battery life is critical, embedded DRAM offers significant advantages.

As the SRAM cell has been scaled down to 90nm geometries, the node capacitance within the cross coupled latch has decreased to sub-fF levels. As a result, the cell is highly susceptible to soft errors from alpha particles or cosmic rays. In contrast, DRAM cells typically have 20fF or more in storage capacitance and are virtually immune to soft errors. To address this problem in SRAM, parity bits and error detection and correction circuitry must be employed, adding further area overhead, power dissipation, and latency. Embedded DRAM provides a more robust and reliable solution.

DRAM Architecture Basics

Figure 1 shows the core circuitry of a DRAM, including cells, sense amplifier, and datapath. In the folded bitline architecture a wordline connects a memory cell to either BL or BL* in each column of the array. The other bitline acts as a reference during sensing. Wordlines WL are driven to Vpp, a voltage supply higher than Vdd, to allow a full Vdd level to be written to the cell capacitor through the n-channel access transistor. Here, the sense amplifier is shared between top and bottom arrays, as selected by an isolation signal ISO which is also driven to a Vpp level. Prior to sensing, the bitlines and sense amplifiers are precharged to Vdd/2 by the equalization signals EQ. At the beginning of an active row cycle, the appropriate EQ signals are de-asserted and a single wordline is raised to Vpp sharing the charge in the cell capacitor with the bitline capacitance. The bitline to cell capacitance ratio is typically 5:1, so the cell signal is attenuated significantly. Sense amplifiers are then activated by simultaneously ramping sense clocks PR and PS* from the Vdd/2 precharge state to Vdd and Vss respectively, to power the cross coupled latch. Depending on the polarity of the data, one of the bitlines will swing to Vdd and the other to Vss. This restores the full data to the memory cell. In Figure 1 a differential bidirectional databus DB/DB* is shared by two columns of sense amplifiers. In a column read operation the previously precharged databus is connected to the sense amplifier selected by the column address Yj to drive data to the I/O circuitry at the edge of the array. In a column write operation the databus overpowers the selected sense amplifier.

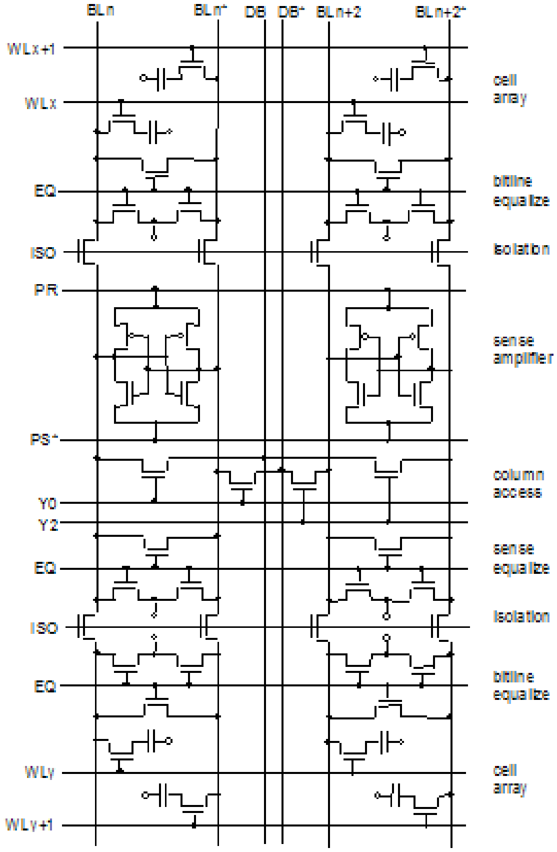

Figure 1. Shared Sense Amplifiers with Wide Databus and Folded Bitline DRAM Cell Array

Although the databus is shown running between the 2 columns, it actually runs over the bitlines in a higher level of metal so it does not cost any additional chip area. Typical commodity DRAM databusses run parallel to the wordlines through the sense amplifier area as shown in Figure 2. Multiple arrays are activated and a small number of bits are provided by each array. This provides sufficient data to feed the limited I/O bandwidth of a standalone memory chip, but only makes use of a small portion of the bits activated during the row cycle.

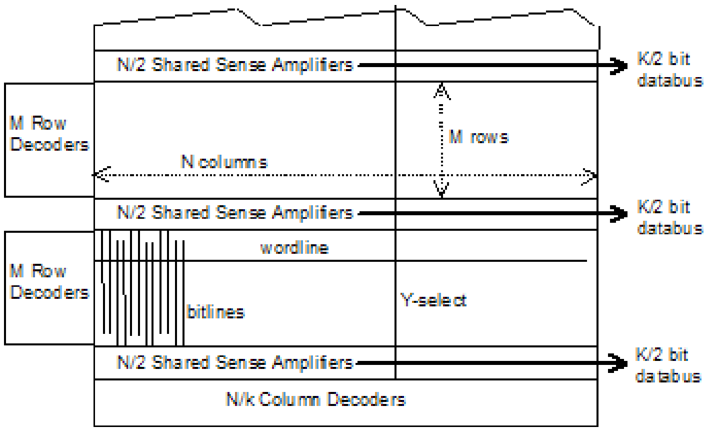

Figure 2. Conventional Databus Architecture

The wide databus architecture in which the databusses run parallel to bitlines is shown at a higher level of abstraction in Figure 3. This architecture is well suited to embedded applications because a significant fraction of the total number of bits sensed in a single row cycle can be accessed outside the memory array. Row decoders drive M wordlines or rows in each array. Multiple arrays are stacked up to share databusses and I/O circuitry. In an array with N columns or bitline pairs, N/2 sense amplifiers are staggered on either side of the array to relax the pitch requirement to 4 bitlines. Because the sense amplifiers are shared, adjacent arrays cannot be activated at the same time. This architecture supports very large data widths for high bandwidth. With aggressive layout it is possible in most target processes to have as many as N/2 databusses running over the array. In practice it is more efficient to have N/4 which would also permit separate read and write databusses for higher speed operation.

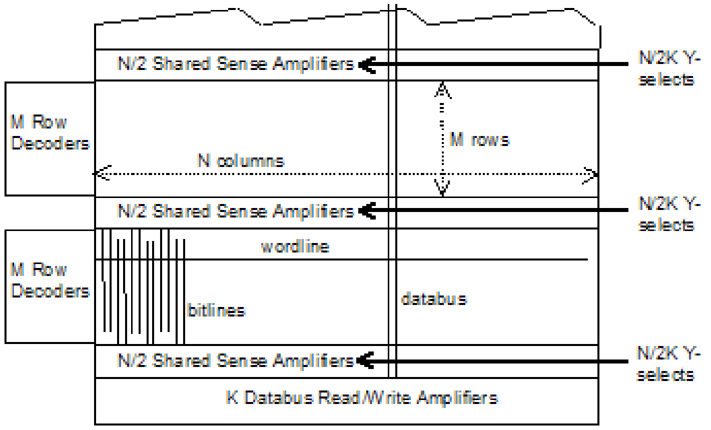
N/2 Shared Sense Amplifiers
N/2K Y-selects
M Row Decoders
M rows
N columns
N/2 Shared Sense Amplifiers
N/2K Y-selects
M Row Decoders
wordline
bitlines
databus
N/2 Shared Sense Amplifiers
N/2K Y-selects
K Databus Read/Write Amplifiers

Figure 3. Wide Databus Architecture

The wide databus architecture serves as the basis for a variety of embedded DRAM macrocells meeting very different requirements according to the end application. Here we will explore 3 different configurations: high bandwidth, high speed with low latency, and low power. For purposes of illustration, a state-of-the-art 90nm embedded DRAM process will be assumed.

High Bandwidth Macrocell

The wide databus architecture is ideally suited to latency tolerant applications that require high bandwidth. The dimensions of individual memory cell arrays are limited by performance considerations. To maintain an adequate cell to bitline capacitance ratio for reliable sensing, the number of wordlines in the array is limited to 256. The RC time constant of the wordline in an array with 1024 columns will allow a complete row cycle to be completed in 16ns, significantly faster than commodity DRAM. Note that actual cell arrays will include several additional rows and columns for redundant elements used to replace faulty bits detected during manufacturing test. Databus loading limits the number of arrays that can share common datapath circuitry. For 500MHz column cycle operation, eight arrays can be stacked to create a 4Mb segment as shown in Figure 4.

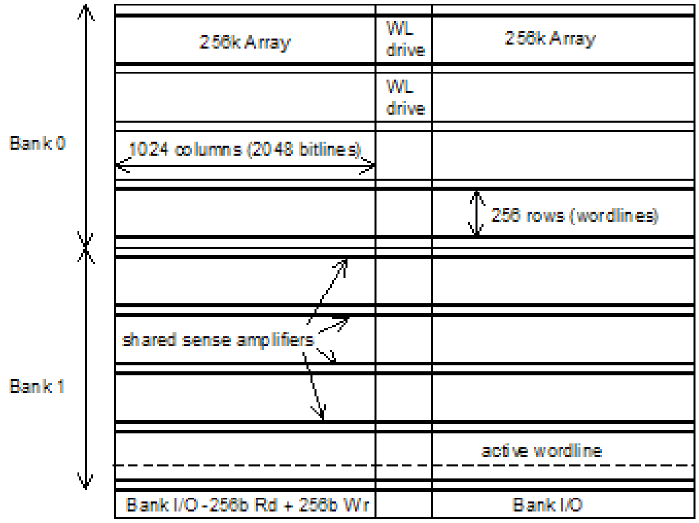

Figure 4. High Bandwidth 4Mbit Segment

An extra row of sense amplifiers in the middle of the macrocell divides the arrays into 2 independent banks. Row decoders and wordline drivers enable arrays on either side of the central spine. Four of these 4Mb segments are placed side by side with shared control circuitry to create a 16Mb macrocell. This macrocell provides 2048 bits of I/O at 500MHz for a sustained aggregate bandwidth of 128Gbytes/second. Figure 5 shows 5-1-1-1 interleaved bank operation providing full utilization of the I/O resources. Bank 0 is activated first, and a burst of 4 words appears on the I/O bus after an initial latency of 5 clocks. Bank 1 is activated 4 cycles after Bank 0 to produce a second burst of data that follows the first without a gap. Bank 0 is automatically precharged in anticipation of a further instruction at tRC=16ns. This architecture achieves very high bandwidth while also achieving high cell efficiency exceeding 50%. Cell efficiency is the ratio of the area occupied by memory cells as a percentage of total macro area.

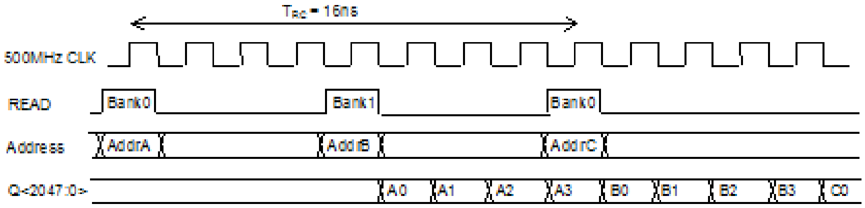

Figure 5. Interleaved Bank Operation

High Speed Macrocell

The embedded DRAM architecture can be optimized for high speed when low latency is a requirement at the cost of additional area overhead. First, the unit cell array size is reduced to minimize both wordline and bitline RC time constants. A significant portion of a conventional DRAM row cycle is devoted to cell restore, charging the cell capacitor to a full Vdd level to maximize the bitline sensing signal and data retention time. If the application can support a reduced refresh interval, and bitlines have been shortened to reduce the bitline RC time constant, we can trade off partial cell restore for faster row cycle operation. An acceptable sensing signal can be maintained because the lower capacitance of a shorter bitline will attenuate less the partially restored cell data. A cell array having 128 rows and 256 columns (512 bitlines) can support a 4ns row cycle in 90nm technology. Four of these arrays are combined to create a single bank. Sixteen banks form a high speed 2Mbit macrocell as shown in Figure 6.

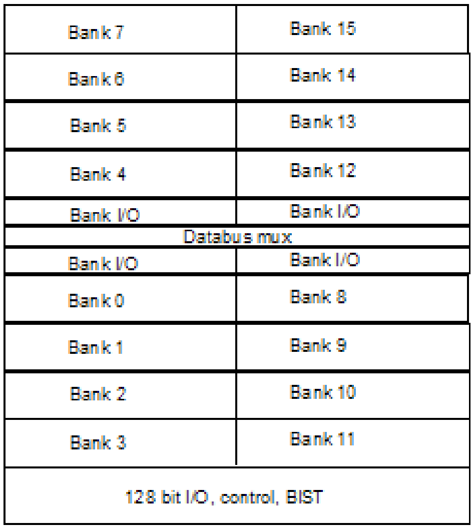

Figure 6. High Speed 2Mbit Macrocell

The high speed architecture provides 1GHz column operations over separate 128 bit read and write datapaths. The synchronous interface enables 3-1-1-1 access. The bank is automatically precharged after the read or write column cycle is completed. A large number of banks reduces the possibility that two successive memory accesses will be made to the same bank, allowing a higher hit rate. If a bank collision does occurs there will be a 3 cycle delay to permit row precharge in the accessed bank. The cost of achieving high performance is the additional area penalty for an increased number of sense amplifiers and decoders to support a small array size, and the additional control circuitry to support a large number of banks. A further benefit of small arrays and multiple banks is the minimization of the active power consumed in charging and discharging bitlines every 4ns. The cell efficiency for this architecture, where speed and latency are pushed to the practical limit, is in the 30% range.

Low Power Macrocell

The final embedded DRAM configuration to be explored is one optimized for both low active power and low standby power. To minimize active power only a small number of columns should be activated during a row cycle. The wide databus architecture easily supports an active bitline to databus ratio of 4:1 to achieve this objective. A hierarchical row decoder scheme which allows a small segment of columns to be activated without excessive area penalty is shown in Figure 7. The wordline decoder is divided into 2 sections, a global wordline decoder (GWL) which selects a group of 4 local wordlines, and the local wordline decoder (LWL), which makes the final 1 of 4 selection. In this way only 512 columns are sensed in a row cycle to minimize active power. Four of the 2Mbit segments are stacked to create an 8Mb asynchronous low power macrocell. A 6ns access time and cycle time can be achieved with a 1.0v supply. This power-optimized architecture achieves roughly 50% cell efficiency while reducing active power to only 25mW.

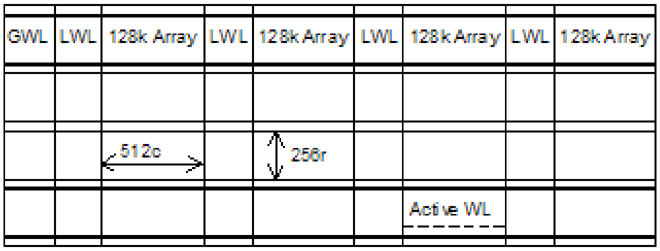

Figure 7. 2Mbit Low Power Segment with Hierarchical Wordline Structure

Low standby power is also critical in battery powered mobile applications. Subthreshold leakage at the 90nm node is highly dependant on the type of transistor employed. High speed transistors will consume significant power in standby while low power transistors will not be fast enough to meet the speed target. For this design we have selected standard transistors and appropriate circuit techniques to minimize standby power. Furthermore, the low power macrocell operates down to 0.7v in sleep mode to further reduce leakage. In sleep mode the macrocell maintains data through periodic self-refresh operation. Internal voltage supplies are activated with a small duty cycle to cut most of the regulator current consumption. As a result, 35?W standby power for the 8Mb macrocell is achieved.

Summary

The three different 90nm macrocells presented here have been optimized to meet the diverse requirements of different applications. Table 1 highlights the key performance indicators for each macrocell. The high bandwidth configuration meets the demanding performance requirements of graphics and networking applications. In these applications the problem of getting data on and off the chip often limits overall system performance. By eliminating the need for memory I/O interfaces, more of the chip pin and power budget is available for application interface. The cost in silicon area to achieve >100GB/s bandwidth is modest since high cell efficiency is achieved, although the active power requirements in excess of 1W will limit this level of performance to desktop and other mains powered applications.

The high speed architecture demonstrates that DRAM technology is appropriate for low latency applications, replacing all but the very highest performance SRAM. The price paid for low latency and 1GHz throughput is lower density, although there remains more than a 2x advantage over SRAM. Standby power is also high, due to the use of high speed transistors in some of the speed critical paths, and the need to keep much of the circuitry activated in anticipation of an active cycle. An additional advantage over SRAM in processor main memory and cache applications is the SER immunity of the DRAM cell.

The low power configuration achieves orders of magnitude lower standby power than the equivalent SRAM for long battery life in mobile applications. DRAM self refresh current is significantly lower than SRAM subthreshold and gate tunneling current in 90nm technologies. Active power is also minimized through appropriate segmentation of the array, while maintaining cell efficiency in the 50% range for low cost.

A single embedded DRAM architecture cannot properly address the divergent requirements of computing, graphics, mobile, and consumer applications. It is not possible to achieve the highest bandwidth, the lowest latency, and the lowest power with same macrocell design. Although a general purpose macrocell may satisfy some applications, a more focused solution will provide tremendous benefits for performance hungry products. Embedded DRAM technology is here, ready to be unleashed in the most demanding System-On-Chip applications.