A 900MHz 2.25MByte Cache with On Chip CPU
- Now in SOI/Cu

J. Michael Hill
Jonathan Lachman

Good Morning.

I will be presenting the 2.25MByte cache on-board Hewlett-Packard’s latest PA-RISC CPU.

Outline

- Technology Overview/Motivation for SOI
- SRAM Design Issues in SOI with Local Interconnect
- Design Improvements
- Redundancy
- Performance
- Conclusions

My presentation will include brief descriptions of

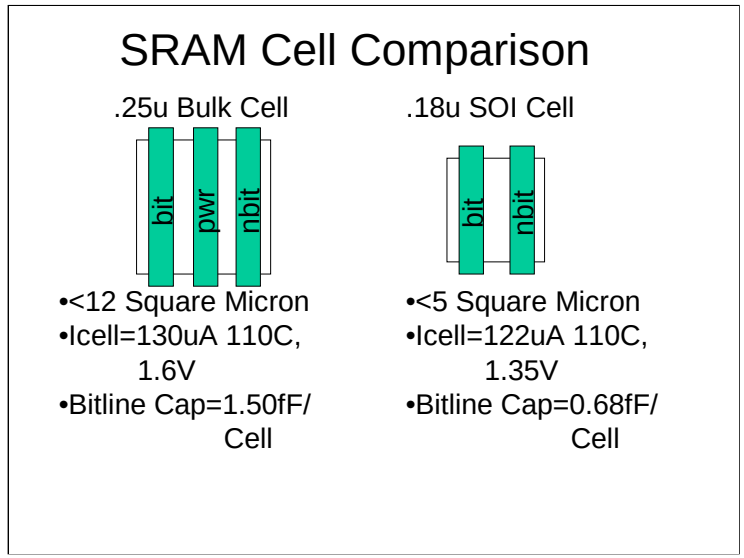
- The technology, and a motivation for using SOI
- SRAM Design issues faced in SOI with local interconnect
- Improvements in the design of the SRAM required to achieve parity in the speed improvement obtained in the CPU core
- Our redundancy methodologies
- Hardware based performance results
- And finally some conclusions

Design/Technology Comparison

Technology	0.25u Bulk, Al	0.18u SOI, Cu
Metal Layers	5	7, Plus Local Interconnect
Cache Size	1.5MByte	2.25MByte (50% Increase)
Designed Operating Freq	500MHz	800MHz (60% Increase)
Die Size	468 mm^2	306 mm^2

The product containing the 2.25MByte cache presented here was ported from a quarter micron 5 level metal bulk technology to a .18u SOI technology with 7 layers of metal and local interconnect. The cache was increased in bit count by 50%. The target operating frequency increased by 60%.

The die size is reduced in area by 34%



The .18u SOI RAM cell is approximately 46% smaller in area than the .25u bulk cell. Approximately 1/5 of this reduction is due to the use of local interconnect. Overall, the new cell reduces bit line capacitance by approximately 45%. 2/3 of this improvement is due to the reduction in wire capacitance. Only ¼ of the improvement is due to the reduction in junction capacitance.

In addition, the read current of the .18u SOI cell is 94% that of the .25u bulk cell.

With the reduction in bit line capacitance, and the nearly equivalent read current, one would expect a significant improvement in rate of bit line differenatial voltage development.

However, note that the .18u cell removes the power wire from between the bit lines. This greatly increases the coupling between bit lines, and will be discussed in detail in later slides.

Digital Logic Performance in SOI

Floating Body/Tied Body
Fall Delay, Rise Delay

rf fr	Inverter	2 Input NAND	2 Input NOR	2 Input Dyn AND
Nom Vt	0.85 0.83	0.79 0.77	0.79 0.82	0.74 0.71
Low Vt	0.89 0.90	0.83 0.81	0.81 0.84	0.77 0.74

Normalized Mean Improvement for Floating Body = 19%

The impact of the floating body on gate delays is presented in this chart. Each entry represents the ratio of the propagation delay for a floating body implementation of a gate versus a bulk like implementation. In the bulk-like implementation, the spice netlist was modified such that the accessible bodies were tied to the appropriate power supply.

The mean improvement introduced by using floating body gates is 19%, but there is a significantly greater improvement for dynamic gates.

Outline

- Technology Overview/Motivation for SOI
- SRAM Design Issues in SOI with Local Interconnect
- Design Improvements
- Redundancy
- Performance
- Conclusions

I will now discuss SRAM design issues

SRAM Design Issues in .18u SOI

- History Effects on Sense Amp Strobe/Output Enable Timing
- Increased Bit Line Coupling Due To:
  - Decrease in Intrinsic Capacitance (Jcap)
  - Use of Local Interconnect Removes PWR Line Between bit and nbit
- Bit Line Differential Slew Rate Degradation Due to State of Bits in Column
  - Read: Source/Drain to Body Cap
  - Write: Bipolar Injection
- Cell Stability

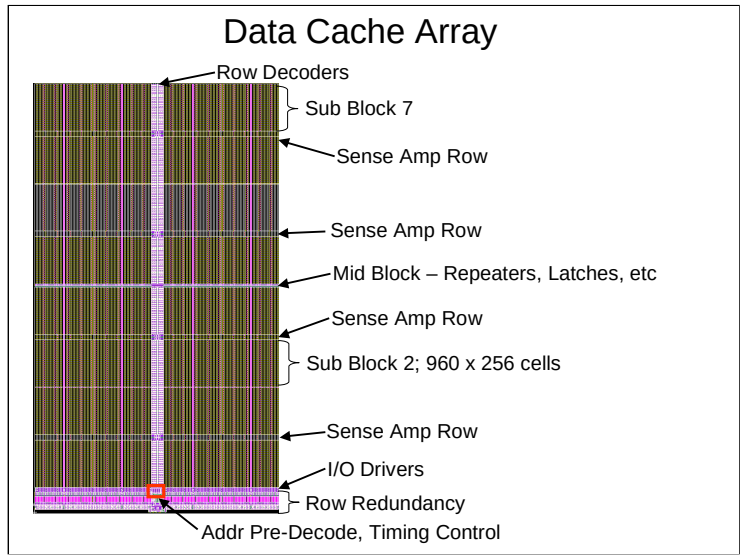
The design of SOI SRAMs poses significant challenges in the goal of matching the expected speed increase of the CPU.

The floating body of SOI gates introduces a delay that is a function of a gates usage pattern, called history dependent delay. For timing delay generators such as those for sense amp or output enable strobes, this edge placement uncertainty must be minimized and properly accounted for in the design timing analysis.

SOI transistors have lower source/drain capacitance than their bulk counterparts. For a RAM cell column, the read slew rate improvement from this reduction in bit line capacitance is offset by the increased bit line coupling caused by the loss of the center power strap between the bit lines.

The speed of reads and writes are dependent upon the state of the RAM cells in the accessed column due to effects unique to SOI: Drain to body capacitance modulation and bipolar current injection. However, these effects are swamped by the bit line coupling.

Finally, access patterns can modulate the body voltage of the RAM cell pass and pull-down transistors. A higher body voltage with the attendant drop in Vt of the pass FET and a low body voltage on the pull down device with a corresponding increase in Vt, can reduce the stability of the RAM cell. Compensating for this loss of stability has the potential to reduce the RAM cell performance.



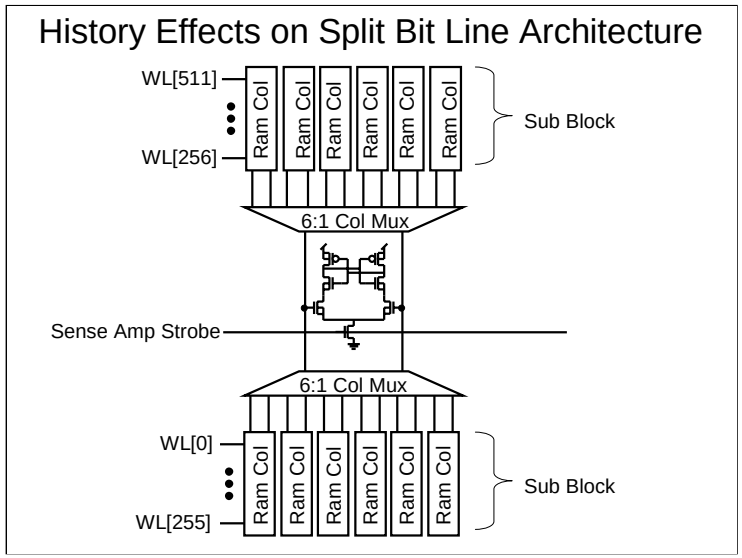
This is a plot of the data cache data array. The array consists of 8 sub blocks of memory cells. Address and data in repeaters, and data output latches reside in the mid-block region. I/O drivers reside at the bottom of the array.

Note that there are four sense amp rows. The sense amp enable signal is generated in the lower center of the array, and routed up the center of the array. A final sense amp strobe is generated for each sense amp row. Two sub blocks share a common sense amp row.

The Addresses are pre-decoded in the lower center of the array, and the final row decode is distributed up the center of the array.

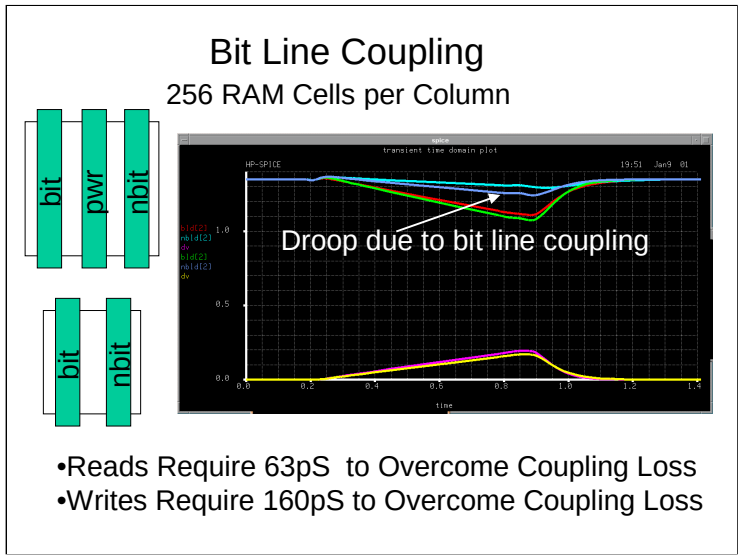
The output buffer is a small signal amplifier whose strobe must be timed to the the availability of data driven from the sense amp.

In SOI designs, the first switching transition of a gate can be shown to be its slowest. Subsequent switching cycles have a reduced propagation delay.



This slide details the split bit line architecture.

Consider the case where the bottom sub block is being repeatedly read. The sense amp strobe is asserting every cycle, hence it is at its minimum propagation delay. Now the address changes and the upper sub block is read. The sense amp strobe still has its minimum propagation delay, but the word line decode path, being switched for the first time, will respond with its maximum propagation delay. Therefore, the signal at the sense amp will be minimized. It is in this simulation environment that the targeted sense amp offset voltage must be obtained. Without any mitigation, this effect would have added 80pS to the sense amp strobe edge placement. A similar effect on the output amplifier strobe would have added an additional 40pS to the access time.



As previously mentioned, the migration to .18u SOI with local interconnect permits the RAM cell area to shrink by almost 60% from the .25u generation and the per cell capacitance to be reduced by 45%. As the slide shows, however, the migration eliminates a power wire between the bit lines. The elimination of this power connection doubles the coupling between bit lines within a cell.

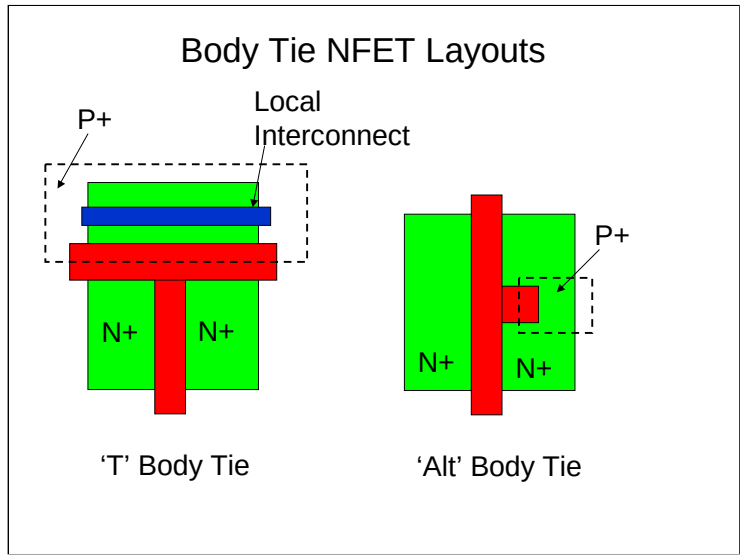
The spice plot shows the bit lines and their differential voltage during reads of a 256 bit column. The increase in coupling between bit lines within a column can be seen by the increase in pull down on the high side bit line shown in blue. The reduction in differential voltage development can be seen on the lower set of curves. The magenta curve shows the differential without coupling, and the yellow curve shows the differential with coupling.

To compensate for this coupling, the sense amp strobe was delayed by an additional 63pS. The droop is even more pronounced during a write, requiring a 160pS increase in active write time to allow the bit line differential to reach and then exceed a desired level long enough to ensure sufficient write margin.

Outline

- Technology Overview/Motivation for SOI
- SRAM Design Issues in SOI with Local Interconnect
- Design Improvements
- Redundancy
- Performance
- Conclusions

I will now discuss the SRAM design improvements necessary to match the speed improvement expected in the CPU



History dependent effects can be nearly eliminated by providing body ties for all the transistors in a path.

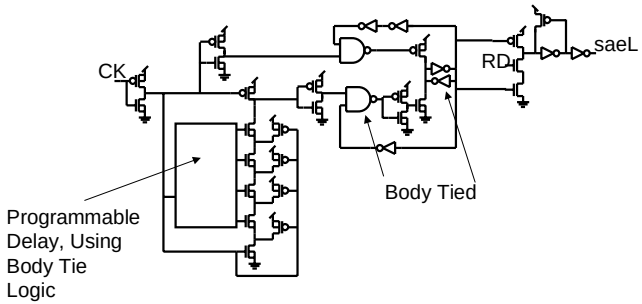
Body tie transistors are formed in two ways.

The T type body tie shown on the left uses a wide poly stripe to isolate the source and drains from the body contact region. The body tie region is of the opposite dopant type from that of the source drains. This provides ohmic contact to the body region under the true gate area. This design style provides maximum flexibility in that the body can be treated as an independent node. This flexability comes at the cost of a fairly large increase in gate capacitance.

The alternate body tie, shown on the right, uses a poly tab extending into the source region. Around this tab an area is doped opposite of that of the source drains, providing an ohmic contact to the body region under the true gate area and is shorted to the source by the silicide. This design style shorts the body to the source, has a much lower impact on gate capacitance, and is smaller. Also, by placing the body contact in the center of the transistor, the distance from the contact to the far ends of the body is cut in half, reducing the RC of the body connection.

These transistor layouts are used extensively in the sense amp and delay generators.

Sense Amp Enable Generation



I will now present the sense amp strobe in two parts.

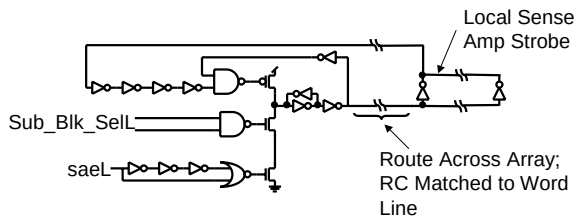
The first stage shown here generates a global sense amp enable signal, called saeL that is routed up the array and provides the enable to each sense amp row. The sense amp enable is formed in two stages, a programmable delay element and a driver that matches the driver of the first stage of address decoders. Three features of this design help mitigate history dependent delays.

First, transistors in the delay stages have their bodies tied to their sources or appropriate power rail.

Second, the delay portion of the circuit fires every clock cycle. The final driver stage is gated with a read signal. Hence, only the final stage is subject to a delay which is a function of the read/write pattern applied to the part. Note also that the transistors in the final driver stage are not body tied.

Third, body tied interstitial pre-charge transistors are placed in the pull down stack of the delay generator. While these transistors do not directly improve the history dependent delay of the design, they do help spice predict accurate delays. By forcing the interstitial nodes to known values every cycle, spice need not be run for many clock cycles to arrive at the quiescent interstitial voltages. They secondarily will provide a more consistent delay, pre-charging the interstitials to VDD.

Sense Amp Strobe Generation

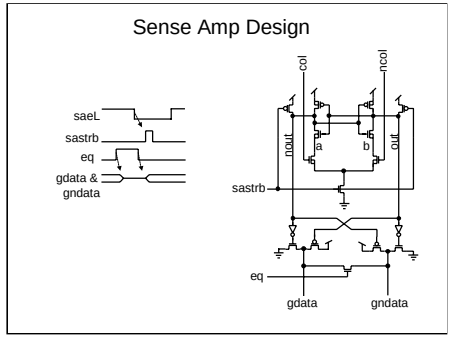


The enable signal from the previous slide is routed to each sense amp row where the final sense amp strobe is generated by the circuit shown in the slide.

In order to minimize the generation of differential body potentials in the sense amp, it is desirable to hold the sense amp in its evaluate state for as short a time as possible. The second stage of the sense amp strobe generates a self timed pulse, producing a minimal width sense amp strobe pulse.

The saeL output from the first stage is connected to a one shot, generating a short pulse on a dynamic pull down tree. If the sub block select is asserted, this pulse will evaluate the dynamic gate, and an active low ‘global’ edge will be sent across the RAM array. At 25% and 75% across the array, the global edge is buffered by inverters. The output of the inverters is the final, local, sense amp strobe which connects to the sense amps. This strobe runs the width of the array, and is also brought back to the second stage strobe generator. The local strobe is then used to precharge the dynamic gate, shutting off the sense amp strobe pulse.

This circuit will power up with the sense amps disabled, without the need of a reset signal.



I will now describe the sense amp.

While the sense amp strobe is de-asserted, precharge devices pull to VDD the outputs of the gain stage, labeled ‘out’ and ‘nout’ as well as the interstitial nodes labeled ‘a’ and ‘b’. The pre-charge devices on nodes ‘a’ and ‘b’ are omitted in the figure for clarity. With ‘out’ and ‘nout’ at VDD, the output ports gdata and gndata are in a high impedance state.

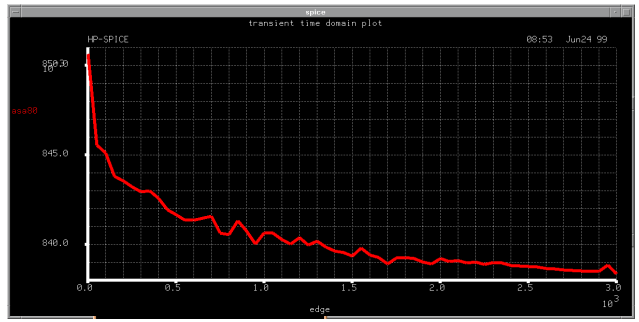
Early in a read cycle, while the sense amp strobe is de-asserted, an equalization pulse is applied to the global data outputs of the sense amp, bringing gdata and gndata to approximately VDD/2. The equalization is disabled immediately prior to the assertion of the sense amp strobe.

Differential data is applied to the signals col and ncol from the accessed memory cell. When the sense amp strobe is asserted, nodes ‘a’ and ‘b’ fall at slightly different rates dependent upon the voltage on col and ncol. The cross coupled stage amplifies the small differential between nodes ‘a’ and ‘b’ producing differential rail to rail signals on nodes ‘out’ and ‘nout’. With nodes ‘a’ and ‘b’ switched to power and ground, differentials will be imposed on the bodies of the transistors in the gain stage. This differential must be eliminated when the sense amp strobe is de-asserted so as to not contribute to offset voltage.

The low going out or nout signal causes the output stage to drive differential data onto gdata and ngdata. Note that the sense amp strobe also controls the time during which the output stage actively drives the global data signals. A latch at the mid-block region of the array holds the global data signals to the rails after the sense amp is disabled.

The sense amp uses alternate body tie style transistors in its differential gain stage. The lower body contact impedance of the alternate style body tie devices reduces the time necessary to pre-charge the source and bodies of the gain stage devices to the positive rail, eliminating an increased offset voltage due to body voltage mismatch. The devices also switch faster than had the bodies been contacted to the appropriate power rails. As described in the previous slide, a pulsed sense amp strobe minimizes the time during which differential body voltages can develop on the gain stage devices, further reducing the probability of body voltage mismatch.

Sense Amp Strobe Delay vs. Cycle Count

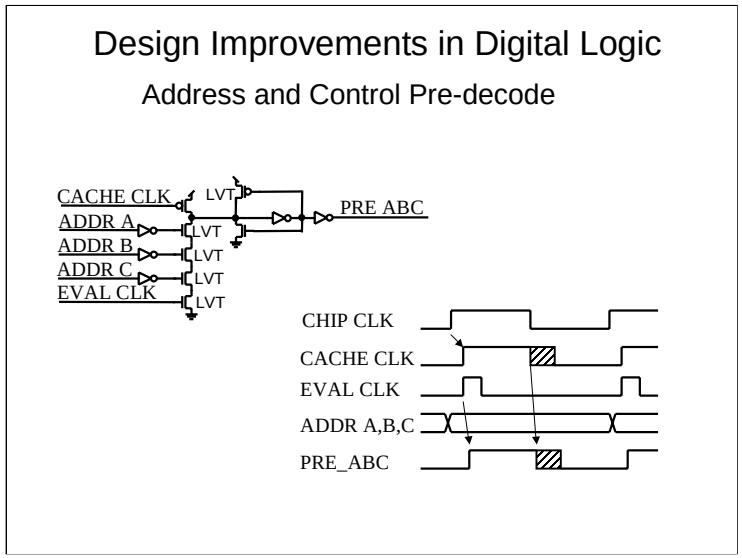


- 50pS Budgeted for History Dependent Delay in Sense Amp Strobe
- 25pS Budgeted for History Dependent Delay in Output Enable

This slide shows the delay through the sense amp strobe network for the first 3000 clock cycles, for the case of no body ties in the sense amp enable generator. It can be seen that the sense amp strobe delay decreases by approximately 15pS over the 3000 clock cycles shown. The difference between the maximum and the minimum propagation delay is 64pS.

By using body tie transistors in the sense amp enable stage the delay change is reduced by 78% to 14pS.

To ensure conservative design margin, 50Ps was budgeted for the history dependence in the sense amp strobe, and 25pS for the output enable.

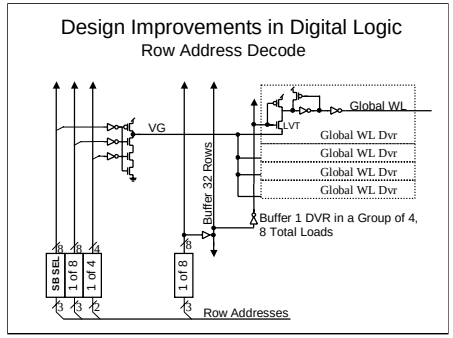


I will now describe the address and control pre-decode logic.

All address and control signal pre-decoding is performed with 3 data input dynamic NAND gates. The CACHE CLK controls the pre-charge device of these pre-decoders and thereby the active portion of a read or write cycle. The width of the CACHE CLK is programmable. The bottom clock input is controlled by a narrow evaluate clock whose rising edge is coincident with that of the CACHE CLK.

The use of this narrow evaluate clock with static address and control inputs reduces any hold time concerns.

The pull down stack of the dynamic gate is comprised of low Vt NFETs. To overcome any noise sensitivity due to the use of low Vt FETs in a dynamic gate, the PFET holder is also a low Vt FET, with a width ratio between the NFET pull down and holder PFET of 10:1. In addition, note that the inputs to the gate are buffered locally, effectively eliminating ground offset between the inputs and the NFET gates.



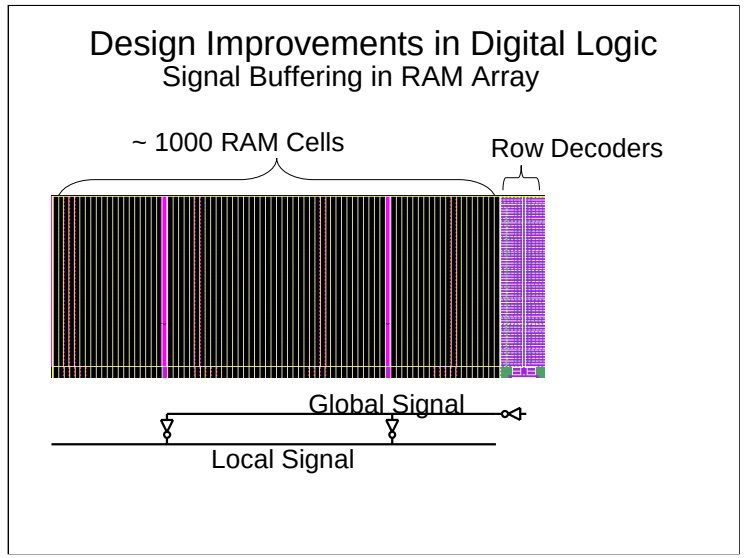
I will now describe the remainder of the row decode path.

At the bottom of this slide are the pre-decoders described in the previous slide. The most significant 8 bits of row address are combined and pre-decoded as 2 groups of 3 and 1 group of 2 into 20 signals, of which only 3 go active in any cycle. The highest order three bits of row addresses, referred to as the sub-block select signals, enable one sub block. These sub block select signals drive the clock input of the dynamic gates that perform further row decoding.

The output of the second stage of decoding forms a virtual ground signal used in the final row decode stage. This virtual ground connects to eight word line drivers, the four shown in this figure plus four more that drive the left side word lines. Note that only one word line driver on the left, and one word line driver on the right can go active in a cycle, reducing the current sunk through the virtual ground.

The pre-decoded outputs of the three least significant row address bits represent the greatest load on the pre-decoder outputs. These signals are buffered every 32 rows. The outputs of each buffer connects to 8 word line drivers. The propagation delay of these buffers matches the delay of the virtual ground, minimizing the overall delay. By using the virtual ground we have eliminated one stage of logic in the row decode path.

The global word line driver uses a low Vt FET in its pull down.



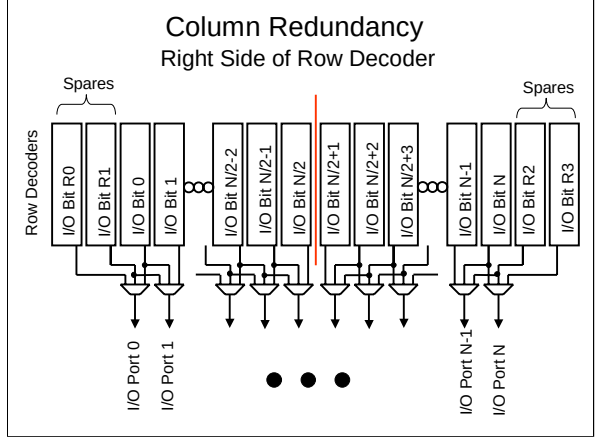
The word line must drive approximately 1000 ram cells. As shown in the slide for the left side of an array, the global word line drives across the array, and is buffered at 25 and 75% of the array width using inverters. The output of these inverters form the local word line that connects to the RAM cells.

This buffering scheme is used for the word lines and all control signals that cross the array width. The delay of control signals is matched to that of the word line to minimize skew across the array.

Outline

- Technology Overview/Motivation for SOI
- SRAM Design Issues in SOI with Local Interconnect
- Design Improvements
- Redundancy
- Performance
- Conclusions

In order to improve yield, several redundancy techniques were utilized.



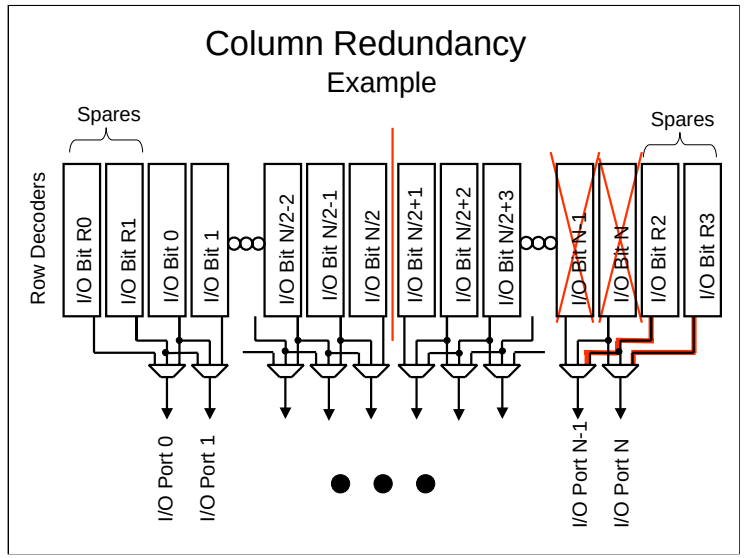
I will now describe our column redundancy scheme.

This figure shows the right side of an array. For the largest cache array,there are 76 I/O ports, connected to 80 I/O bits. Each I/O bit contains 2K rows by 12 bits and all associated column muxes and sense amps. The 80 I/O bits are logically split in half with each having two spare I/O bits. The two spare I/O bits in a half can logically replace any two I/O bits within their half by shifting data around defective I/O bits.

It was desirable to have four redundant elements, each of which that could replace any I/O bit in an array. However, this would have required 5 to 1 input and output muxes, and would have degraded performance due to the extra capacitance on the mux ports. By splitting the array in half, and providing a pair of redundant elements per half, we have reduced the muxes to a 3 to one, with a very slight reduction in the efficacy of the redundancy.

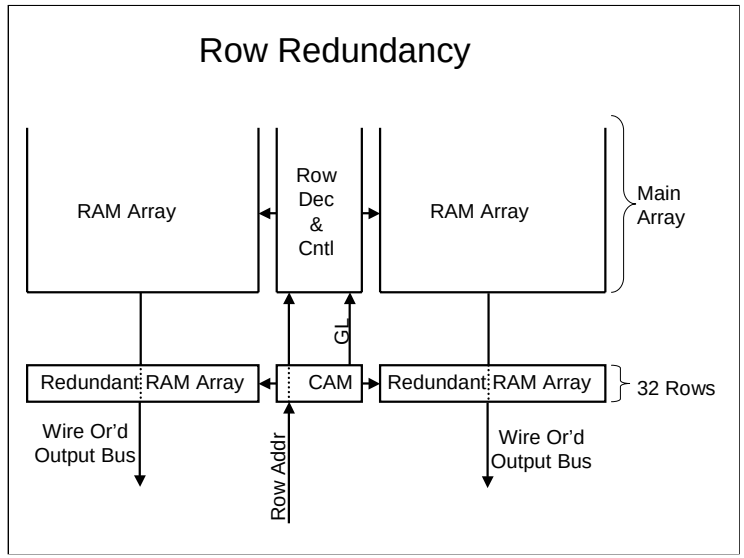
During wafer testing the defective memory structures are identified using saturating counters placed at each output port. The counter is incremented each time an I/O port fails for a given column within the I/O port. If the count reaches seven, the position of the failure is encoded and later stored in laser programmable fuses. The encoded values are then used to program the muxes shown in the figure. Each output mux can select its input from the I/O bit immediately above, one I/O away, or two I/O’s away.

A similar set of muxes steer input data.



As an example, the I/O slices labeled Bit N-1 and bit N are assumed to be defective. The output mux corresponding to port N-1 is programmed to accept its input from spare block 2, 2 I/O's away. Similarly, the output mux for I/O bit N selects spare block 3.

It can be shown that this arrangement will permit the replacement of any single I/O or pair of I/O's, adjacent or not.



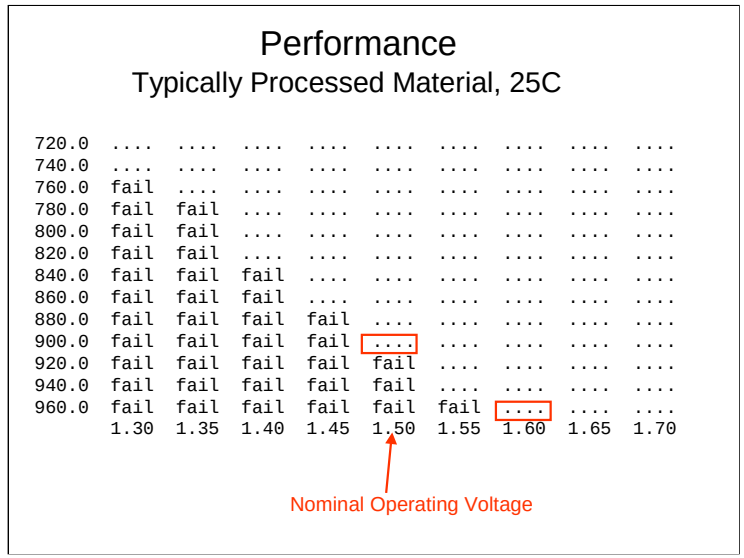
In addition to the column redundancy, a row redundancy block was added to each major cache block. The outputs of the main RAM array are wire-ore'd with the outputs of the redundant RAM array. Up to 32 rows may be replaced by the row redundancy block.

After a column repair has been performed, the BIST test is run again. The row address of any failure is stored in a CAM structure. On subsequent accesses of a formerly failing address, the CAM will match, enabling the redundant RAM array three-state drivers to drive the output bus. The CAM also generates an enable, GL, that three-states the main array outputs. By placing the row redundancy block adjacent to the main array and simply wire-or'ing the outputs, the addition of row redundancy had little impact on read timing.

Outline

- Technology Overview/Motivation for SOI
- SRAM Design Issues in SOI with Local Interconnect
- Design Improvements
- Redundancy
- Performance
- Conclusions

I will now present some typical performance results.



This is a shmoo plot from a nominally processed die at an ambient temperature of 25 degrees C. As indicated by the highlighted point, at a supply voltage of 1.5 volts the cache will operate at over 900MHz.

Subsequent testing shows that the device will operate at above 960MHz at 1.6V, with 960 MHz being at the limit of the test hardware used for cache characterization to date.

Outline

- Technology Overview/Motivation for SOI
- SRAM Design Issues in SOI with Local Interconnect
- Design Improvements
- Redundancy
- Performance
- Conclusions

In conclusion….

Conclusions

- Difficult to Match SRAM Performance to 'Typical' Digital Improvements in .18u SOI
  - Coupling due to smaller SRAM cell
  - History Dependent Delay
  - Differential Body Voltage in Sense Amp

- Accounting for the above, and optimizing design yield:
  - 2.25MByte
  - Greater than 900MHz Operation
  - Consuming < 7.1 Watts @1.5V, 900MHz

When migrating a design from bulk to SOI, while increasing the cache size by 50%, it is difficult to match the anticipated improvement in performance of the CPU. The difficulty is primarily due to

- Coupling due to the smaller SRAM cell
- History dependent delay
- Differential body voltage in Sense amp, leading to higher offset voltage

Accounting for the above, and optimizing the digital portions of the design, yield

- A 2.25MByte single cycle access cache
- Greater than 900MHz operation
- Consuming <7.1W at 1.5V, 900MHz

That concludes our presentation. Thank you for your attention!