



A Step-by-Step Design and Analysis of Low Power Caches for Embedded Processors

Mahmoud Ben Naser and Csaba Andras Moritz
Department of Electrical and Computer Engineering
University of Massachusetts, Amherst
Jan 21, 2005

Motivation

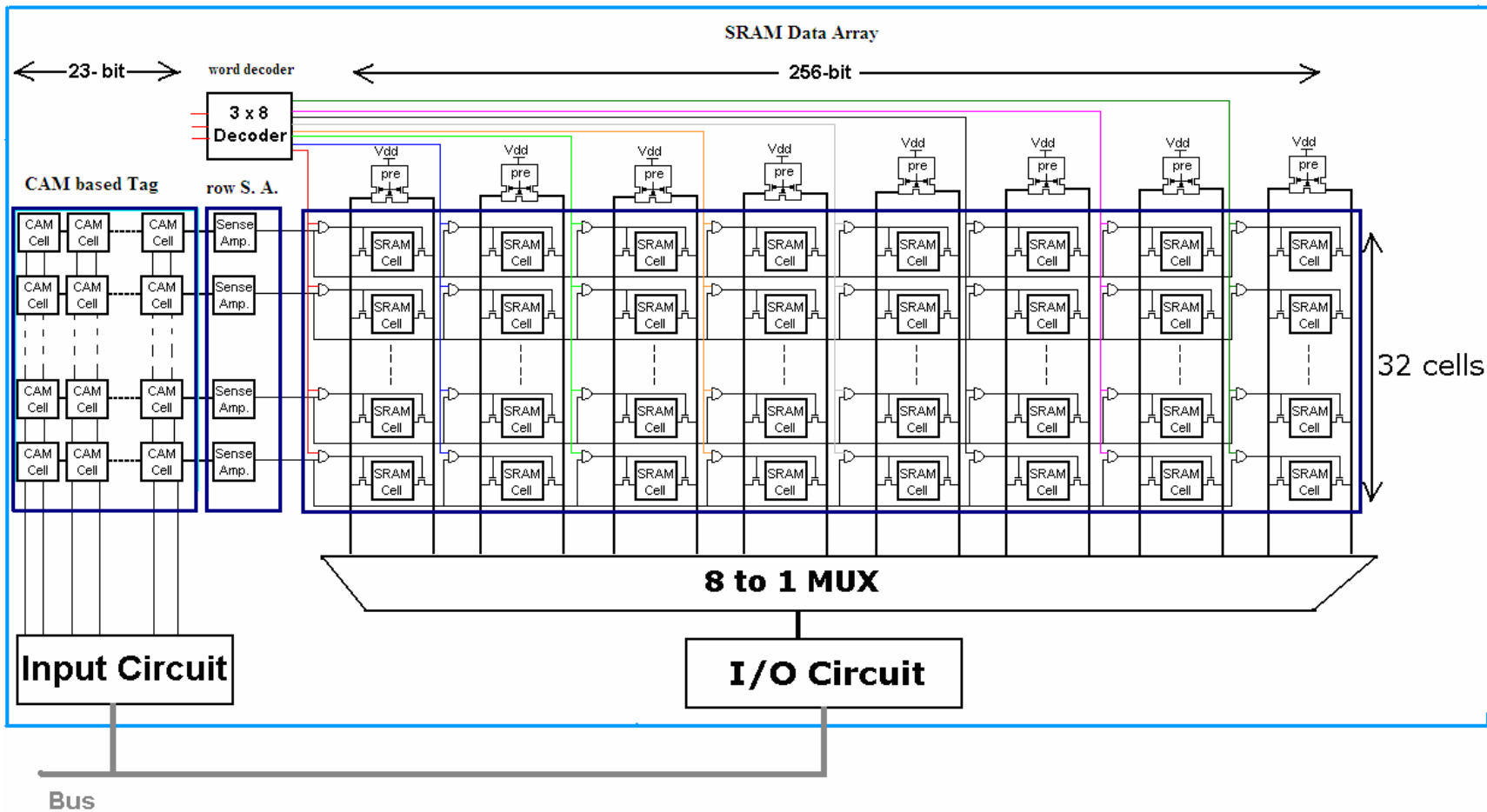
- Caches can consume 50%+ of total chip energy and account for a large fraction of the total chip area in embedded processors.
 - Example: 50% power for ARM10TDMI in 0.13micron TSMC at 400MHz.
- While there has been lots of work focusing on the cache subsystem there are many questions still unanswered.
 - How do different design styles affect overall cache power and delay in deep sub-micron technology?
 - How much power is consumed in the auxiliary cache circuits?
 - What is the fraction of power consumed in different parts when all components are using state-of-the-art low-power circuits?
 - How much power reduction is possible at the circuit level?



Talk Outline

- Overview Cache Organization
- Design Techniques to Reduce Power Consumption
 - Dynamic Power Reduction
 - Leakage Reduction Techniques
- Results
- Conclusion

Cache Organization Overview



Address Decoder Design

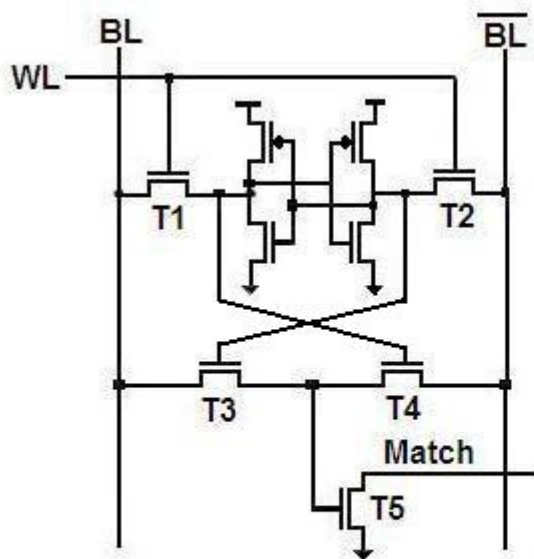
- Decoder Design involves choosing the optimal circuit and figuring out their sizing.

Decoder Style	# of stages	Delay	Power
6-input Dynamic NOR	1	0.14 ns	128 μW
2-input SNAND- 3-input DNOR	2	0.12 ns	105 μ W
3-input DNOR-2-input SAND	2	0.22 ns	37.5 μ W
2-input DNAND- 3-nput SNOR	2	0.33 ns	28.4 μ W
6-input Dynamic NAND	1	0.20 ns	48.5 μ W
3-input DNAND- 2-input SNOR	2	0.20 ns	22.9 μW
6-input Static NAND	1	0.22 ns	38.3 μ W

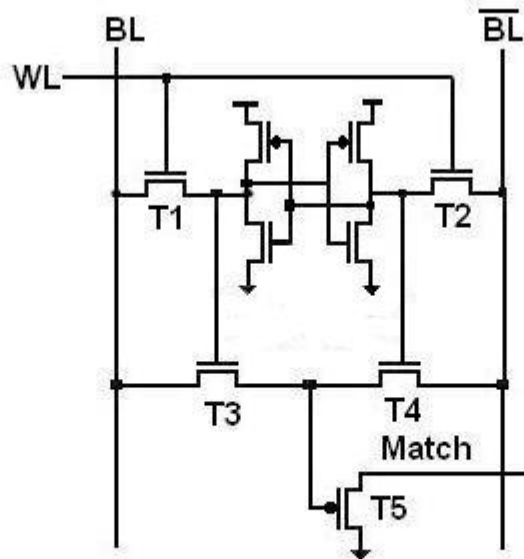
- 82% average power saving from using the 3- input DNAND- 2-input SNOR over the 6-input NOR decoder with some degradation in delay.

Tag Array Design

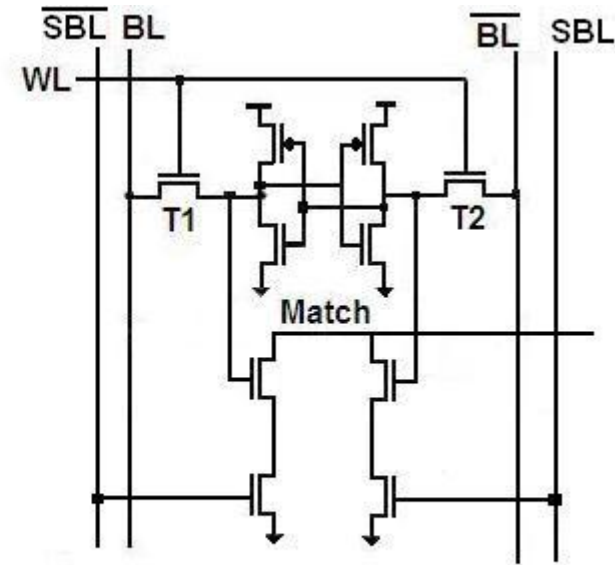
- CAM provides a unique exclusive fast data search function by accessing data by its content rather than its memory location.
- Slight advantage to CCAM at the targeted design point



NCAM



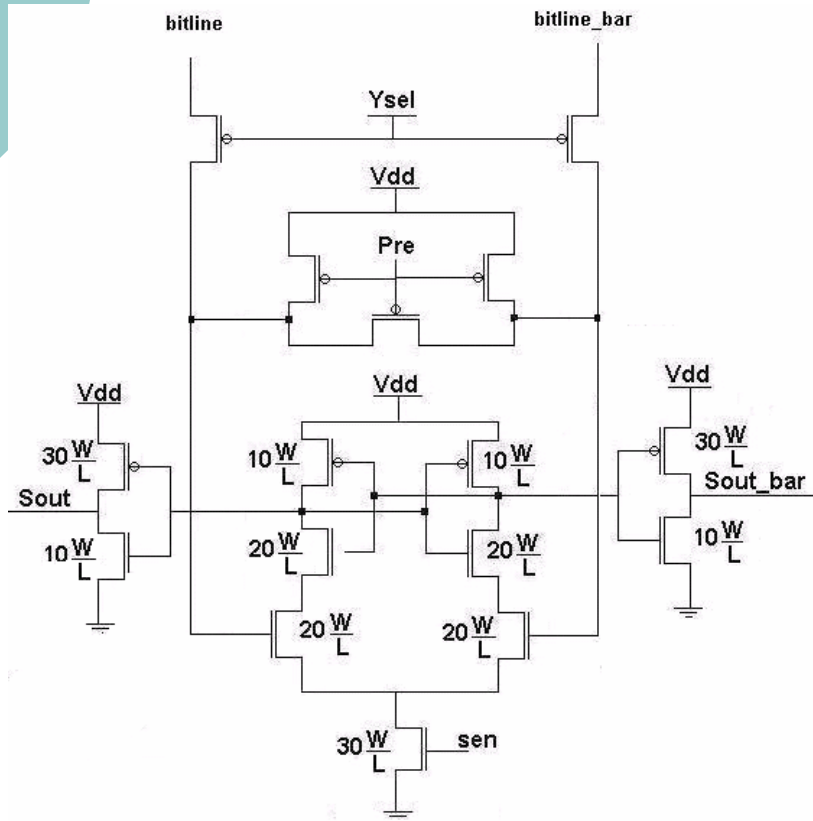
PCAM



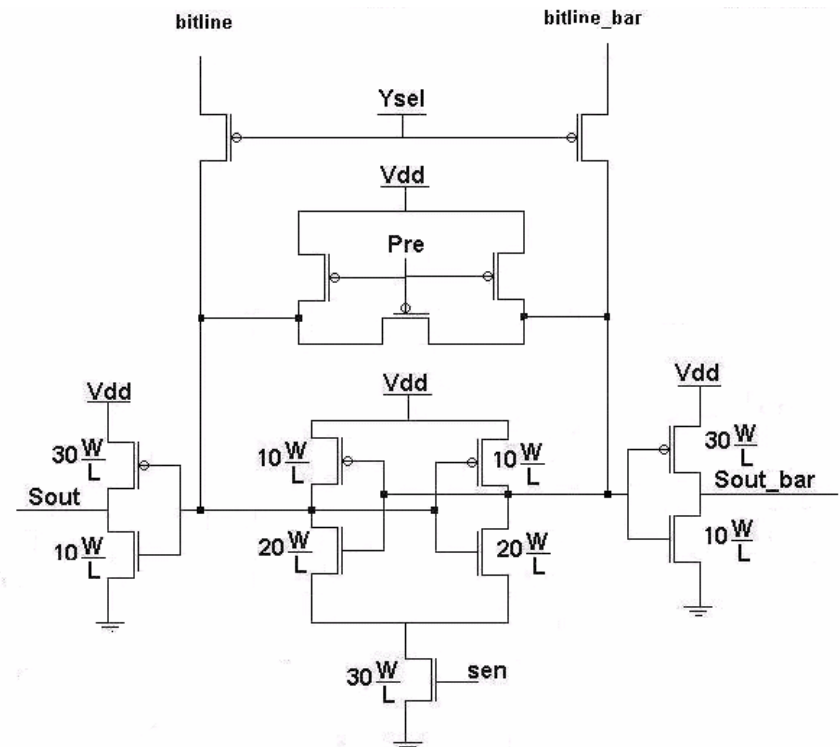
CCAM

Low Power Sense Amplifier

- A key component in the periphery of a cache is the sense amplifier.



Alpha Latch Sense Amplifier



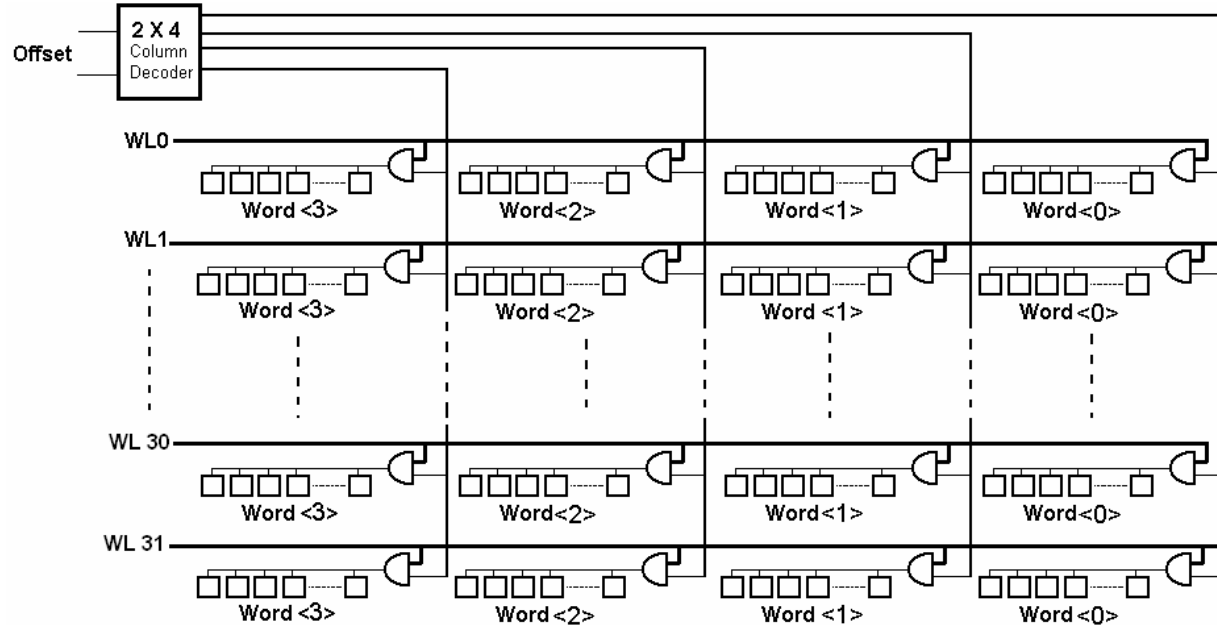
Cross-Coupled Inverter Latch

Low Power Sense Amplifier

- The Alpha latch isolates the output node from the bitlines thereby providing low swing in the bitlines as opposed to CCil.

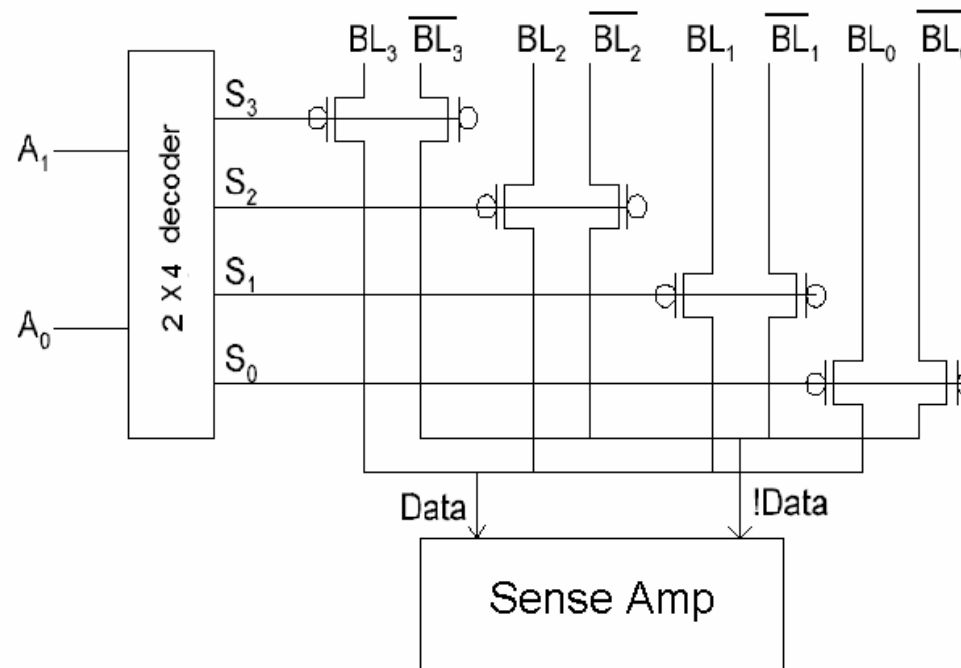
Sense amplifier Circuit	Temperature			
	T=75°C		T=100°C	
	Delay	Power	Delay	Power
Cross-Coupled Inverter Latch	0.70 ns	9.45 mW	0.72 ns	11.3 mW
Alpha latch	0.72 ns	9.17 mW	0.74 ns	10.7 mW

Divided/Local Wordline



- To evaluate the divided wordline technique, we apply it on a 16K-byte 32-way set associative cache with 32-byte cache lines.
- Our power measurements show that the overall cache power consumption is reduced by 10% with the divided wordline technique.

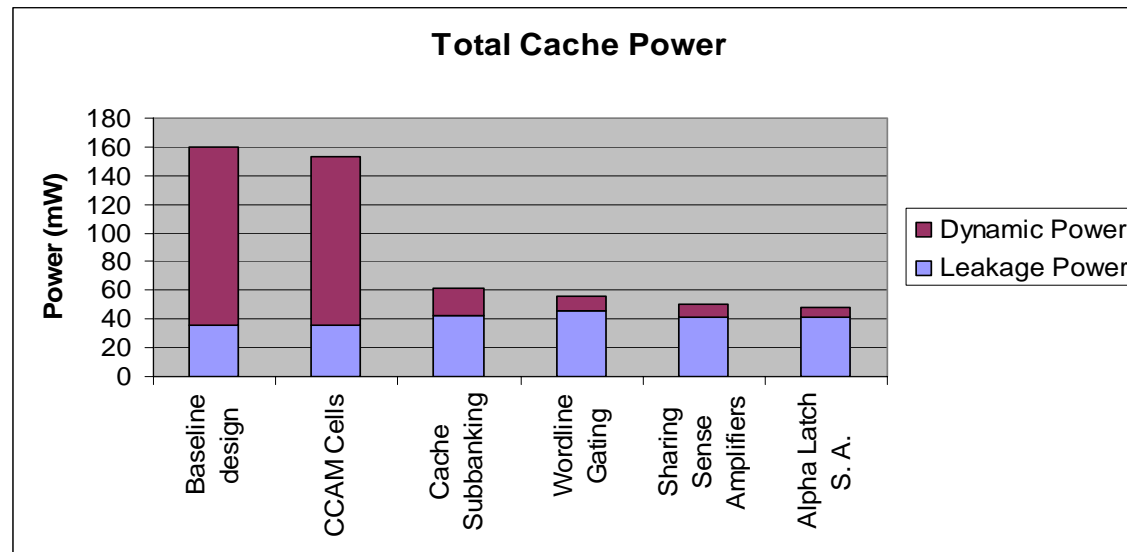
Sharing Sense Amplifiers



- For example, a 32-byte cache line can be partitioned into eight words in which there are 32 sense amplifiers shared among 256 pairs of bitlines.
- The power savings from sharing sense amplifiers (we have evaluated the Alpha latch) can be as high as 8-11% of the total cache power.

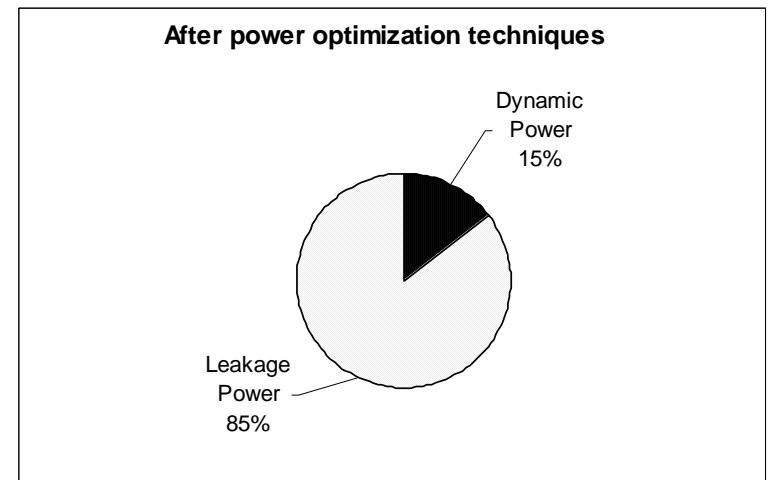
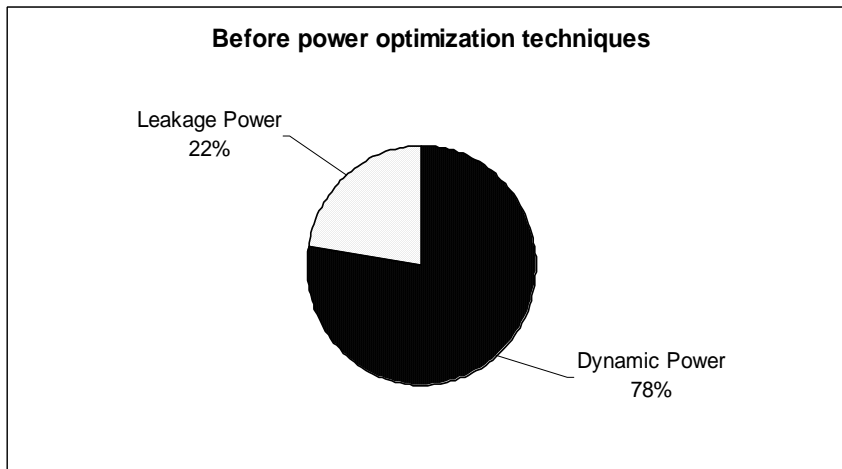
Power Optimization Results

- The Cadence tool and HSPICE simulation was used to evaluate the performance and power dissipation of a 16 KB cache.
 - Target is 1GHz design in 70nm BPTM
- The baseline design is a 16 KB cache with no active power optimization techniques using NCAM cell and CCil sense amplifier.



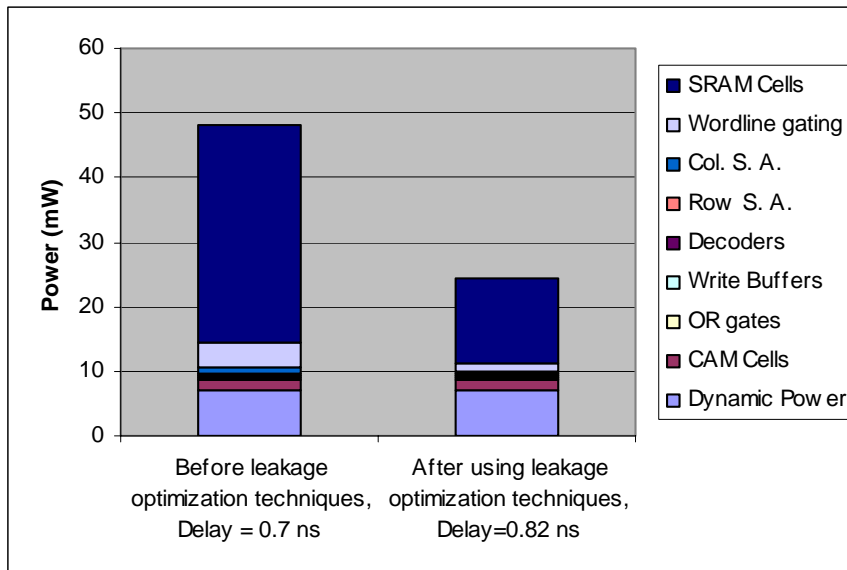
Power Optimization techniques

- After dynamic power optimizations are done how big is the leakage?



Leakage Reduction Techniques

- We have **evaluated** several leakage optimization techniques: e.g., dual Vt, stacked “sleep” transistors, and supply voltage reduction.

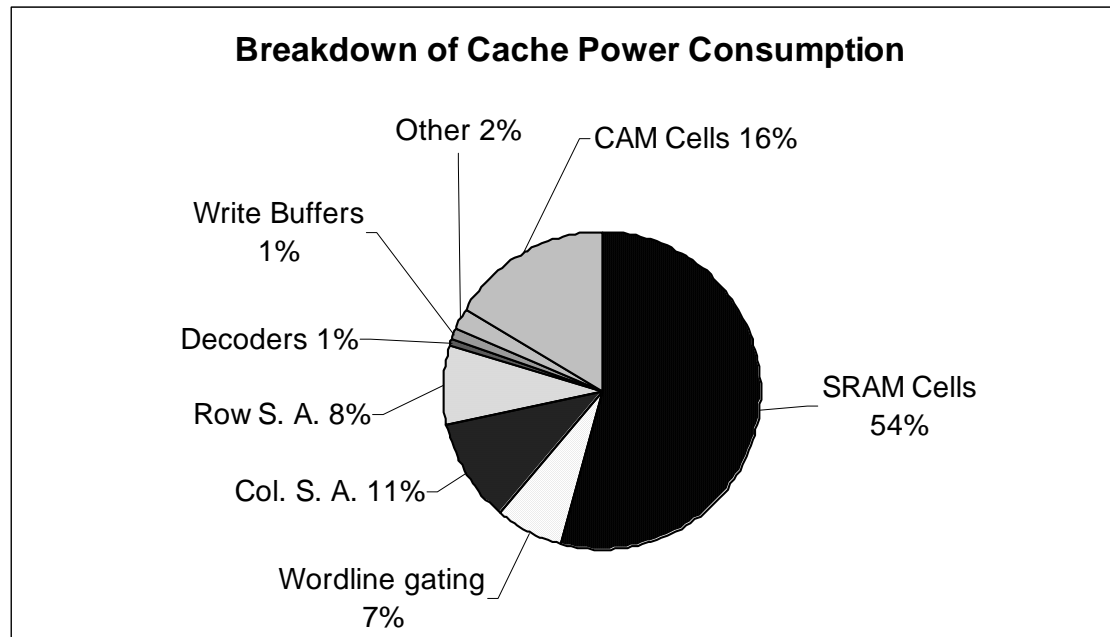


Cache Component	Leakage Technique
SRAM Cell	Asymmetric Cell
Wordline gating	High-Vt
Sense Amps.	stacked “sleep” transistors

- The overall cache leakage power consumption is reduced by **58%** with the **dual-vt and stacked** techniques at T= 75 C.
- 1GHz design limits the leakage reduction. Slower design points could have much larger fraction of the leakage component reduced

Cache Power Breakdown

- After all active and leakage power optimization techniques
- In a relatively high performance 70nm design leakage in the SRAM is still dominant
- Can be further reduced with architectural techniques, e.g., leakage can be optimized further in idle banks



Conclusion

- We explored a complete custom low power cache, at the circuit level, suitable for an embedded microprocessor running at 1GHz in 70nm BPTM.
- The choice of components and optimizations have a very significant impact on power consumption and are affected by the performance objective and technology node.
 - At 400MHz a key contributor might be the CAM active power
 - At 1GHz the part that require additional attention is SRAM leakage
- By applying many state-of-the-art low power techniques, at the circuit level alone, the cache power consumption is reduced by 6X for the same speed and functionality.
 - A challenge to use correct power consumption assumptions in architectural level studies



Thank You!