

Tutorial Outline

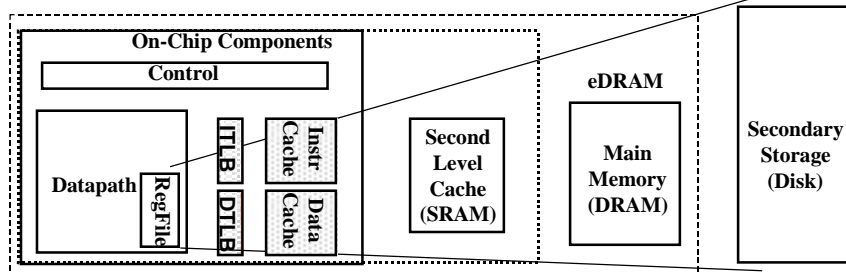
8:30 - 8:45	Introduction and motivation
8:45 - 9:05	Sources of power in CMOS designs
9:05 - 9:30	Power analysis tools and techniques
9:30 - 10:30	Gate & functional unit design issues & techniques
10:30 - 10:50	BREAK
10:50 - 12:15	Architectural level issues and techniques
12:15 - 1:30	LUNCH
1:30 - 2:30	Low power memory system design
2:30 - 3:30	Software level issues and techniques
3:30 - 3:50	BREAK
3:50 - 4:30	Software level issues and techniques, con't
4:30 - 4:45	Future challenges

ISCA Tutorial: Low Power Design

Memories.1

©MJIVN, PSU, 2000

Typical Memory Hierarchy



DEC 21164a ($2.0V_{dd}$, 0.35μ , 400MHz, 30W max)

–caches dissipate 25% of the total chip power

DEC SA-110 ($2.0V_{dd}$, 0.35μ , 233MHz, 1W typ) – no L2 on-chip

–I\$ (D\$) dissipate 27% (16%) of the total chip power

ISCA Tutorial: Low Power Design

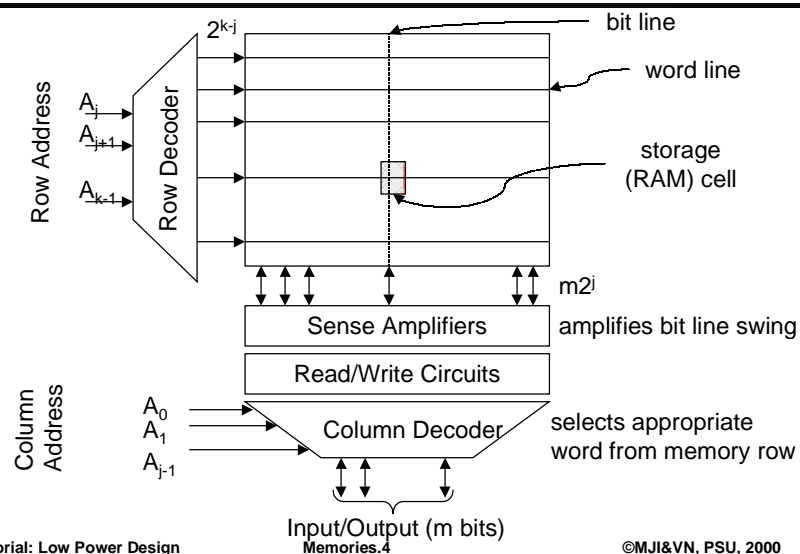
Memories.2

©MJIVN, PSU, 2000

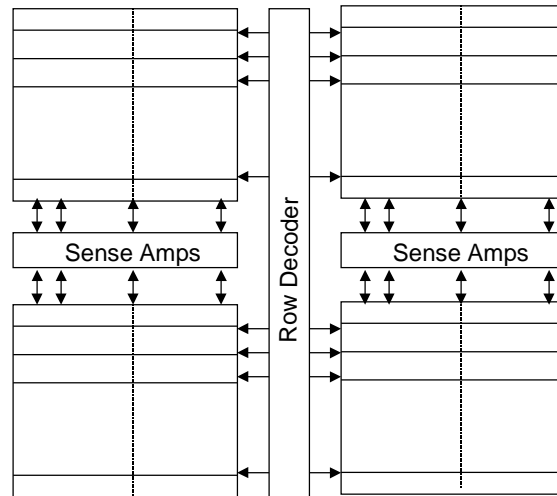
Importance of Optimizing Memory System Energy

- Many emerging applications are data-intensive
- For ASICs and embedded systems, memory system can contribute up to 90% energy
- Multiple memories in future System-on-chip designs

2D Memory Architecture



2D Memory Configuration



ISCA Tutorial: Low Power Design

Memories.5

©MJIVN, PSU, 2000

Sources of Power Dissipation

- **Active Power Sources** Negligible at high frequencies
- $$P = V_{dd} \cdot I_{dd}$$
- $$I_{dd} = m \cdot I_{act} + m \cdot (n-1) \cdot I_{ret} + (n+m) \cdot C_{de} \cdot V_{int} \cdot f + C_{pt} \cdot V_{int} \cdot f + I_{dcp}$$
- $(n+m) = 2$ for CMOS NAND decoders
 Virtually independent of operating frequency
- m** - number of columns
n - number of rows
V_{dd} - External power supply
I_{act} - Effective current of active cells
I_{ret} - Data retention current of inactive cells
C_{de} - Output node capacitance of each decoder
V_{int} - Internal Supply Voltage
C_{pt} - total capacitance in periphery
I_{dcp} - Static current of Column circuitry, Diff Amps

ISCA Tutorial: Low Power Design

Memories.6

©MJIVN, PSU, 2000

DRAM Energy Consumption

- I_{dd} increases with m and n
- Destructive Readout characteristics of DRAM requires bit line to be charged and discharged with a large Voltage Swing, V_{swing} (1.5 - 2.5 V)

$$I_{dd} = [m \cdot C_{BL} V_{swing} + C_{pt} \cdot V_{int}] f + I_{dcp}$$

Reduce charging capacitance - C_{pt} , $m \cdot C_{BL}$

Reduce external and internal voltages - V_{dd} , V_{int} , V_{swing}

Reduce static current - I_{dcp}

DRAM Reliability Concerns

- Signal to Noise Characteristics requires bit line capacitance to be small

$$\text{Signal, } V_s = (C_s / C_{BL}) \cdot V_{swing}$$

C_s - Cell capacitance

Reducing is C_{BL} beneficial

Reducing is V_{swing} detrimental

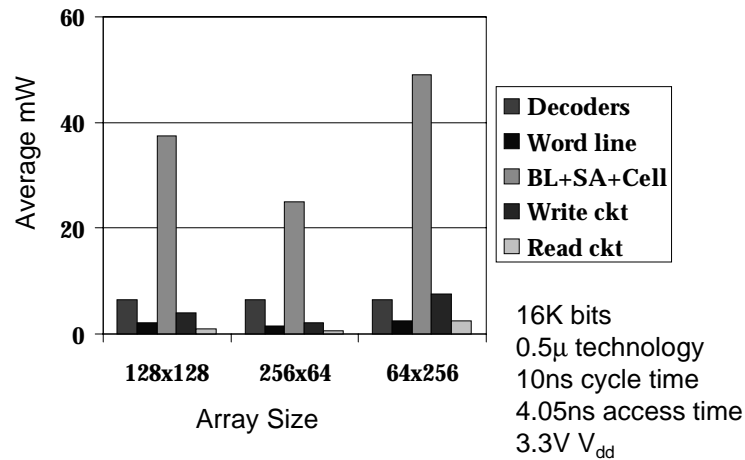
SRAM Design

- $I_{dd} = [m \cdot I_{DC} \Delta t + C_{pt} \cdot V_{int}] f + I_{dcp}$
- **Signal to Noise not so serious**
- **Both SRAM and DRAM have evolved to use similar techniques**

Data Retention Power

- **In data retention mode, memory has no access from outside and data are retained by the refresh operation (for DRAMs)**
- $I_{dd} = [m \cdot C_{BL} V_{swing} + C_{pt} \cdot V_{int}] (n/t_{ref}) + I_{dcp}$
- t_{ref} **is the refresh time and increases with reducing junction temperature**
- I_{dcp} **can be significant in this mode**

SRAM Power Budget



ISCA Tutorial: Low Power Design

Memories.11

©MJIVN, PSU, 2000

Low Power SRAM Techniques

- Standby power reduction
- Operating power reduction
 - » memory bank partitioning
 - » SRAM cell design
 - » reduced bit line swing (pulsed word line and bit line isolation)
 - » divided word line
 - » bit line segmentation
- Can use the above in combination!

ISCA Tutorial: Low Power Design

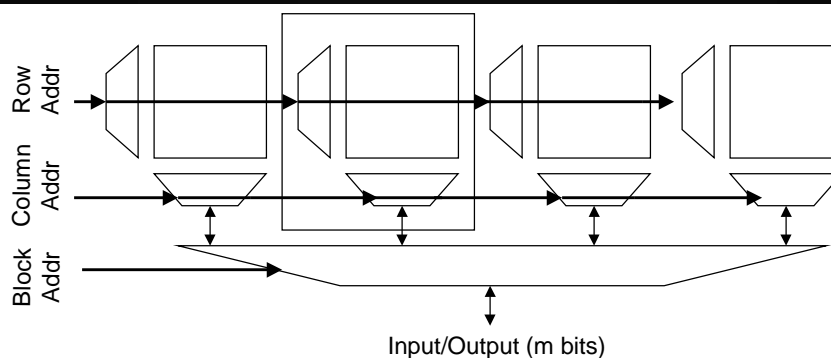
Memories.12

©MJIVN, PSU, 2000

Memory Bank Partitioning

- Partition the memory array into smaller banks so that only the addressed bank is activated
 - » improves speed and lowers power
 - » word line capacitance reduced
 - » number of bit cells activated reduced
- At some point the delay and power overhead associated with the bank decoding circuit dominates (2 to 8 banks typical)

Partitioned Memory Structure

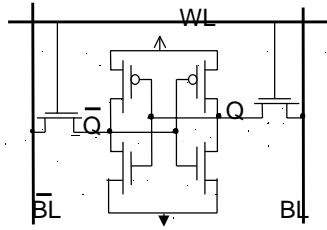


Advantages:

1. Shorter word and/or bit lines
2. Block addr activates only 1 block saving power

SRAM Cell

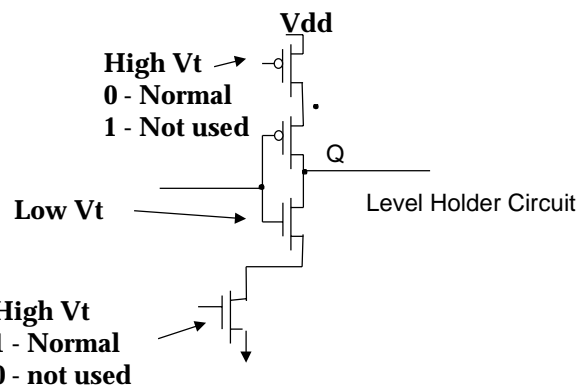
- 6-T SRAMs cell reduces static current (leakage) but takes more area



- Reduction of V_{th} in very low V_{dd} SRAMs suffer from large leakage currents

» use multiple threshold devices (memory cells with higher V_{th} to reduce leakage while peripheral circuits use low V_{th} to improve speed)

Switched Power Supply with Level Holding



- Multi V_t device by changing Well voltages; V_t high during standby & low otherwise

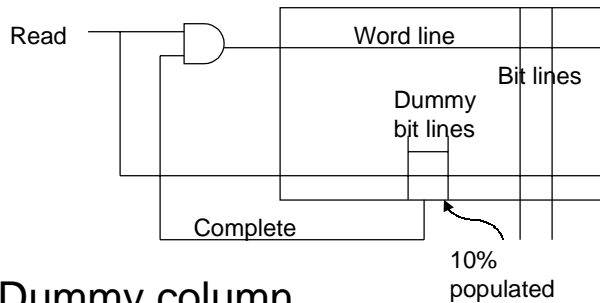
Reduced Bit Line Swing

- Limit voltage swing on bit lines to improve both speed and power
 - » need sense amp for each column to sense/restore signal
 - » isolate memory cells from the bit lines after sensing (to prevent the cells from changing the bit line voltage further) - pulsed word line
 - » isolate sense amps from bit lines after sensing (to prevent bit lines from having large voltage swings) - bit line isolation

Pulsed Word Line

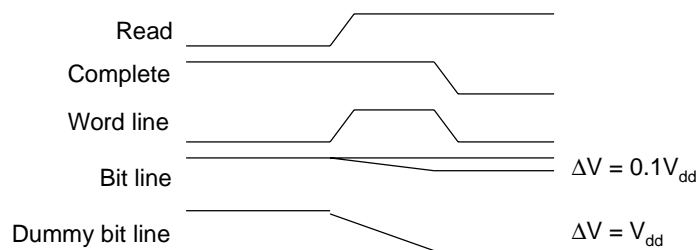
- Generation of word line pulses very critical
 - » too short - sense amp operation may fail
 - » too long - power efficiency degraded (because bit line swing size depends on duration of the word line pulse)
- Word line pulse generation
 - » delay lines (susceptible to process, temp, etc.)
 - » use feedback from bit lines

Pulsed Word Line Structure



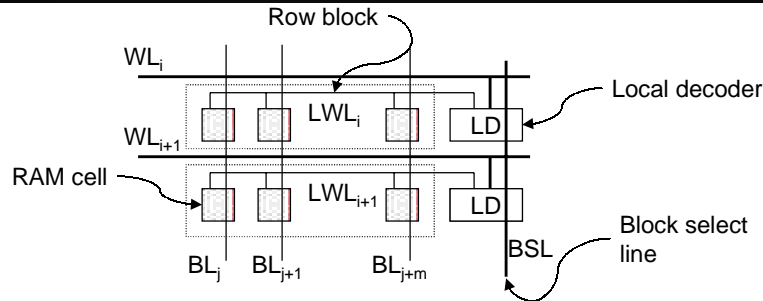
- **Dummy column**
 - » height set to 10% of a regular column and its cells are tied to a fixed value
 - » capacitance is only 10% of a regular column

Pulsed Word Line Timing



- **Dummy bit lines have reached full swing and trigger pulse shut off when regular bit lines reach 10% swing**

Divided Word Line Structure



- Load capacitance on word line determined by number/size of local decoder
 - » faster word line (since smaller capacitance)
 - » now have to wait for local decoder delay

Cells/Block

- How many cells to put in one block?
 - » Power savings best with 2 cells/block
 - fewest number of bit lines activated
 - » Area penalty worst with 2 cells/block
 - more local decoders and BSL buffers
 - » BSL logic
 - need buffers to drive each BSL
 - 4 and 16 cells/block BSLs are the enable inputs of the column decoder's last stage of 2x4 decoders
 - 2 (8) cells/block need a NOR gate with 2 (8) inputs from the output of the column decoder

DWL Power Reduction

Cells/block	Write Operations			Read Operations		
	128x128	256x64	64x256	128x128	256x64	64x256
2	77.0%	68.5%	78.4%	80.1%	71.6%	82.9%
4	75.5%	65.5%	77.2%	79.1%	68.3%	82.0%
8	73.1%	60.3%	75.8%	76.6%	62.9%	80.3%
16	67.2%	49.8%	72.6%	70.2%	51.9%	76.7%

From Chang, 1997

ISCA Tutorial: Low Power Design

Memories.25

©MJl&VN, PSU, 2000

DWL Area Penalty

Cells/block	128x128	256x64	64x256
2	25.5%	24.6%	24.8%
4	19.2%	18.5%	18.4%
8	17.0%	16.5%	16.2%
16	15.4%	14.8%	14.5%

ISCA Tutorial: Low Power Design

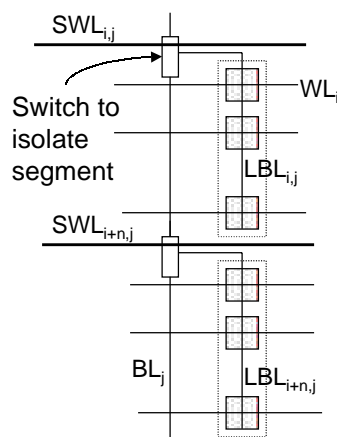
Memories.26

©MJl&VN, PSU, 2000

Bit Line Segmentation

- RAM cells in each column are organized into blocks selected by word lines
- Only the memory cells in the activated block present a load on the bit line
 - » lowers power dissipation (by decreasing bit line capacitance)
 - » can use smaller sense amps

Bit Line Segmented Structure



- Address decoder identifies the segment targeted by the row address and isolates all but the targeted segment from the common bit line
- Has minimal effect on performance

Cache Power

- On-chip I\$ and D\$ (high speed SRAM)
 - » DEC 21164a ($2.0V_{dd}$, 0.35μ , 400MHz, 30W max)
 - I/D/L2 of 8/8/96KB and 1/1/3 associativity
 - caches dissipate 25% of the total chip power
 - » DEC SA-110 ($2.0V_{dd}$, 0.35μ , 233MHz, 1W typ)
 - I/D of 16/16KB and 32/32 associativity (no L2 on-chip)
 - I\$ (D\$) dissipate 27% (16%) of the total chip power
- Improving the power efficiency of caches is critical to the overall system power

Cache Energy Consumption

- Energy Dissipated by Bitlines: precharge, read and write cycles
- Energy Dissipated by Wordlines: when a particular row is being read or written
- Energy Dissipated by Address Decoders
- Energy Dissipated by Peripheral Circuit - comparators, cache control logic etc.
- Off-Chip Main Memory Energy is based on per-access cost

Analytical Energy Model Example

- On-chip cache

$$\text{Energy} = E_{\text{bus}} + E_{\text{cell}} + E_{\text{pad}} + E_{\text{main}}$$

...

$$E_{\text{cell}} = \beta * (\text{wl_length}) * (\text{bl_length} + 4.8) * (\text{Nhit} + 2 * \text{Nmiss})$$

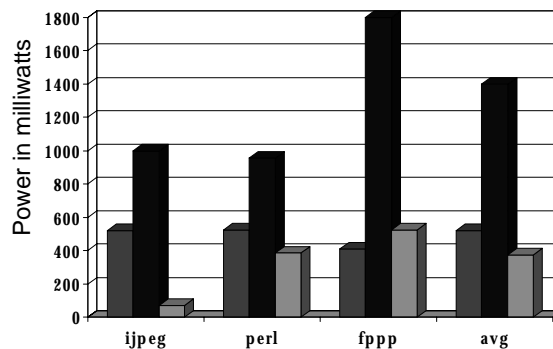
$$\text{wl_length} = m * (\text{T} + 8\text{L} + \text{St})$$

$$\text{bl_length} = C / (m * \text{L})$$

Nhit = number of hits; Nmiss = number of misses;

C = cache size; L = cache line size in bytes; m = set associativity; T = tag size in bits; St = # of status bits per line; $\beta = 1.44e-14$ (technology parameter)

Cache Power Distribution



Base Configuration:

- 4-way superscalar
- 32KB DM L1 I\$
- 32KB, 4-way SA L1 D\$
- 32B blocks, write back
- 128KB, 4-way SA L2
- 64B blocks, write back
- 1MB, 8-way SA off-chip L3
- 128B blocks, write thru

Interconnect widths

- 16B between L1 and L2
- 32B between L2 and L3
- 64B between L3 and MM

From Ghose, 1999

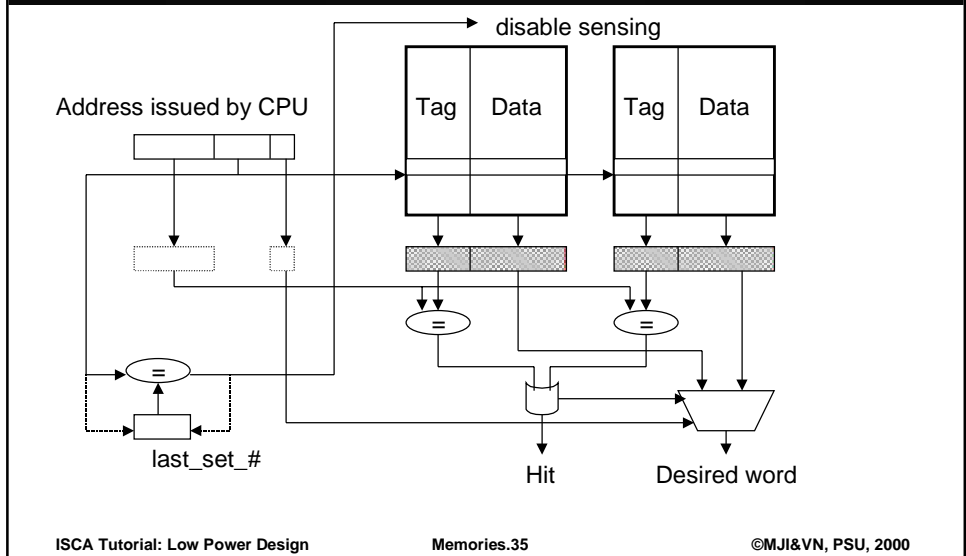
Low Power Cache Techniques

- SRAM power reduction
- Cache block buffering
- Cache subbanking
- Divided word line
- Multidivided module (MDM)
- Modifications to CAM cell (for FA cache and FA TLB)

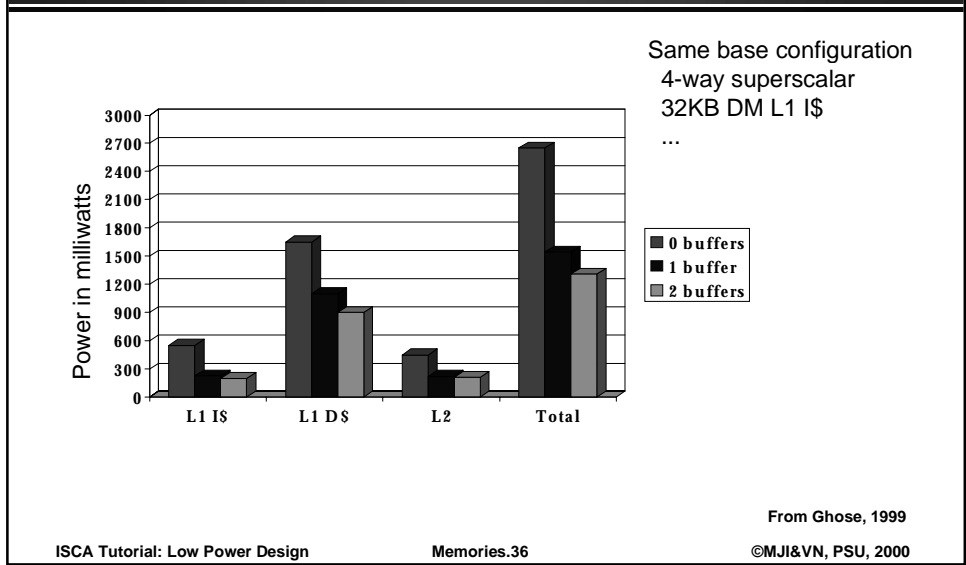
Cache Block Buffering

- Check to see if data desired is in the data output latch from the last cache access (i.e., in the same cache block)
- Saves energy since not accessing tag and data arrays
 - » minimal overhead hardware
- Can maintain performance of normal set associative cache

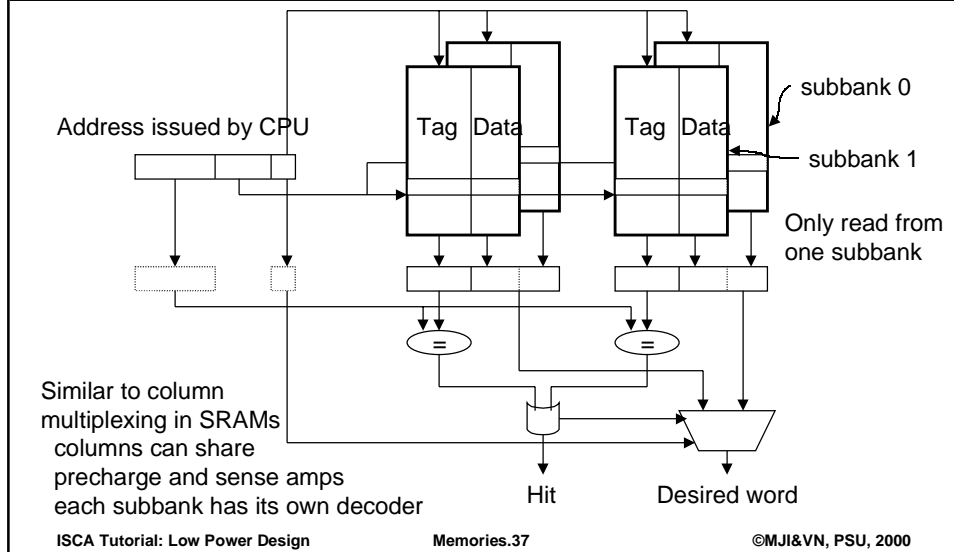
Block Buffer Cache Structure



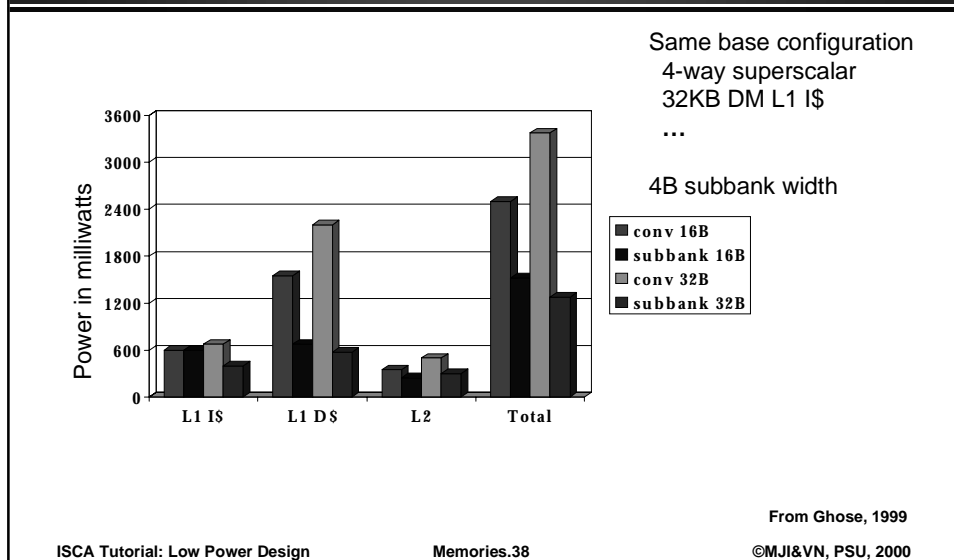
Block Buffering Performance



Cache Subbanking



Subbanking Performance

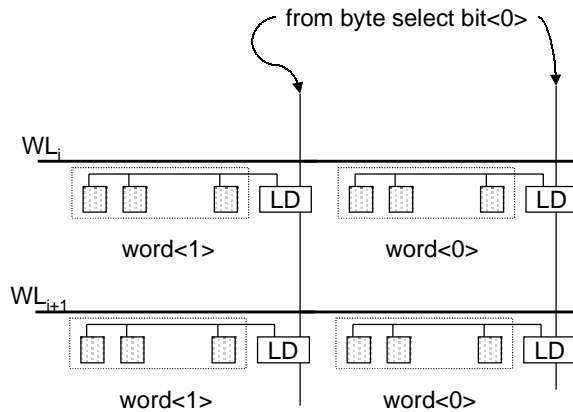


Divided Word Line Cache

Same goals as subbanking

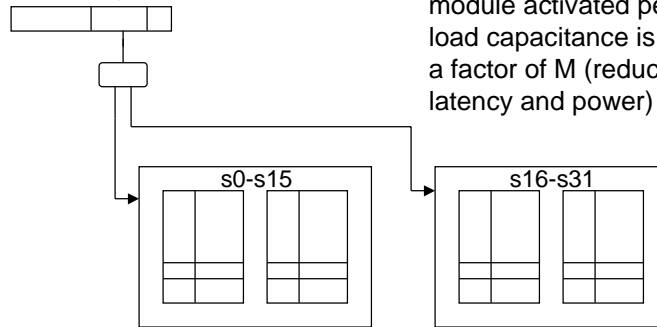
reduce # of active bit lines

reduce capacitive loading on word and bit lines



Multidivided Module Cache

Address issued by CPU



With M modules and only one module activated per cycle, load capacitance is reduced by a factor of M (reduces both latency and power)

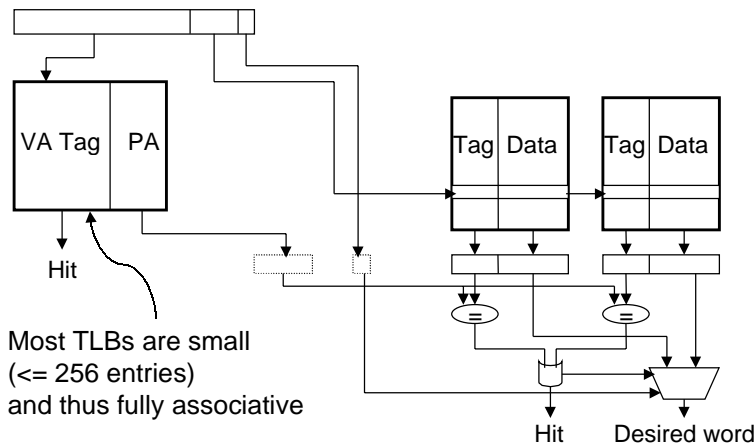
Can combine multidivided module, buffering, and subbanking or divided word line to get the savings of all three

Translation Lookaside Buffers

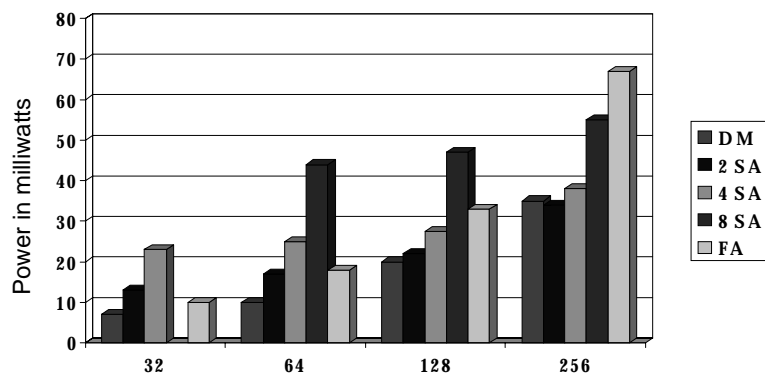
- Small caches to speed up address translation in processors with virtual memory
- All addresses have to be translated before cache access
 - » DEC SA-110 ($2.0V_{dd}$, 0.35μ , 233MHz, 1W typ)
 - I\$ (D\$) dissipate 27% (16%) of the total chip power
 - TLB 17% of total chip power
- I\$ can be virtually indexed/virtually tagged

TLB Structure

Address issued by CPU (page size = index bits + byte select bits)



TLB Power



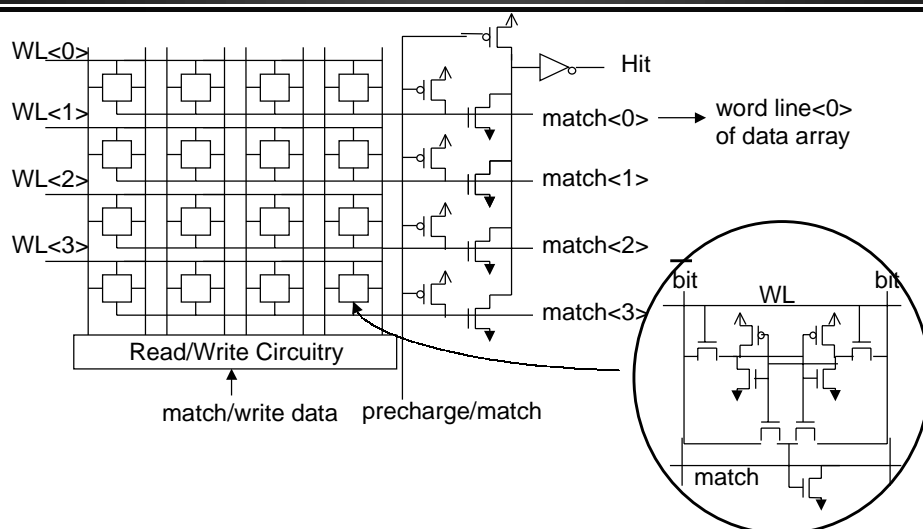
From Juan, 1997

ISCA Tutorial: Low Power Design

Memories.43

©MJIVN, PSU, 2000

CAM Design

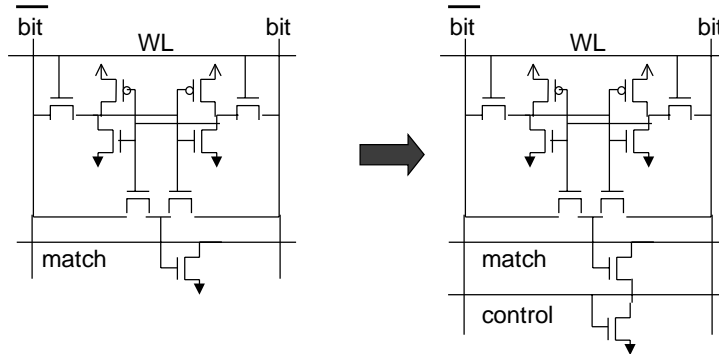


ISCA Tutorial: Low Power Design

Memories.44

©MJIVN, PSU, 2000

Low Power CAM Cell

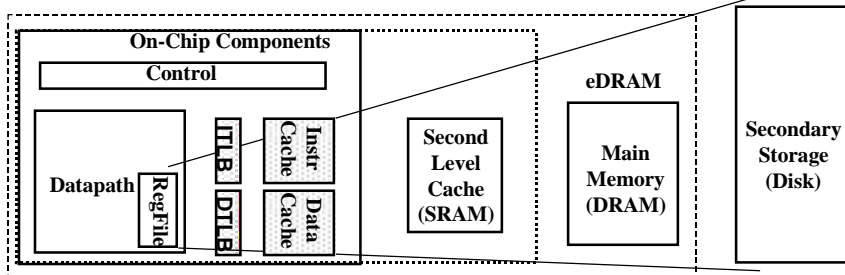


ISCA Tutorial: Low Power Design

Memories.45

©MJIVN, PSU, 2000

Typical Memory Hierarchy



DEC 21164a ($2.0V_{dd}$, 0.35μ , 400MHz, 30W max)

–caches dissipate 25% of the total chip power

DEC SA-110 ($2.0V_{dd}$, 0.35μ , 233MHz, 1W typ) – no L2 on-chip

–I\$ (D\$) dissipate 27% (16%) of the total chip power

ISCA Tutorial: Low Power Design

Memories.46

©MJIVN, PSU, 2000

Low Power DRAMs

- Conventional DRAMs refresh all rows with a fixed single time interval
 - » read/write stalled while refreshing
 - » refresh period $\rightarrow t_{ref}$
 - » DRAM power = $k * (\#read/writes + \#ref)$
- So have to worry about optimizing refresh operation as well

Optimizing Refresh

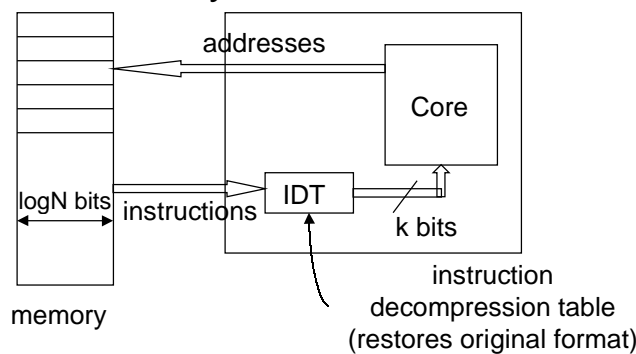
- Selective refresh architecture (SRA)
 - » add a valid bit to each memory row and only refresh rows with valid bit set
 - » reduces refresh 5% to 80%
- Variable refresh architecture (VRA)
 - » data retention time of each cell is different
 - » add a refresh period table and refresh counter to each row and refresh with the appropriate period to each row
 - » reduces refresh about 75%

Application-Specific Memories

- Data and Code Compression
 - » Custom instruction sets: ARM thumb code: interleaving of 32-bit and 16-bit thumb codes
 - » Reduces memory size
 - » Reduces width of off-chip buses
 - » location of compression unit is important
 - » Compress only selective blocks

Hardware Code Compression

- Assuming only a subset of instr's used, replace them with a shorter encoding to reduce memory bandwidth



Other Techniques

- Customizing Memory Hierarchy
 - » Close vs. far memory accesses
 - » Close - faster, less energy consuming, smaller caches
 - » Energy per access increases monotonically with memory size
 - » Automatic memory partitioning

Memory Partitioning

- A memory partition is a set of memory banks that can be independently selected
- Any address is stored into one and only one bank
- The total energy consumed by a partitioned is the sum of the energy consumed by all its banks
- Partitions increasing selection logic cost

Macii, 2000

Scratch Pad Memory

- Use of Scratch Pad Memory instead of Caches for locality
 - » Memory accesses of embedded software are usually very localized
 - » Map most frequent accessed locations onto small on-chip memory
 - » Caches have tag overhead - eliminate by application specific decode logic
 - » Map small set of most frequently accessed addresses to consecutive locations in small memory

Benini 2000

ISCA Tutorial: Low Power Design

Memories.53

©MJl&VN, PSU, 2000

Key References, Memories

- Amrutur, Techniques to Reduce Power in Fast Wide Memories, *Proc. of SLPE*, pp. 92-93, 1994.
- Angel, Survey of Low Power Techniques for ROMs, *Proc. of SLPED*, pp. 7-11, Aug. 1997.
- Chang, Power-Area Trade-Offs in Divided Word Line Memory Arrays, *Journal of Circuits, Systems, Computers*, 7(1):49-57, 1997.
- Evans, Energy Consumption Modeling and Optimization for SRAMs, *IEEE Journal of SSC*, 30(5):571-579, May 1995.
- Itoh, Low Power Memory Design, in *Low Power Design Methodologies*, pp. 201-251, KAP, 1996.
- Ohsawa, Optimizing the DRAM Refresh Count, *Proc. Of SLPED*, pp. 82-87, Aug 1998.
- Shimazaki, An Automatic Power-Save Cache Memory, *Proc. Of SLPE*, pp. 58-56, 1995.
- Yoshimoto, A Divided Word Line Structure in SRAMs, *IEEE Journal of SSC*, 18:479-485, 1983.

ISCA Tutorial: Low Power Design

Memories.54

©MJl&VN, PSU, 2000

Key References, Caches

- Ghose, Reducing Power in SuperScalar Processor Caches Using Subbanking, Multiple Line Buffers and Bit-Line Segmentation, *Proc. of ISLPED*, pp. 70-75, 1999.
- Juan, Reducing TLB Power Requirements, *Proc. of ISLPED*, pp. 196-201, Aug 1997.
- Kin, The Filter Cache: An Energy-Efficient Memory Structure, *Proc. of MICRO*, pp. 184-193, Dec. 1997.
- Ko, Energy Optimization of Multilevel Cache Architectures, *IEEE Trans. On VLSI Systems*, 6(2):299-308, June 1998.
- Panwar, Reducing the Frequency of Tag Compares for Low Power I\$ Designs, *Proc. of ISLPD*, pp. 57-62, 1995.
- Shimazaki, An Automatic Power-Save Cache Memory, *Proc. of SLPE*, pp. 58-59, 1995.
- Su, Cache Design Tradeoffs for Power and Performance Optimization, *Proc. of ISLPD*, pp. 63-68, 1995.