
Samuel Rodriguez
Ph.D. Proposal

University of
Maryland

Department of
Electrical and
Computer
Engineering

# Comparative Analysis of Contemporary Cache Power Reduction Techniques

## Ph.D. Dissertation Proposal

Samuel V. Rodriguez

Samuel Rodriguez
Ph.D. Proposal

University of Maryland

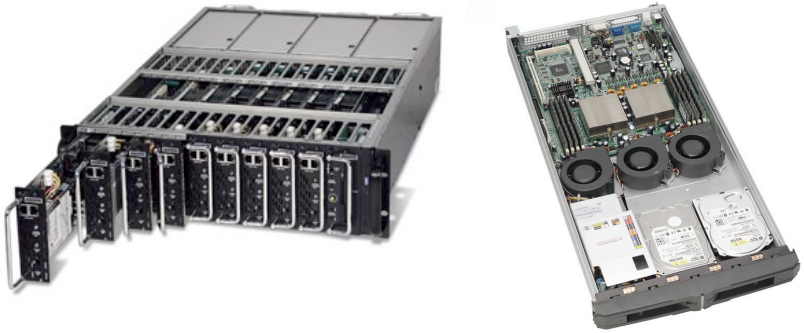
Department of Electrical and Computer Engineering

# Motivation

Power dissipation is important *across the board*, not just portable devices!!

Portable Devices

Mid-end (e.g. Desktops)



High-end

(e.g. servers)

Samuel Rodriguez
Ph.D. Proposal

University of Maryland

Department of Electrical and Computer Engineering

# Motivation

*-Thermal Design Power (TDP) is now a priority specification*

*-AMD currently can't compete in "Thin and light" notebooks because of their higher TDP's*

*-AMD's power advantage in initial dual-core offerings*

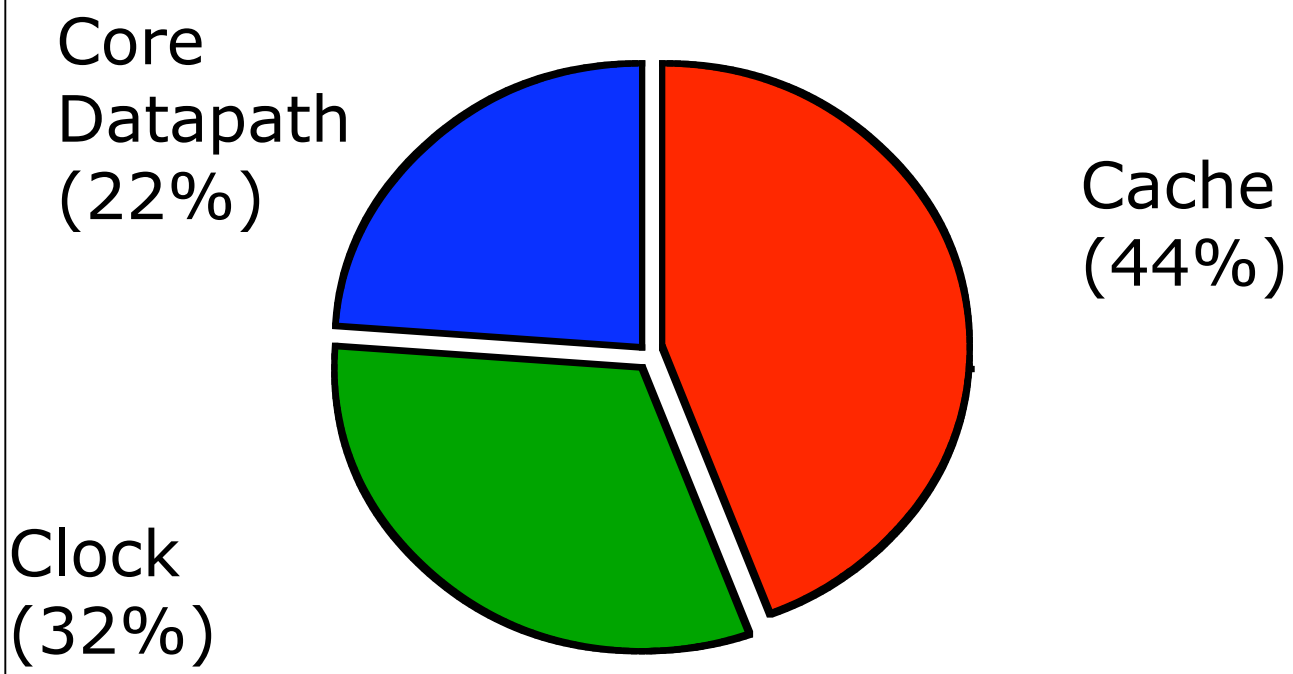
*-An entire Intel Pentium 4 design recently cancelled because of higher than expected TDP's*

Samuel Rodriguez
Ph.D. Proposal

University of Maryland

Department of Electrical and Computer Engineering

# Motivation




- Breakdown of power consumption for a 4-wide 200MHz 3.3V 0.35um processor with 32kB/32kB/1MB caches

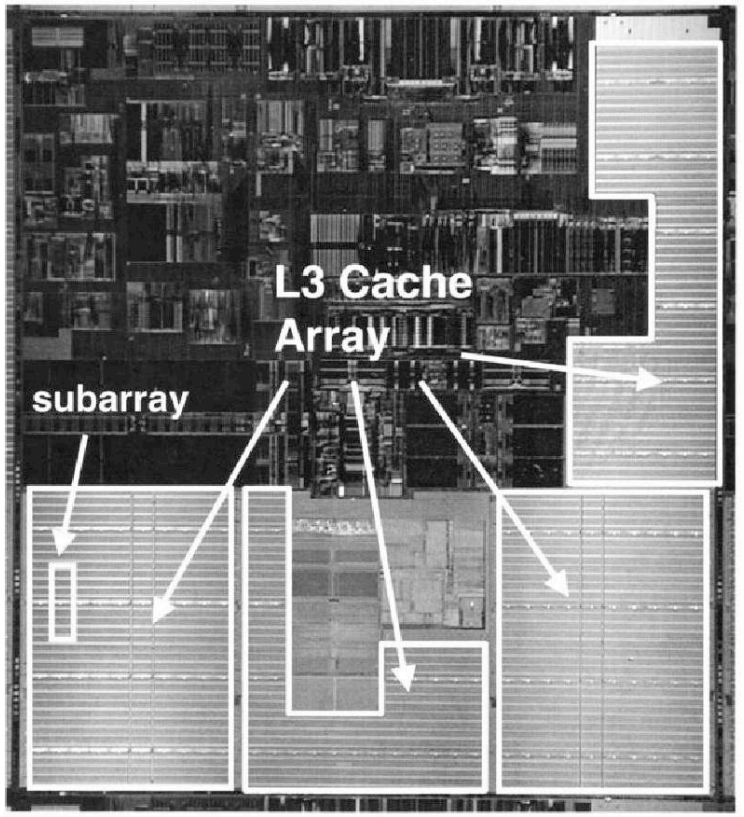
Photograph taken from Gurumurthi

Samuel Rodriguez
Ph.D. Proposal

University of Maryland

Department of Electrical and Computer Engineering

# Motivation




Itanium Die Photo

- Fraction of die area and xsistor count dedicated to caches is increasing

Photograph taken from Weiss2002

Samuel Rodriguez
Ph.D. Proposal

University of Maryland

Department of Electrical and Computer Engineering

# Presentation Outline

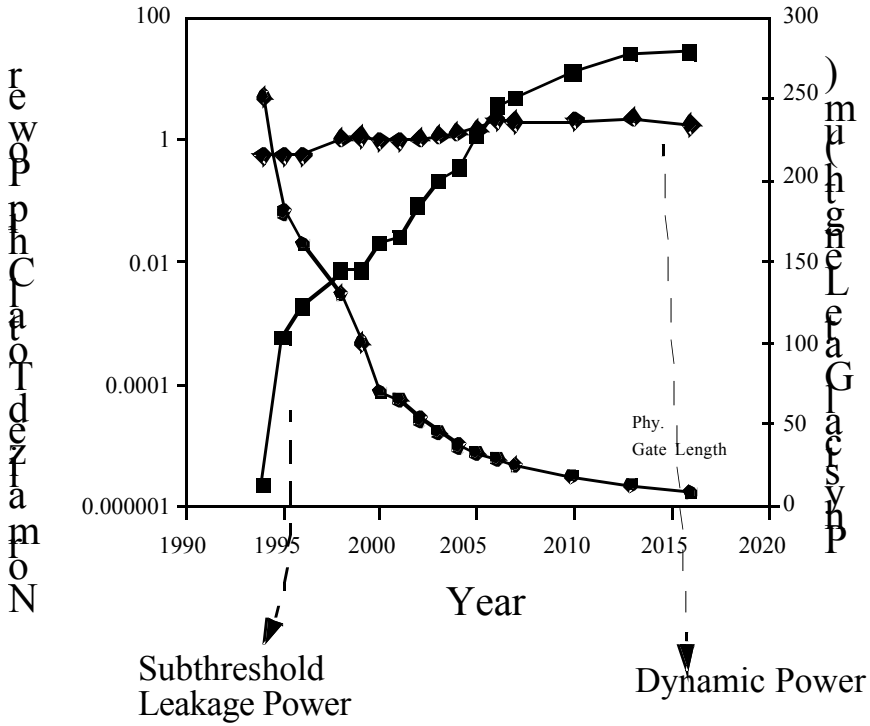
- Motivation (finished)
- Background
  - Power Dissipation
  - Cache/SRAM Implementation
- Contemporary Cache Power Reduction Schemes
- Proposed Work
- Q&A

Samuel Rodriguez
Ph.D. Proposal

University of Maryland

Department of Electrical and Computer Engineering

# Background (Power Dissipation)




- Need to account for both dynamic and static power dissipation!

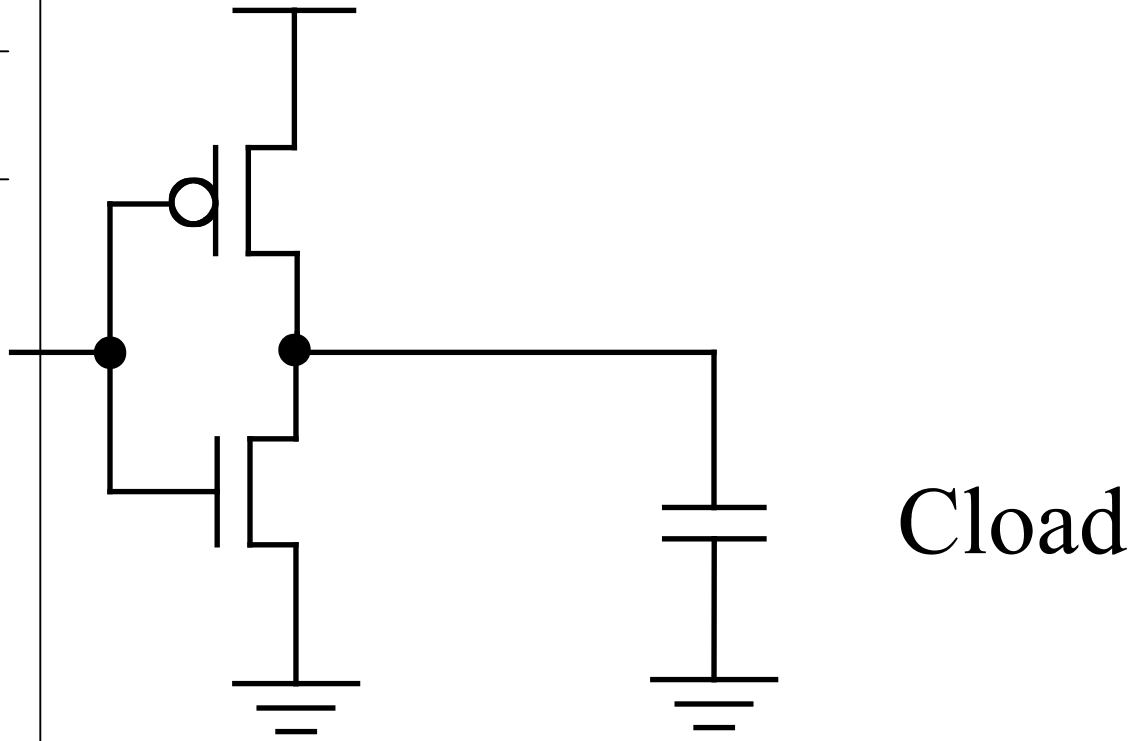
Graph from Kim2004

Samuel Rodriguez
Ph.D. Proposal

University of Maryland

Department of Electrical and Computer Engineering

# Background (Power Dissipation)



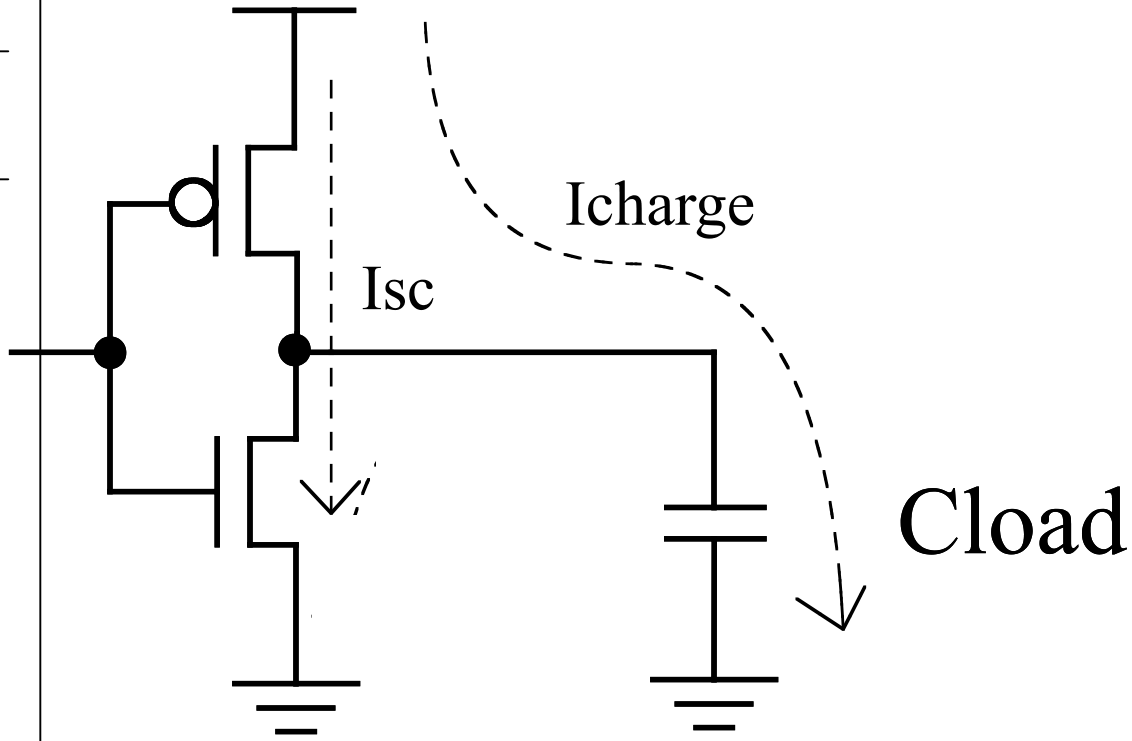

- Causes of dynamic power

Samuel Rodriguez
Ph.D. Proposal

University of Maryland

Department of Electrical and Computer Engineering

# Background (Power Dissipation)



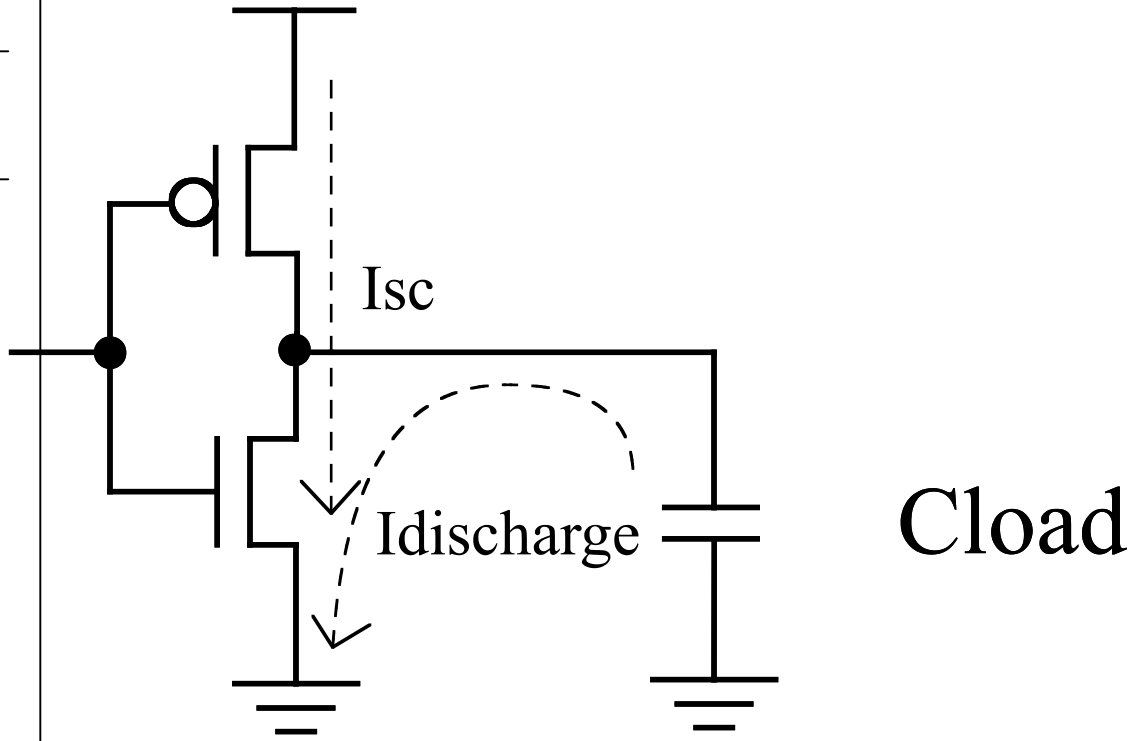

- Causes of dynamic power

Samuel Rodriguez
Ph.D. Proposal

University of Maryland

Department of Electrical and Computer Engineering

# Background (Power Dissipation)




- Causes of dynamic power

Samuel Rodriguez
Ph.D. Proposal

University of Maryland

Department of Electrical and Computer Engineering

# Background (Power Dissipation)

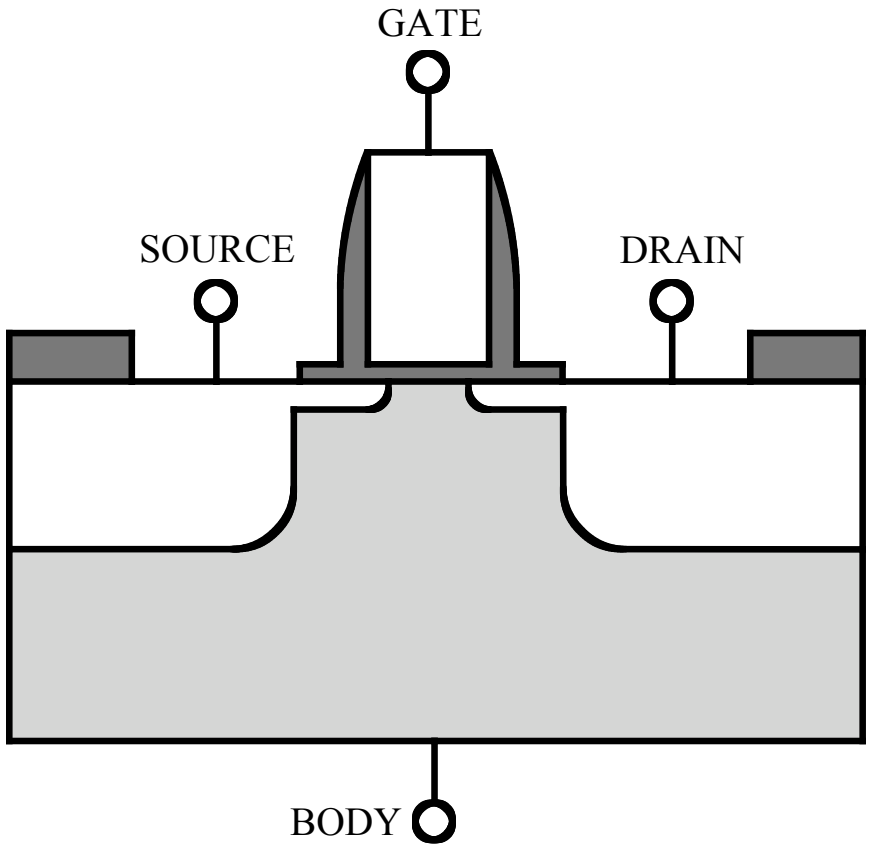
- $Power_{dyn} \propto N \times C \times V_{DD}^2 \times f$
  - ↑↑↑ N : Number of transistors
  - ↓↓ C : Device capacitance
  - ↓ $V_{DD}$ : supply voltage
  - ↑↑ f: Frequency
- Dynamic power trend: slow increase

Samuel Rodriguez
Ph.D. Proposal

University of Maryland

Department of Electrical and Computer Engineering

# Background (Power Dissipation)



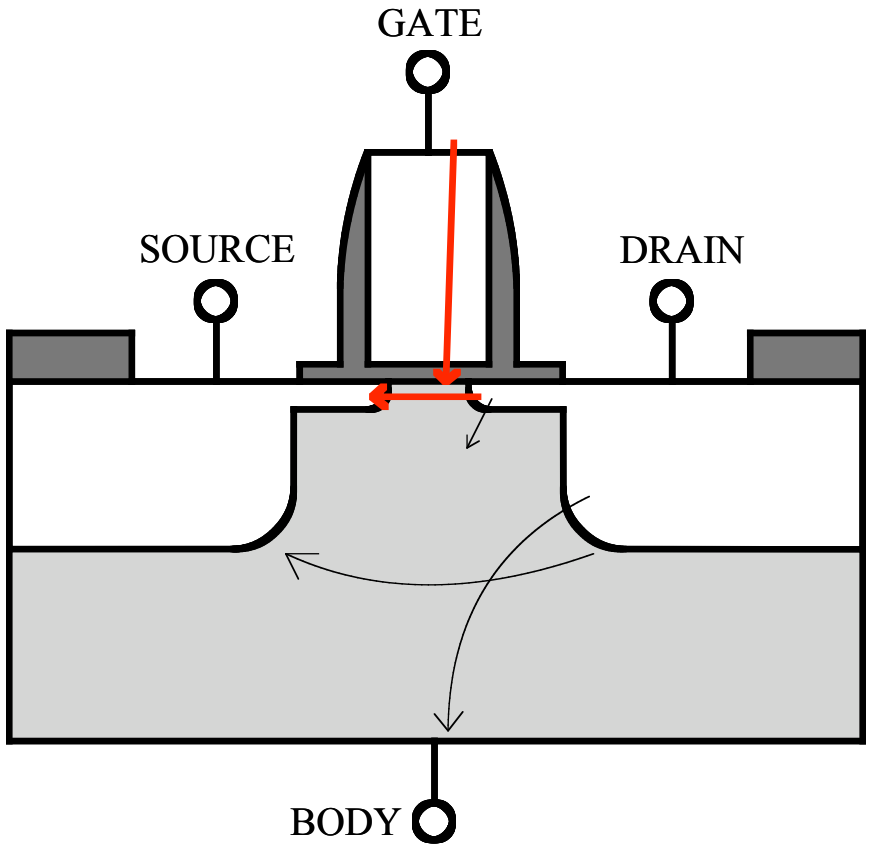

- Causes of static power: leakage currents

Samuel Rodriguez
Ph.D. Proposal

University of Maryland

Department of Electrical and Computer Engineering

# Background (Power Dissipation)




- Subthreshold: 5x per generation
- Gate leakage: *500x* per generation!!!

Samuel Rodriguez
Ph.D. Proposal

University of Maryland

Department of Electrical and Computer Engineering

# Background (Power Dissipation)

- Subthreshold leakage is increasing:

$$I_{d,sat} \propto (V_{gs} - V_{th}) = (V_{DD} - V_{th})$$

- Increase: 5x per generation

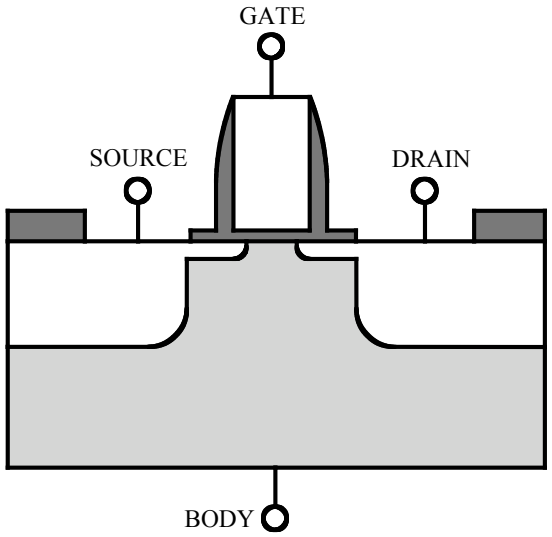
Samuel Rodriguez
Ph.D. Proposal

University of Maryland

Department of Electrical and Computer Engineering

# Background (Power Dissipation)

- Gate leakage




Tox scaling resulting in increased gate leakage caused by oxide tunneling

- Gate leakage: *500x* per generation!!!


Samuel Rodriguez
Ph.D. Proposal

University of Maryland

Department of Electrical and Computer Engineering

# Presentation Outline

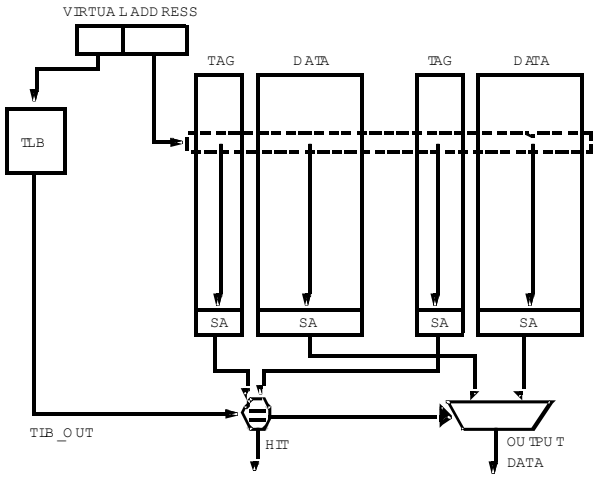
- Motivation (finished)
- Background
  - Power Dissipation
  - Cache/SRAM Implementation
- Contemporary Cache Power Reduction Schemes
- Proposed Work
- Q&A

Samuel Rodriguez
Ph.D. Proposal

University of Maryland

Department of Electrical and Computer Engineering

# Background (Cache Implementation)





2-way set-associative Cache Read

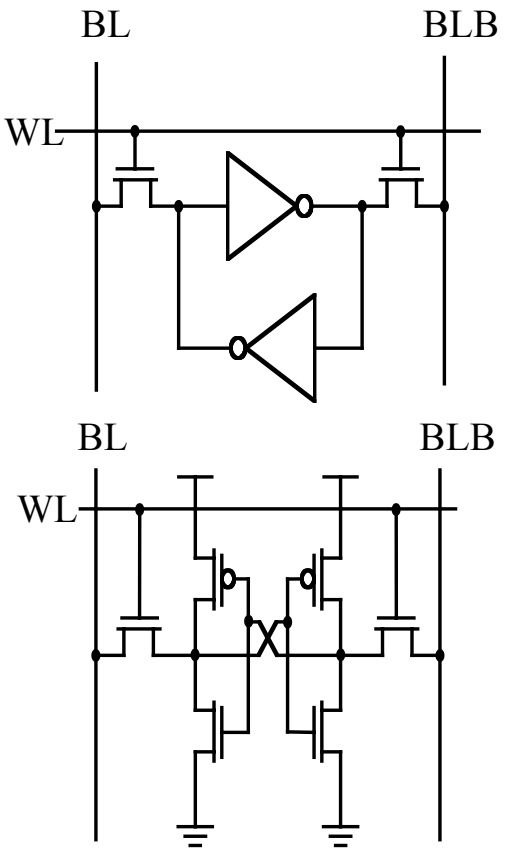
*Note: (external signal)*

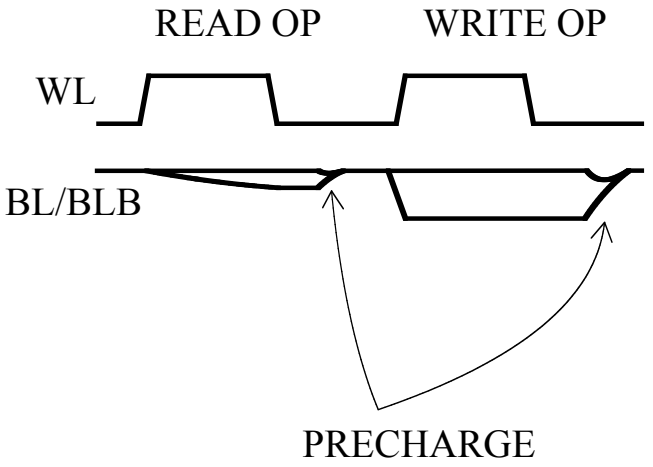
Samuel Rodriguez
Ph.D. Proposal

University of Maryland

Department of Electrical and Computer Engineering

# Background (Cache Implementation)



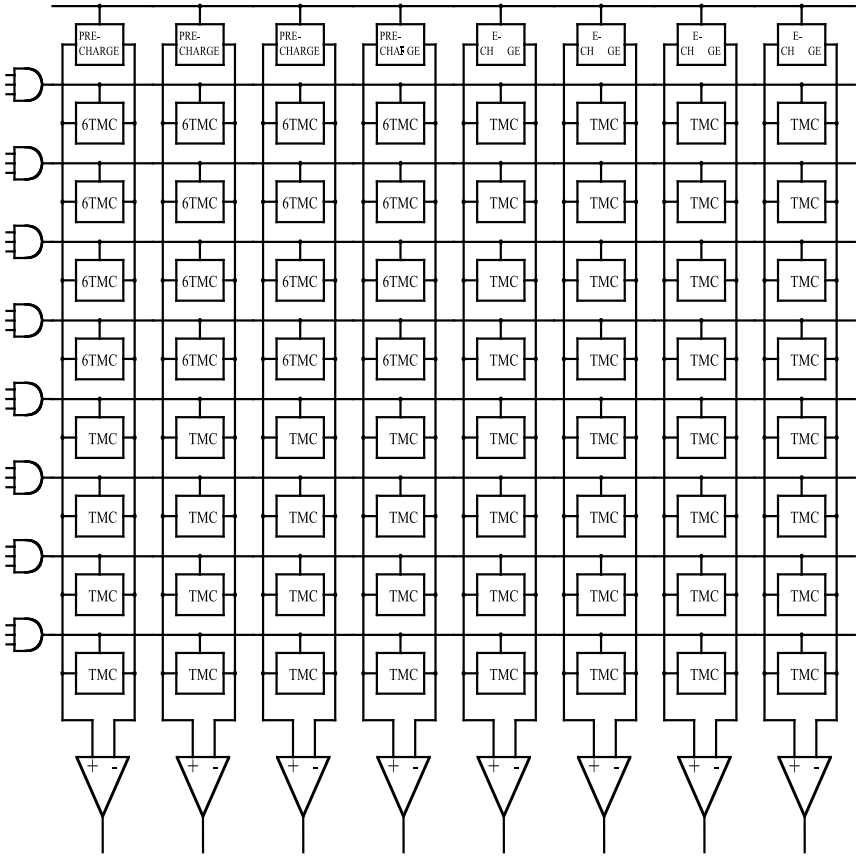

- Full CMOS 6T Memory Cell

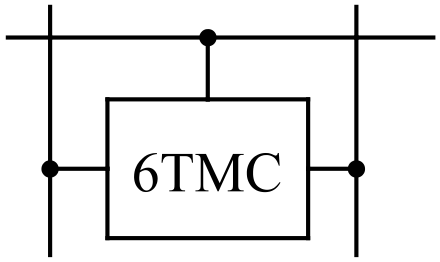
Samuel Rodriguez
Ph.D. Proposal

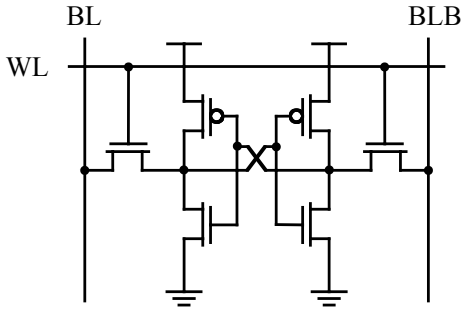
University of Maryland

Department of Electrical and Computer Engineering

# Background (Cache Implementation)



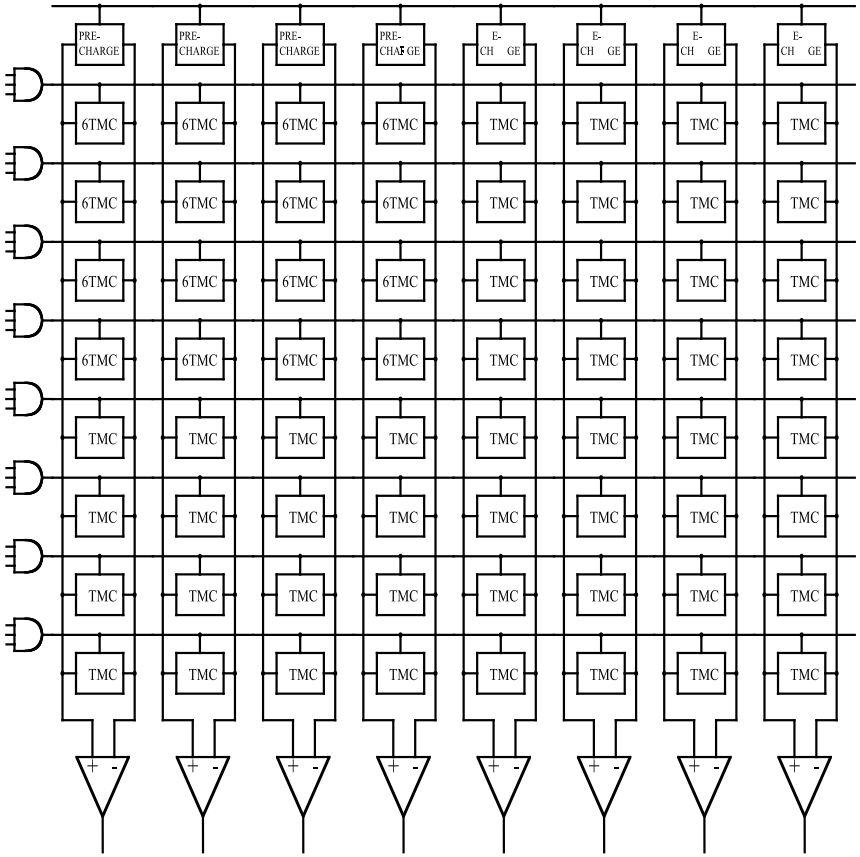

Simplified 8 x 8b SRAM array

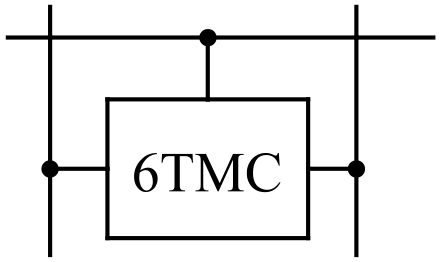
Samuel Rodriguez
Ph.D. Proposal

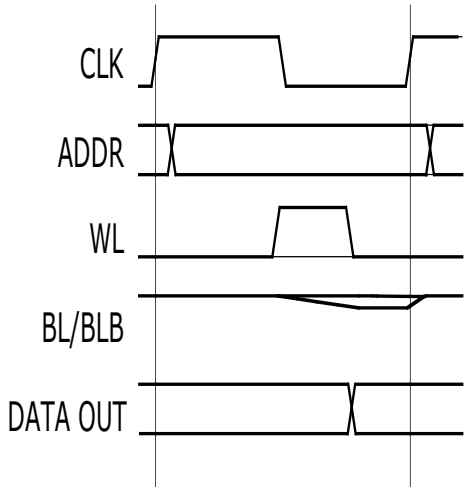
University of Maryland

Department of Electrical and Computer Engineering

# Background (Cache Implementation)



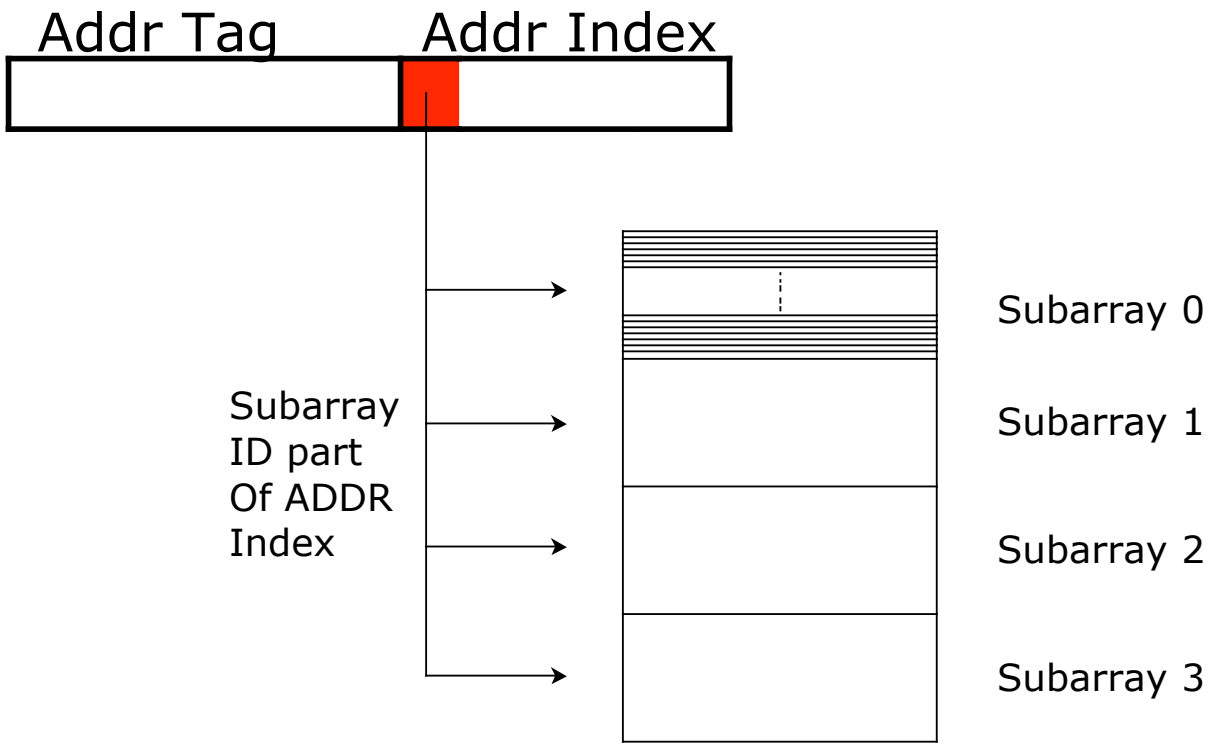

Simplified 8 x 8b SRAM array

Samuel Rodriguez
Ph.D. Proposal

University of Maryland

Department of Electrical and Computer Engineering

# Background (Cache Implementation)




SRAM partitioning: array is often divided into smaller "subarrays"

Samuel Rodriguez
Ph.D. Proposal

University of Maryland

Department of Electrical and Computer Engineering

# Presentation Outline

- Motivation (finished)
- Background (finished)
  - Power Dissipation
  - Cache/SRAM Implementation
- Contemporary Cache Power Reduction Schemes
- Proposed Work
- Q&A

Samuel Rodriguez
Ph.D. Proposal

University of Maryland

Department of Electrical and Computer Engineering

# Cache Power Reduction Techniques

| Scheme | Dynamic / Static? | Est. Power Savings | Exec-time increase? | State-Retentive? |
|---|---|---|---|---|
| Gated-Vdd | Static | N/A * | YES | NO |
| Cache decay | Static | 80% | YES | NO |
| DRG-cache | Static | 39%-59% | NO | YES |
| Drowsy cache | Static | 60-75% | YES | YES |
| Near-OPT precharge | Static | N/A ** | YES | N/A |
| Way-halting | Dynamic | 55% | NO | N/A |
| Data size detection | Dynamic | N/A | NO | N/A |

* - paper only cites 62% energy-delay savings

** - paper only cites 92% reduction of bitline discharge

Samuel Rodriguez
Ph.D. Proposal

University of Maryland

Department of Electrical and Computer Engineering

# Cache Power Reduction Techniques (cont…)

| Scheme | Miss Ratio increase? | Access time increase? | Variable load-hit latency? | μARCH transparent ? | Additional noise problems? |
|---|---|---|---|---|---|
| Gated-Vdd | YES | YES | NO | NO | NO |
| Cache decay | YES | YES | NO | NO | NO |
| DRG-cache | NO | YES | NO | YES | YES |
| Drowsy cache | NO | YES | YES | NO | YES |
| Near-OPT precharge | NO | NO* | YES | NO | NO |
| Way-halting | NO | NO* | NO | YES | NO |
| Data size detection | NO | YES | NO | YES | NO |

* - *With proper design*

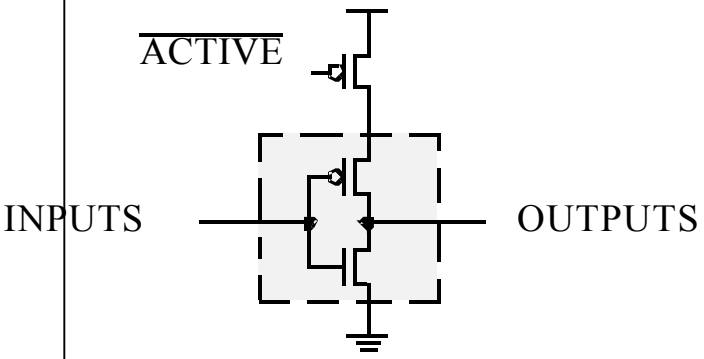
Samuel Rodriguez
Ph.D. Proposal

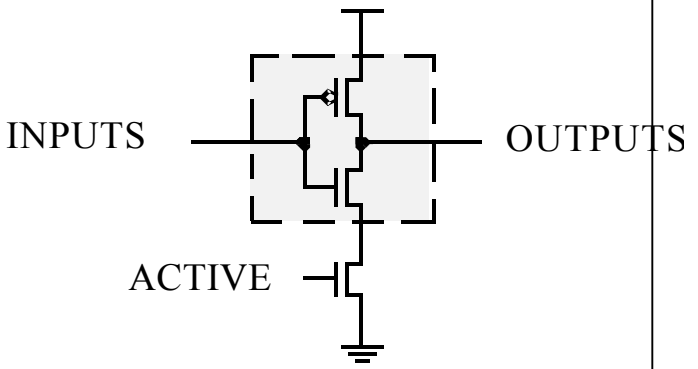
University of Maryland

Department of Electrical and Computer Engineering

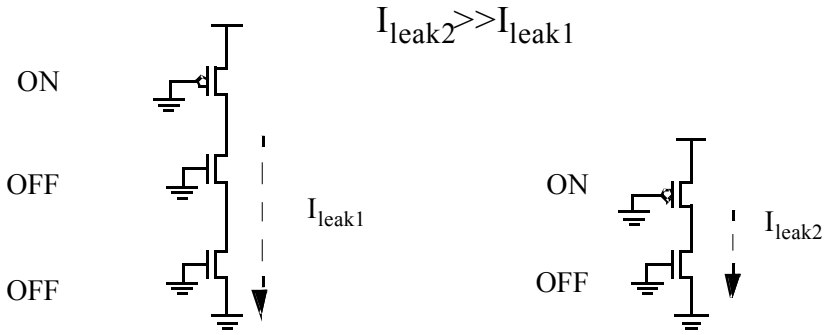
# Cache Power Reduction Techniques

## First four techniques: Supply Gating




## Stacking Effect:



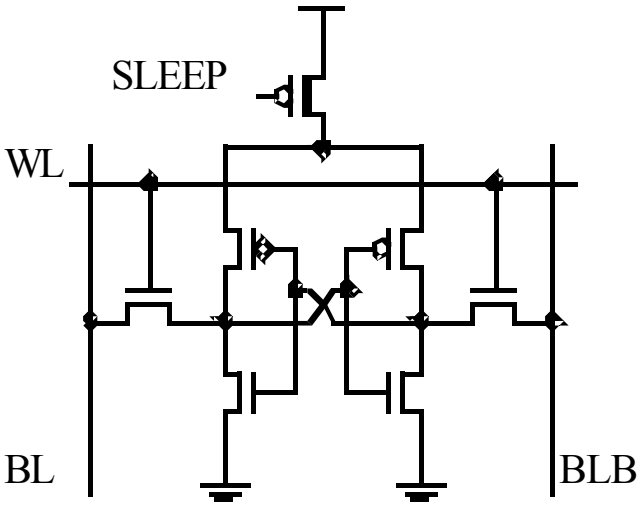
Samuel Rodriguez
Ph.D. Proposal

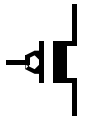
University of Maryland

Department of Electrical and Computer Engineering

# Cache Power Reduction Techniques

## 1. Gated-Vdd (circuit)



High-Vt PMOS

-6TMC can be disabled by the gating transistor, resulting in less leakage

Samuel Rodriguez
Ph.D. Proposal

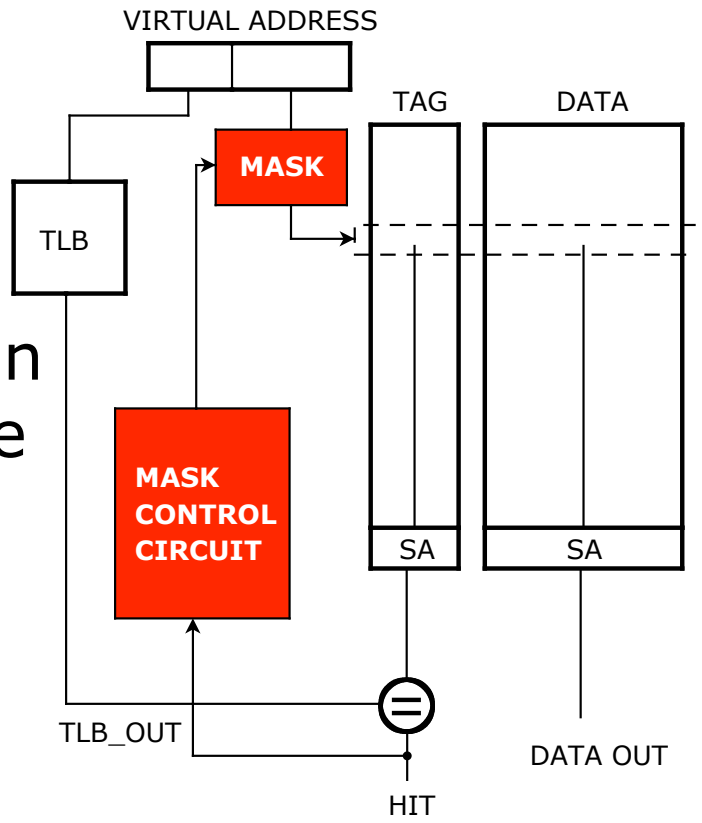
University of Maryland

Department of Electrical and Computer Engineering

# Cache Power Reduction Techniques

## 1. Gated-Vdd (microarchitecture : Dynamically ResIzable [DRI] Cache)

- Mask out part of the index to dynamically resize the cache
- Make this decision based on the cache Hit ratio
- Energy-delay reduced by 62%



Samuel Rodriguez
Ph.D. Proposal

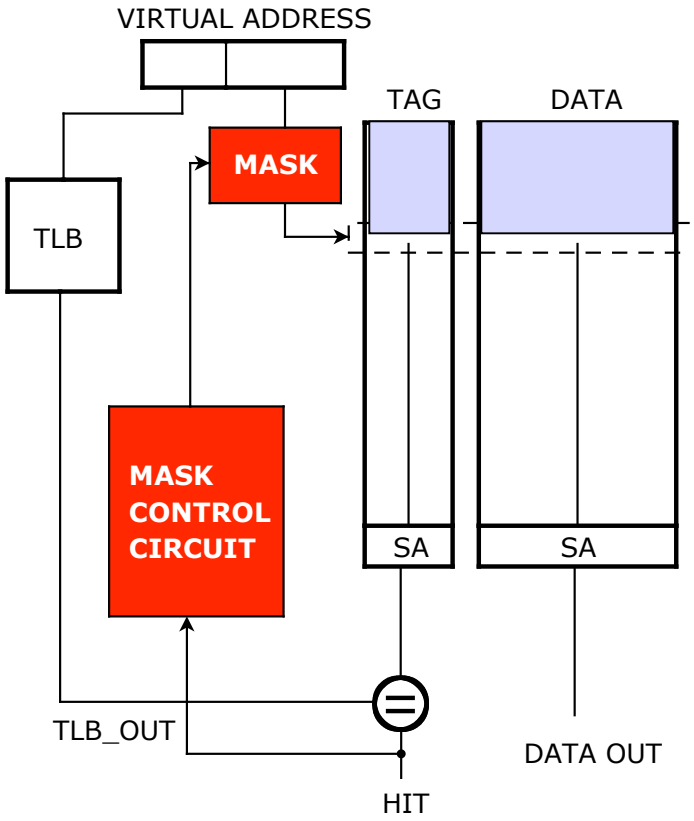
University of Maryland

Department of Electrical and Computer Engineering

# Cache Power Reduction Techniques

## 1. Gated-Vdd (microarchitecture : Dynamically ResIzable [DRI] Cache)

Example:
If MASK removes the upper 2 bits of the index, only the lower _ sets of the cache can be accessed (all other sets are gated off)



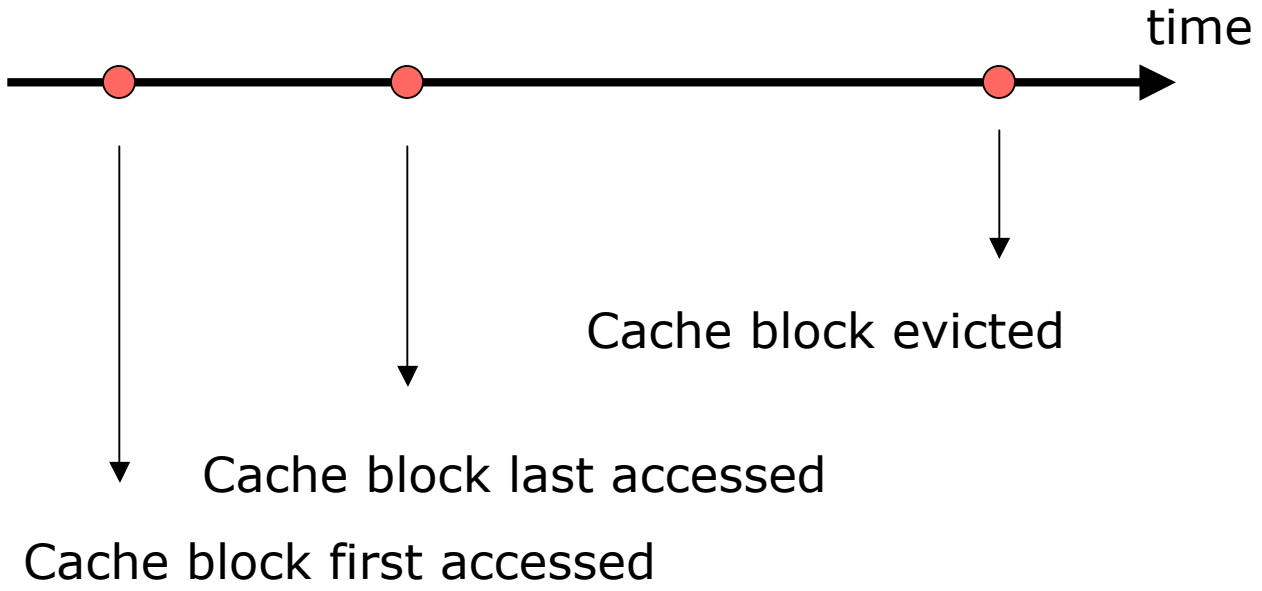
Samuel Rodriguez
Ph.D. Proposal

University of Maryland

Department of Electrical and Computer Engineering

# Cache Power Reduction Techniques

## 2. Cache decay (concept)




-If we turn a cache block's power off right after it is last accessed, we save leakage power *without* any performance penalty

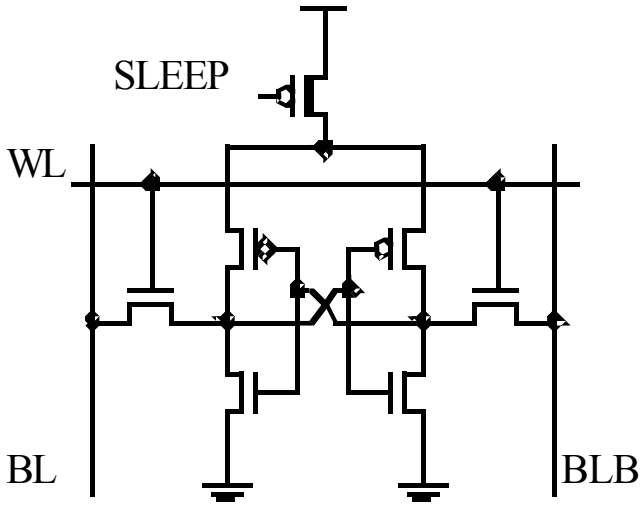
Samuel Rodriguez
Ph.D. Proposal

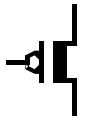
University of Maryland

Department of Electrical and Computer Engineering

# Cache Power Reduction Techniques

## 2. Cache decay (circuit borrows Gated-Vdd techniques)



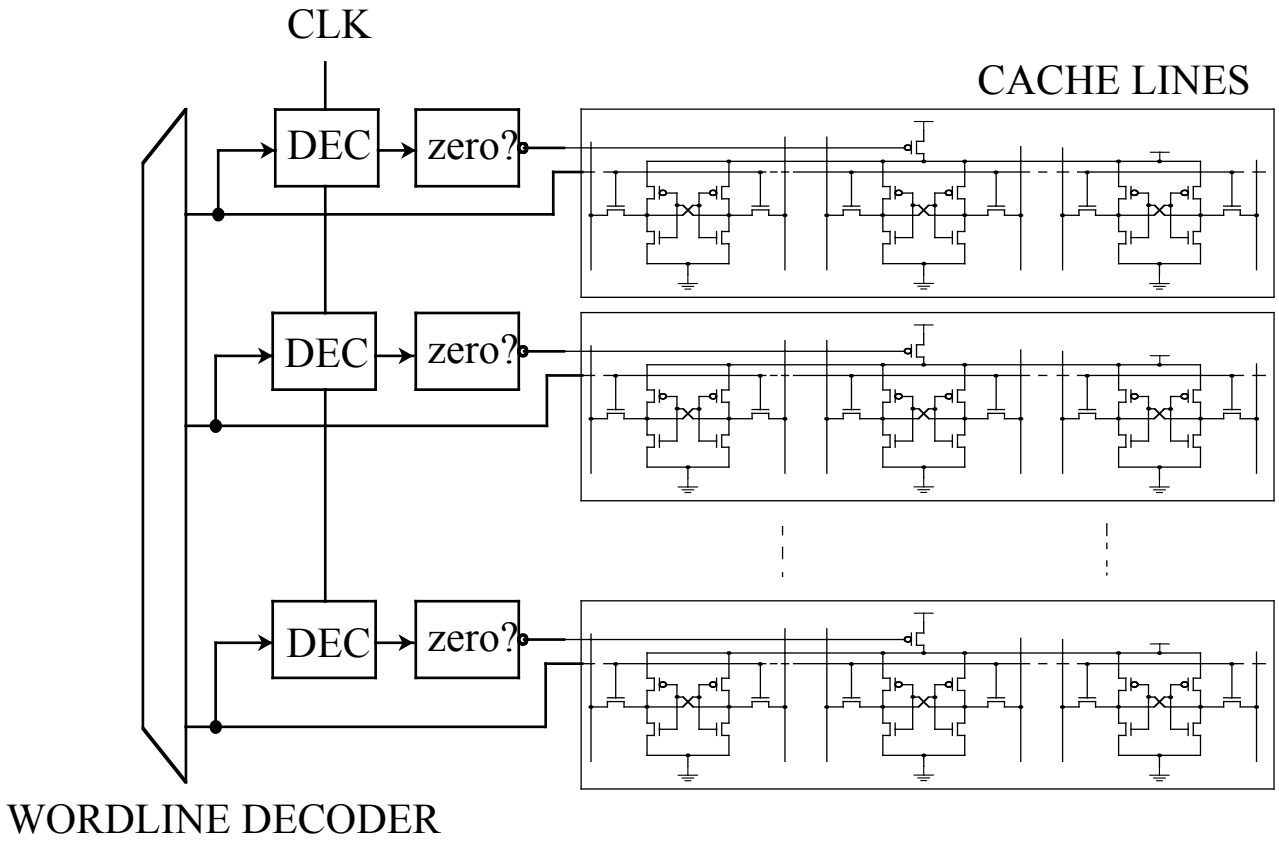
Samuel Rodriguez
Ph.D. Proposal

University of Maryland

Department of Electrical and Computer Engineering

# Cache Power Reduction Techniques

## 2. Cache decay (microarchitecture)




- Static power reduced by 80%!!

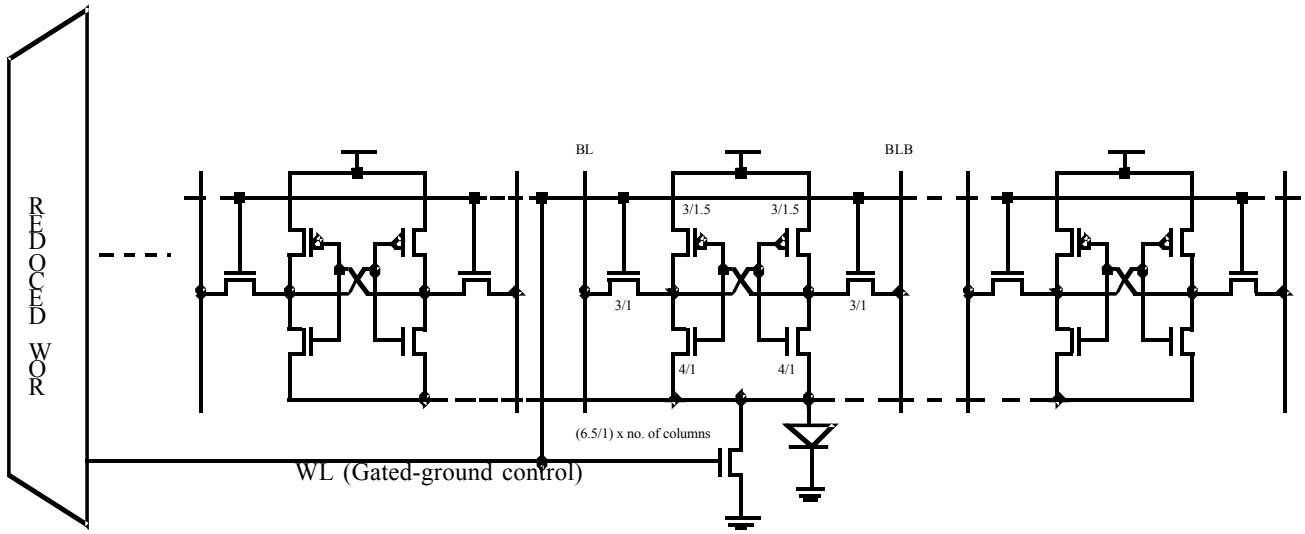
Samuel Rodriguez
Ph.D. Proposal

University of Maryland

Department of Electrical and Computer Engineering

# Cache Power Reduction Techniques

## 3. Data Retention Ground (DRG) (circuit)




-DRG gates the ground of the MC's
-With careful sizing, state can be preserved!
-Technique is transparent!!
-Power is reduced by 39% to 59%

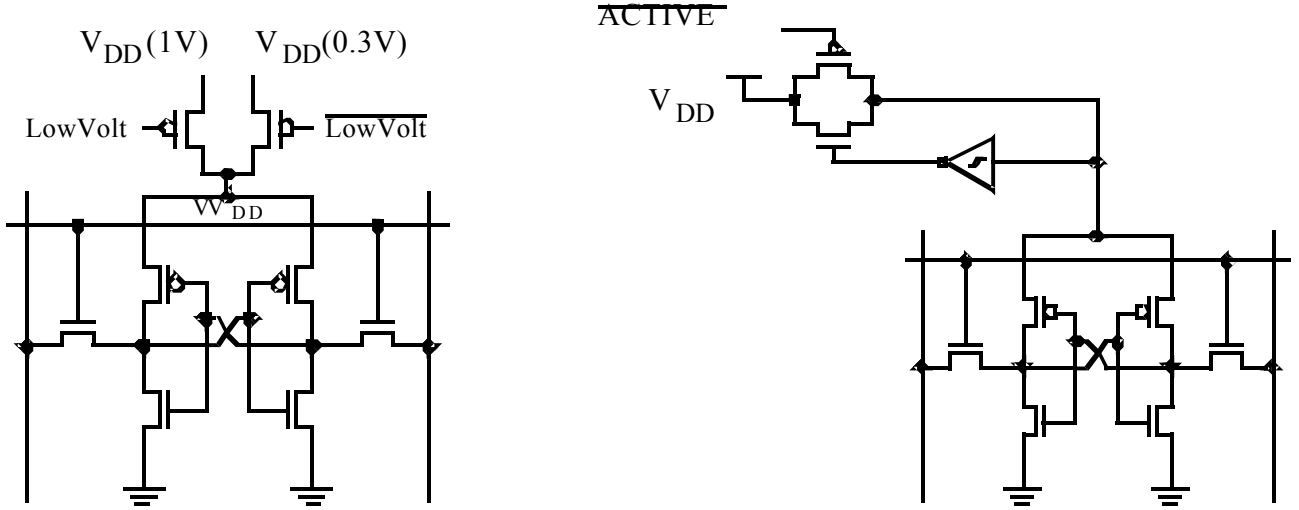
Samuel Rodriguez
Ph.D. Proposal

University of
Maryland

Department of
Electrical and
Computer
Engineering

# Cache Power Reduction Techniques

## 4. Drowsy caches (circuit)




- Drowsy caches (microarchitecture) : Simple algorithm – periodically put *every* cache line into drowsy mode
- Static power reduced by 60% to 75%

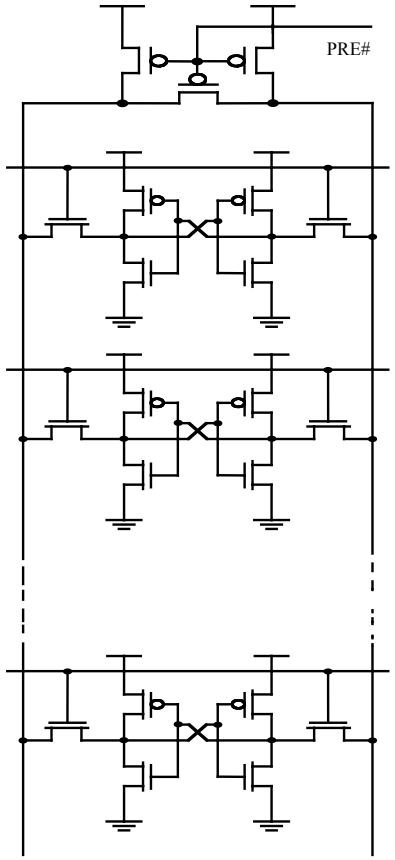
Samuel Rodriguez
Ph.D. Proposal

University of Maryland

Department of Electrical and Computer Engineering

# Cache Power Reduction Techniques

## 5. Near-optimal Precharging




-bitline leakage burns power even in unused cache subarray (additional power is needed during the precharge phase)

-For a given time interval, only a small fraction of subarrays are actually used

-Bitline discharge reduced by 92%

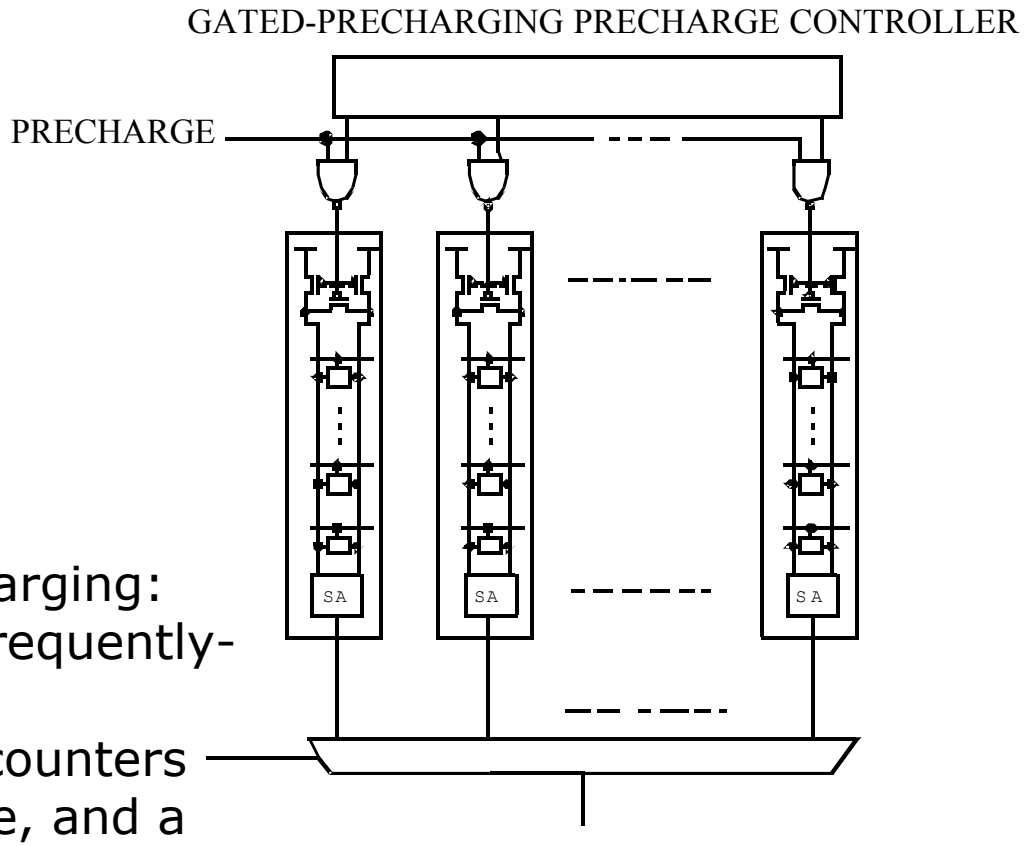
Samuel Rodriguez
Ph.D. Proposal

University of Maryland

Department of Electrical and Computer Engineering

# Cache Power Reduction Techniques

## 5. Near-optimal Precharging




-Near-optimal precharging: stop precharging infrequently-used subarrays
-Microarchitecture: counters to track subarray use, and a system to handle variable load-hit latency

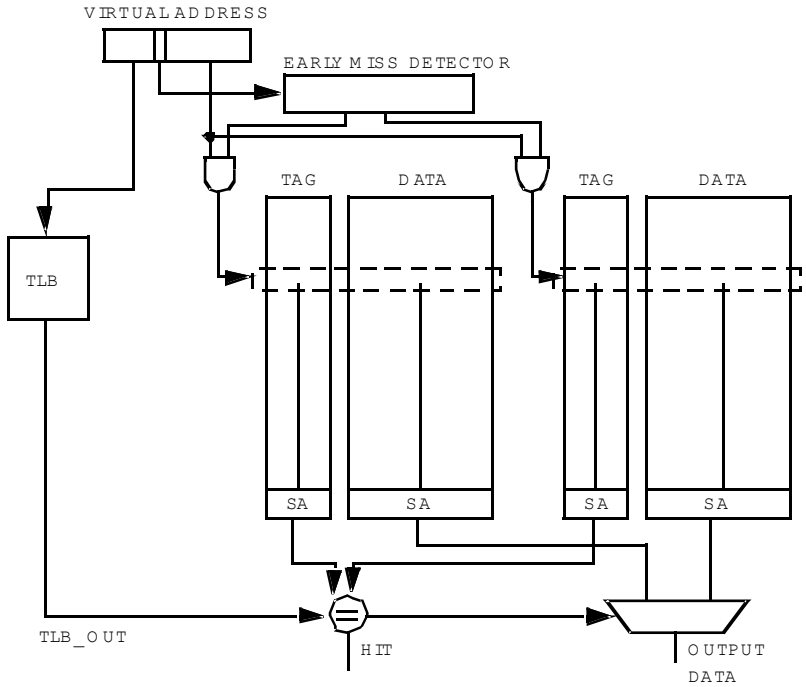
Samuel Rodriguez
Ph.D. Proposal

University of Maryland

Department of Electrical and Computer Engineering

# Cache Power Reduction Techniques

## 6. Way-halting cache




-Perform early miss detection to stop access to cache ways that are certain to miss

-Early miss detection performed by offloading a few tag bits into a faster array that performs tag comparison early in the access

-Power reduced by 55%

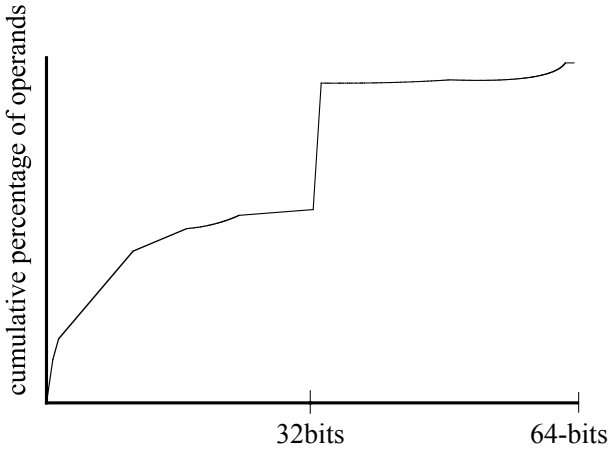
Samuel Rodriguez
Ph.D. Proposal

University of Maryland

Department of Electrical and Computer Engineering

# Cache Power Reduction Techniques

## 7. Data Size Detection




-Not every operand uses up the maximum space provided by the wordlength
(e.g. ~94% of the operands in 64-bit Alpha SpecInt95 benchmarks use 32-bit or less)

-Keep track of this information to turn off the upper bits of the datapath (saving on wordline, bitline and sense-amp power)

Plot from Brooks and Martonosi

Samuel Rodriguez
Ph.D. Proposal

University of Maryland

Department of Electrical and Computer Engineering

# Cache Power Reduction Techniques

| Scheme | Dynamic / Static? | Est. Power Savings | Exec-time increase? | State-Retentive? |
|---|---|---|---|---|
| Gated-Vdd | Static | N/A * | YES | NO |
| Cache decay | Static | 80% | YES | NO |
| DRG-cache | Static | 39%-59% | NO | YES |
| Drowsy cache | Static | 60-75% | YES | YES |
| Near-OPT precharge | Static | N/A ** | YES | N/A |
| Way-halting | Dynamic | 55% | NO | N/A |
| Data size detection | Dynamic | N/A | NO | N/A |

* - paper only cites 62% energy-delay savings

** - paper only cites 92% reduction of bitline discharge

Samuel Rodriguez
Ph.D. Proposal

University of Maryland

Department of Electrical and Computer Engineering

# Cache Power Reduction Techniques (cont…)

| Scheme | Miss Ratio increase? | Access time increase? | Variable load-hit latency? | μARCH transparent ? | Additional noise problems? |
|---|---|---|---|---|---|
| Gated-Vdd | YES | YES | NO | NO | NO |
| Cache decay | YES | YES | NO | NO | NO |
| DRG-cache | NO | YES | NO | YES | YES |
| Drowsy cache | NO | YES | YES | NO | YES |
| Near-OPT precharge | NO | NO* | YES | NO | NO |
| Way-halting | NO | NO* | NO | YES | NO |
| Data size detection | NO | YES | NO | YES | NO |

* - *With proper design*

Samuel Rodriguez
Ph.D. Proposal

University of Maryland

Department of Electrical and Computer Engineering

# Presentation Outline

- Motivation (finished)
- Background (finished)
  - Power Dissipation
  - Cache/SRAM Implementation
- Contemporary Cache Power Reduction Schemes
- Proposed Work
- Q&A

Samuel Rodriguez
Ph.D. Proposal

University of Maryland

Department of Electrical and Computer Engineering

# Proposed Work

- Detailed comparative study of discussed low-power cache techniques (and various combinations)
- Metrics of comparison:
  - Power dissipation (including overheads)
  - Performance penalty (IPC and access time)
  - Die area overhead
  - Complexity

Samuel Rodriguez
Ph.D. Proposal

University of Maryland

Department of Electrical and Computer Engineering

# Proposed Work

- Contributions
  - Every scheme is put on the same playing field
  - Schemes are made up to date with the use of predictive 65nm/45nm technology
  - Improved evaluation accuracy
    - Gate leakage is now accounted for
    - Careful accounting for overheads
    - Use of a state-of-the-art memory system model
  - Data Size Detection is proposed


Samuel Rodriguez
Ph.D. Proposal

University of Maryland

Department of Electrical and Computer Engineering

# Q & A


Samuel Rodriguez
Ph.D. Proposal

University of Maryland

Department of Electrical and Computer Engineering

# Thank You