Design and Implementation of High Performance Application Specific Memory

- 고성능 Application Specific Memory의 설계와 구현 -

M.S. Thesis

Sungdae Choi

Dec. 20th, 2002

Outline

- Introduction
- Memory for Mobile 3D Graphics
- Content-Addressable-Memory (CAM) for Network Memory
- Conclusion

Outline

- **Introduction**
  - Motivation
  - Related Work
  - This Work
- **Memory for Mobile 3D Graphics**
- **Content-Addressable-Memory (CAM) for Network Memory**
- **Conclusion**

Motivation(1/2)

- **Conventional Memory Architecture**
  - Designed for General Applications
    - Not Optimized for Specific Application
  - Long t_{RC} (60~80ns)
    - For Mass Production & Low Cost
    - Bottleneck of the System Performance
  - Only Read / Write operation

Need for Application Specific Memory

Motivation(2/2)

- **Application Specific Memory**
  - "Offer Unique Characteristics to Improve Performance in Particular Application"*
  - Free Design
    - Flexible Control
    - Guaranteed Bandwidth
    - Low Power Consumption
  - Complex Design
    - Both Memory and System Background

* IEEE Circuits & Devices, "High-Speed Memory Architectures for Multimedia Applications", 1997

Application Specific Memory

- **Cache DRAM (CDRAM)***
  - DRAM Density + SRAM Speed
- **Video RAM (VRAM)***
  - Requirement for High-quality Display
- **Media-Chip****
  - Embedded Memory for PC 3D Graphics
- **Content-Addressable-Memory (CAM)*****
  - High-speed Search Function

* IEEE Circuits & Devices, “High-Speed Memory Architectures for Multimedia Applications”, 1997
** Takao Watanabe, “A Modular Architecture for a 6.4Gbyte/s 8Mb DRAM-Integrated Media Chip”, 1997
*** IEEE Potentials, “RAM versus CAM”, 1997

Thesis Work

- **Memories for Mobile 3D Graphics**
  - Embedded in RamP-IV Processor
  - Implemented using 0.16um DRAM Process
- **Content-Addressable-Memory (CAM) for Network Lookup Engine**
  - Ternary Cell Structure
  - Simulation using 0.35um Logic Process

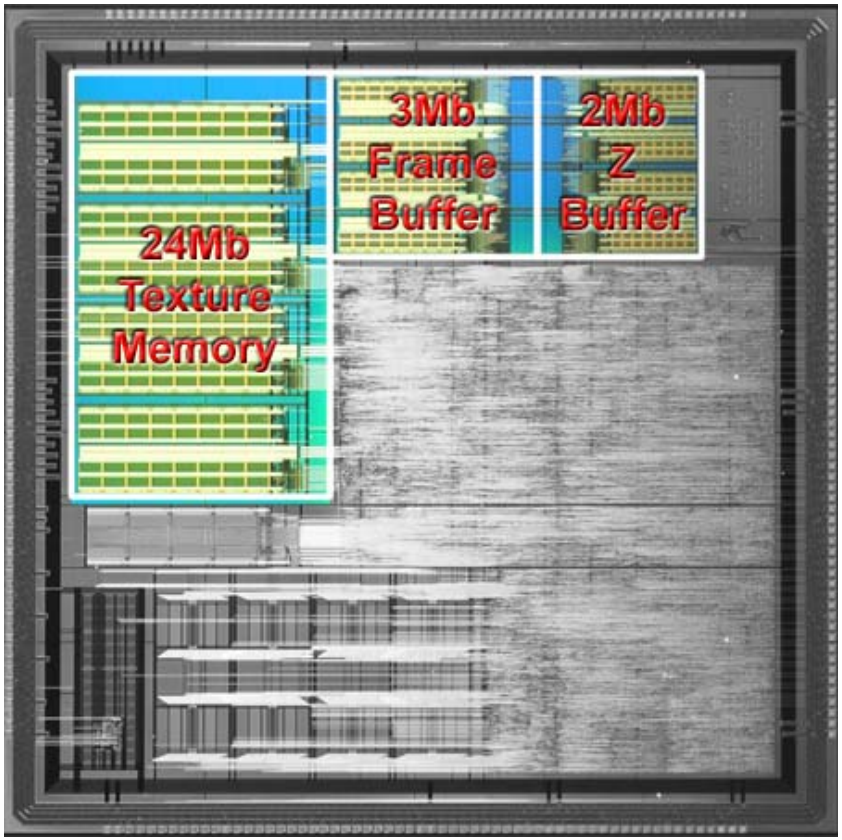
Memory Architecture Contributes the System Performance

Outline

- Introduction
- **Memories for Mobile 3D Graphics**
  - Need for Special Memory
  - Memory Specification
  - Applied Architectures
  - Simulation Results
  - Chip Implementation
- Content-Addressable-Memory (CAM) for Network Memory
- Conclusion

RamP-IV Processor*

- **For Mobile Multimedia Application**

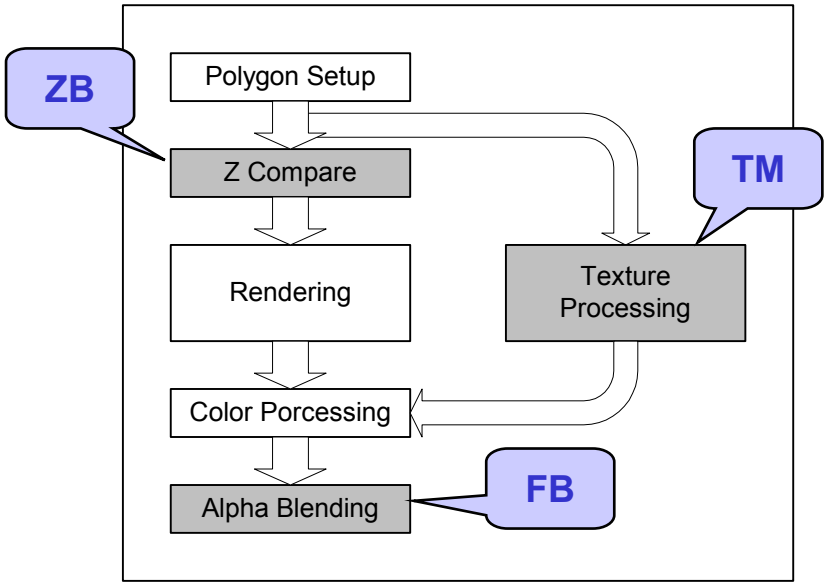


* R. Woo, et. al., "A 210mW Graphics LSI Implementing Full 3D Pipeline with 264Mtexels/s Texturing for Mobile Multimedia Applications“, ISSCC 2003

3D Graphics in RamP-IV

- **Target Performance for High-Quality 3D Graphics**
  - Mobile Application
    - 256 x 256 resolution @ 24bit color
    - 16bit depth-comparison
  - High Image Quality
    - Double Frame Buffering
    - Texture Mapping
  - High Performance
    - Double Pixel Processor
    - 100Mpixel/sec Performance

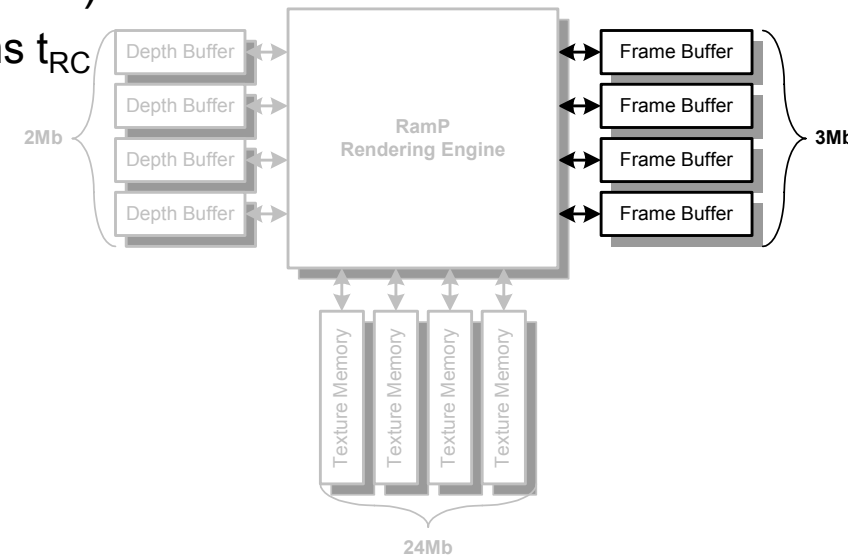
Need for New Memories to Supply the Performance



Memories in the Rendering Engine

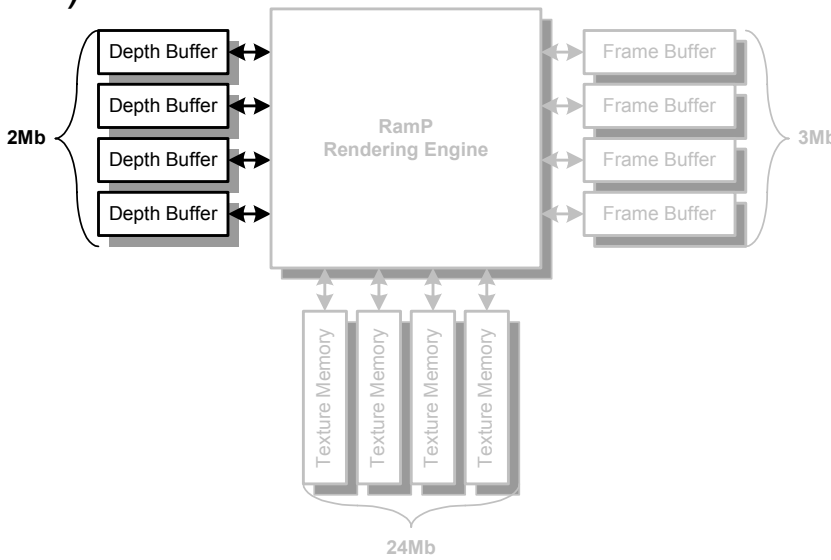
Memories for Target Performance

- **Frame-Buffer**
  - 24bit I/O bus (for 24bit color processing)
  - 3Mb capacity (256 x 256 x 24bit x double FB)
    - Individual 768kb x 4 (for double pixel processing)
  - Read-Modify-Write Operation
    - Simplifies the Pile-line Stage of the Rendering Engine
  - 20ns t_{RC} (for 100Mpixel/s performance)
    - Conventional Memory has 60~80ns t_{RC}



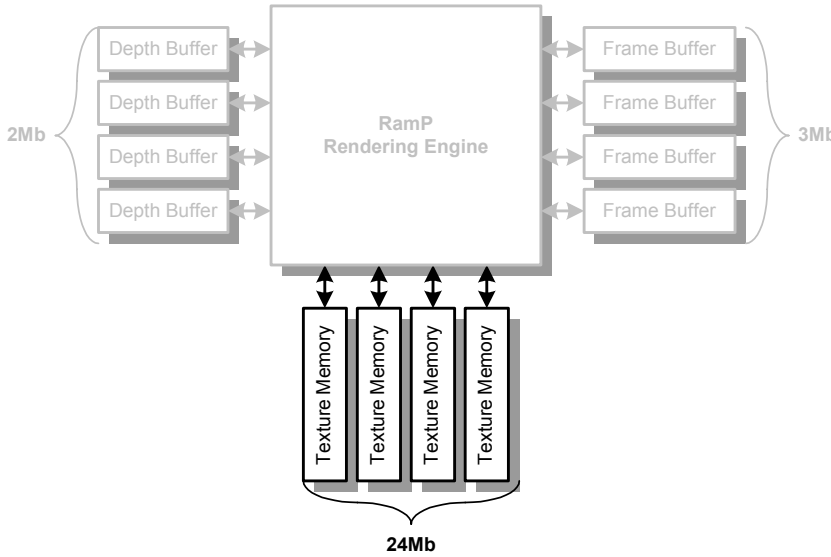
Memories for Target Performance

- **Z-Buffer**
  - 16bit I/O bus (for 16bit depth comparison)
  - 2Mb capacity (256 x 256 x 16bit depth x double FB)
    - Individual 512kb x 4 (for double pixel processing)
  - Read-Modify-Write Operation
    - Simplifies the Pipe-line Stage of the Rendering Engine
  - 20ns t_{RC} (for 100Mpisel/s performance)



Memories for Target Performance

- **Texture Memory**
  - 24bit I/O bus (for 24bit color processing)
  - 20ns t_{RC} (for 200Mtexel/s performance)
  - Read Oriented Operation
  - 24Mb capacity = 6Mb x 4 (for bilinear interpolation)

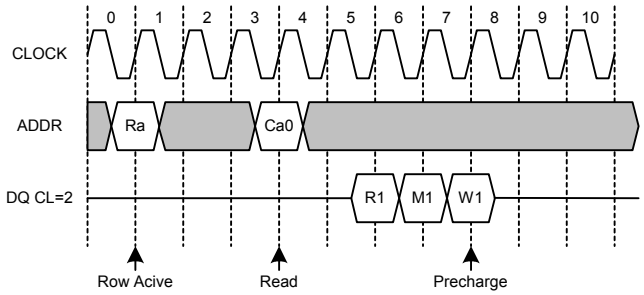


Memory Specification

	Frame-Buffer	Z-Buffer	Texture Memory
t_{RC}	20ns		
Memory Size	768kb	512kb	6Mb
Function	Read-Modify-Write (RMW) Auto Refresh NOP		Read Write Auto Refresh NOP
Control Pin	CLK - Clock Input RMW - RMW Command MASK – Write Mask Signal REF – Auto Refresh Command		CLK READ WRITE REF
Address Pin	15bit		24bit
Data pin	24bit Write Bus 24bit Read Bus	16bit Write Bus 16bit Read Bus	24bit R/W Bus

Design Restriction

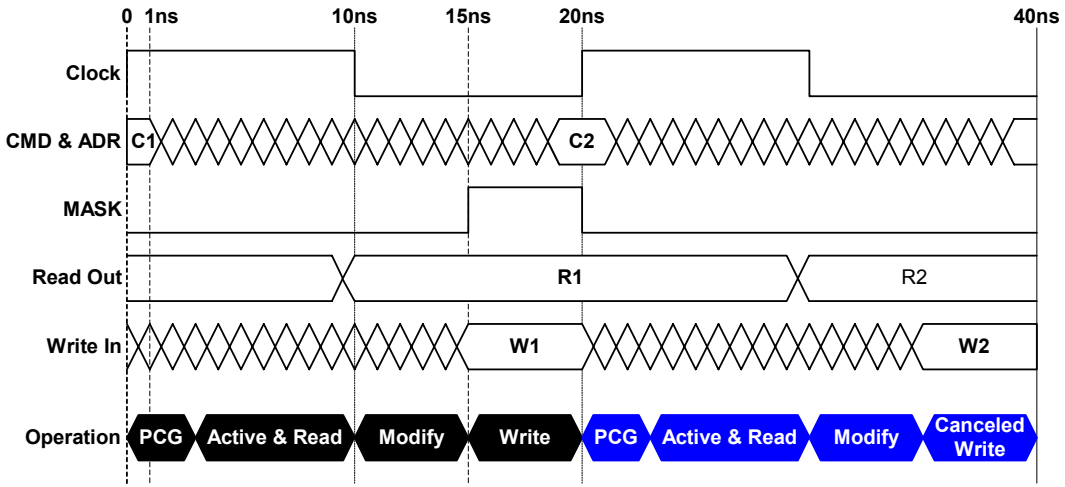
- **<u>"Always Complete RMW within 20ns"</u> for Constant Performance of the Rendering Engine**
  - t_{RC} = 20ns
    - "Precharge - Activation - CMD" within 20ns
  - Burst Access function is useless.
    - Page Access Mode enhances burst access latency.
    - Partial Activation reduces power consumption.
  - Conventional SDRAM interface scheme requires high clock frequency for 20ns t_{RC}



7 cycle for RMW = 350MHz Clock

Architecture(1/5)

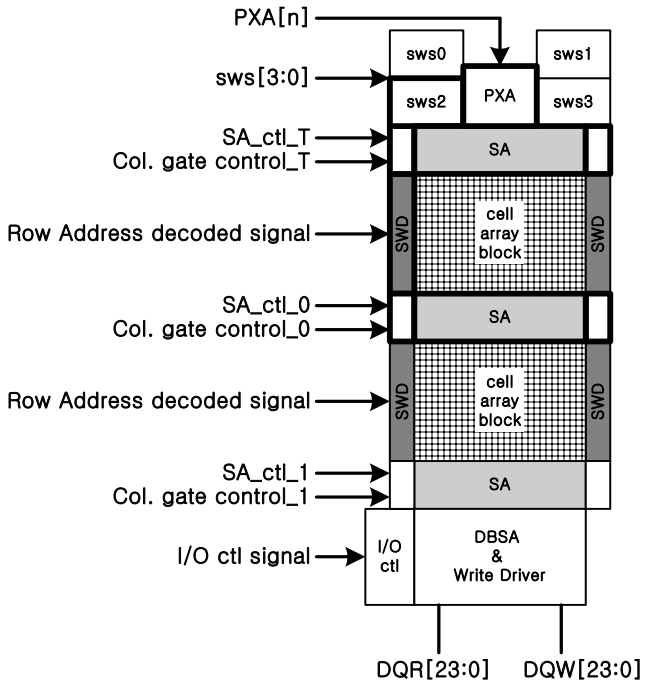
- **Single Clock Operation**
  - Finish “Precharge-Activation-CMD” in one cycle
  - SRAM-like Easy Control
  - Simplifies the pipeline control of the rendering engine
  - Low-Power Consumption due to Low Clock Frequency



FB&ZB RMW Timing Diagram

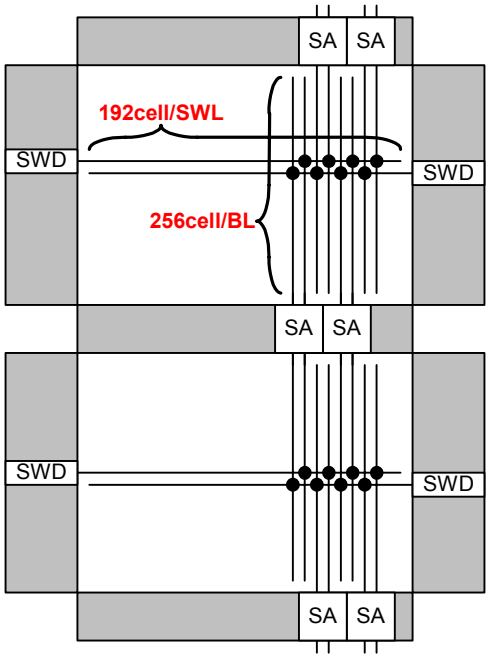
Architecture(2/5)

- **Partial Activation**
  - Consumes lower power than Page-Access-Mode

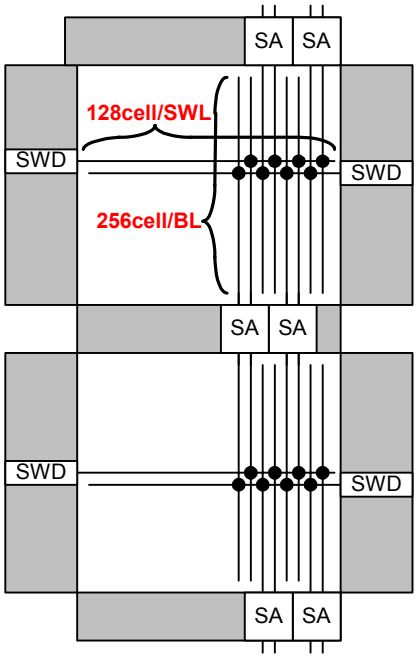


Architecture(3/5)

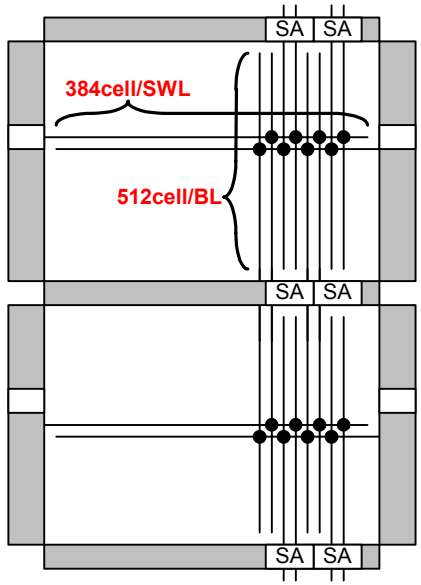
- **Speed-Optimized Cell Architecture**
  - Minimize Wire-Delay



(a) Cell Array of the Frame-Buffer



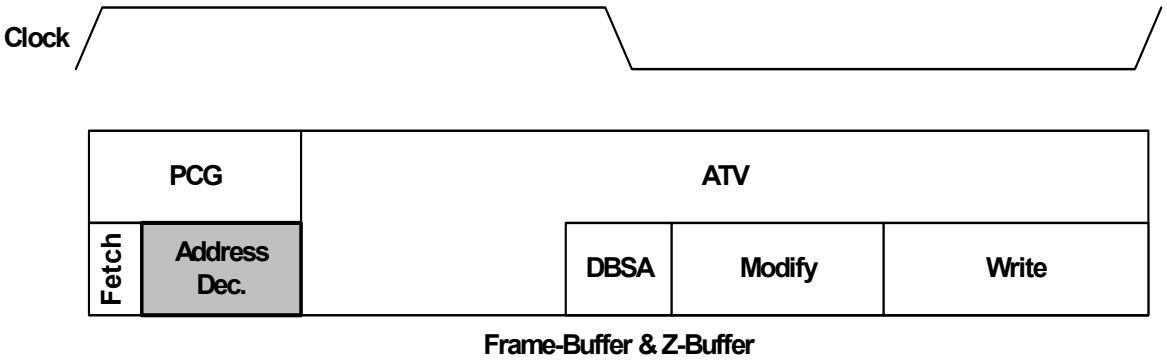
(b) Cell Array of the Z-Buffer



(c) Cell Array of the Texture Memory

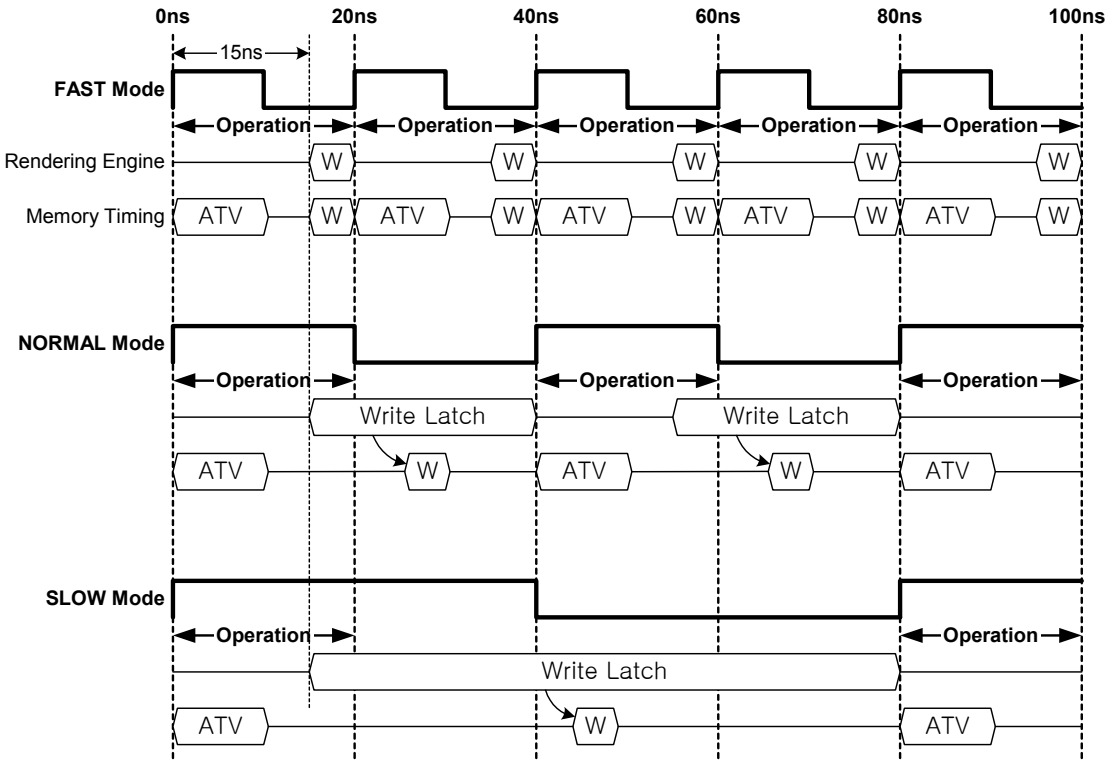
Architecture(4/5)

- **Non-Multiplexed Addressing Scheme**
  - Decodes both row and column address at a time
  - Reduces address decoding time



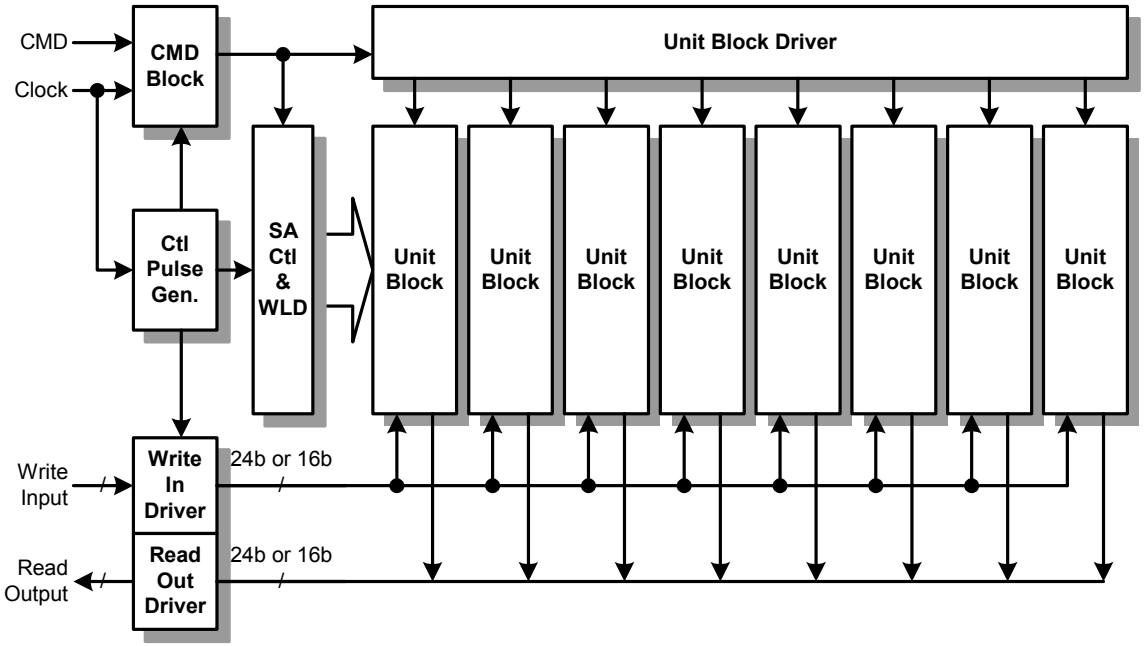
Architecture(5/5)

- **Variable Clock Operation**
  - Support Long-Lasting Operation in the System Level



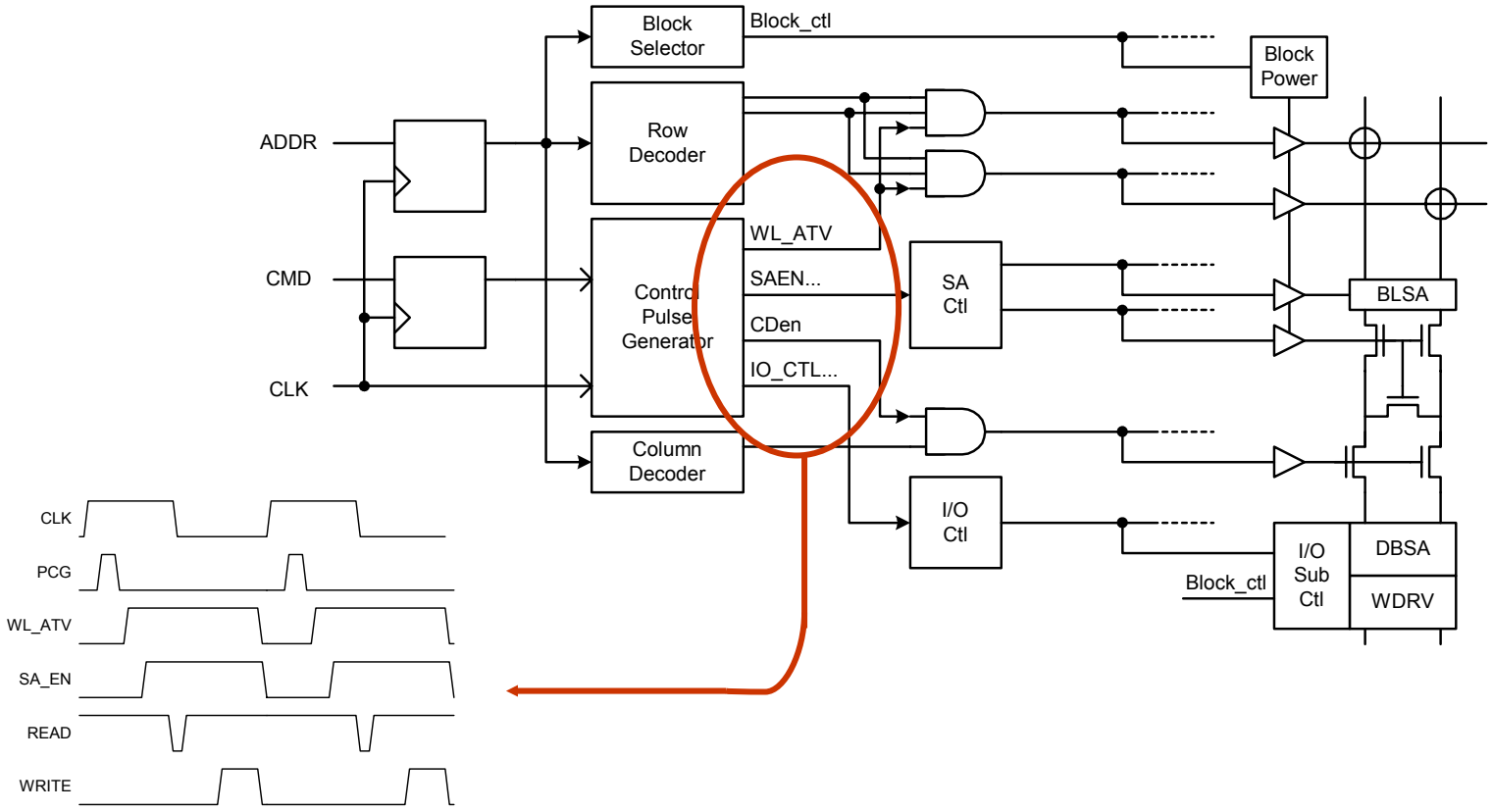
Implementation(1/2)

- **Overall Structure**



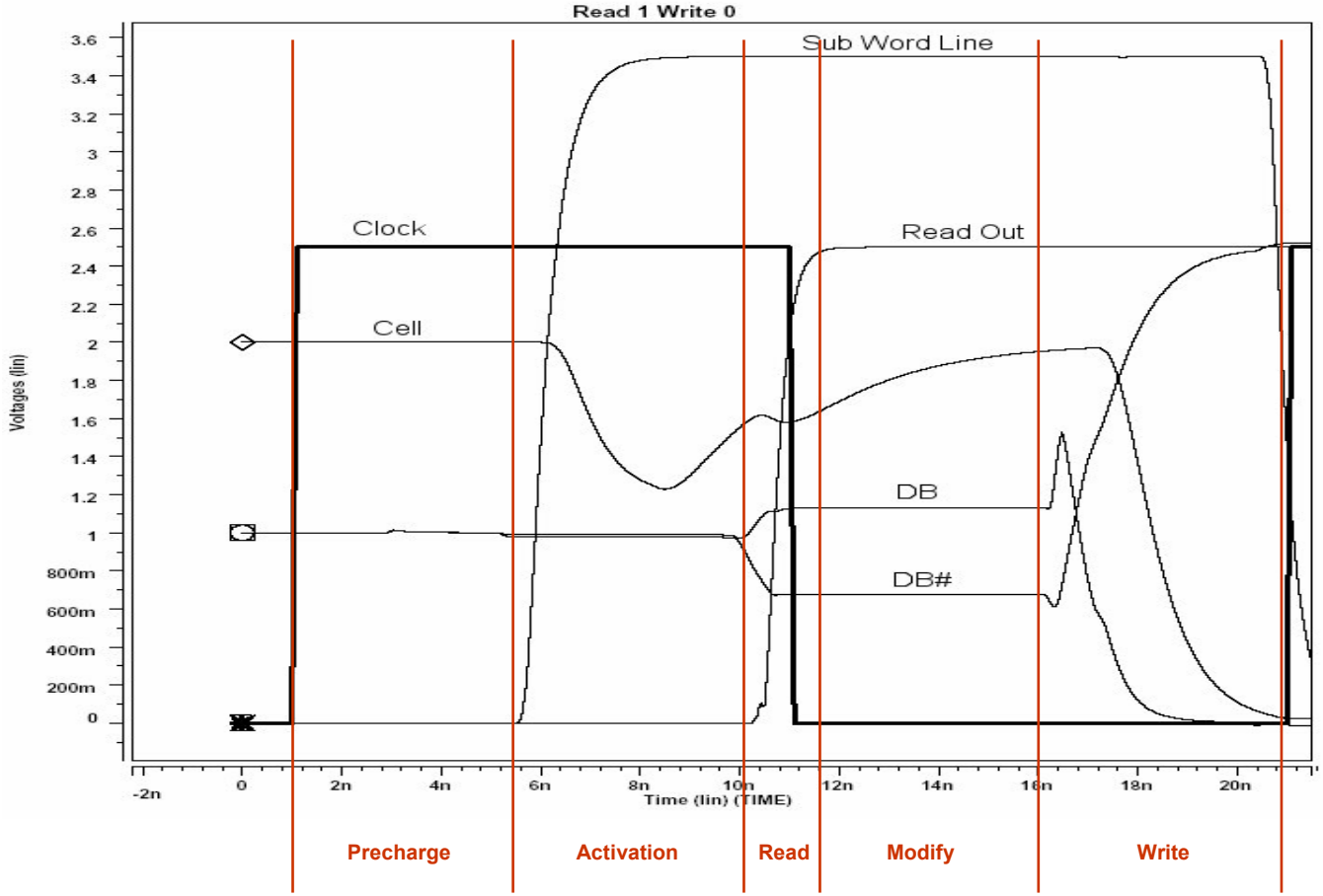
Implementation(2/2)

- **Unit Block Control**



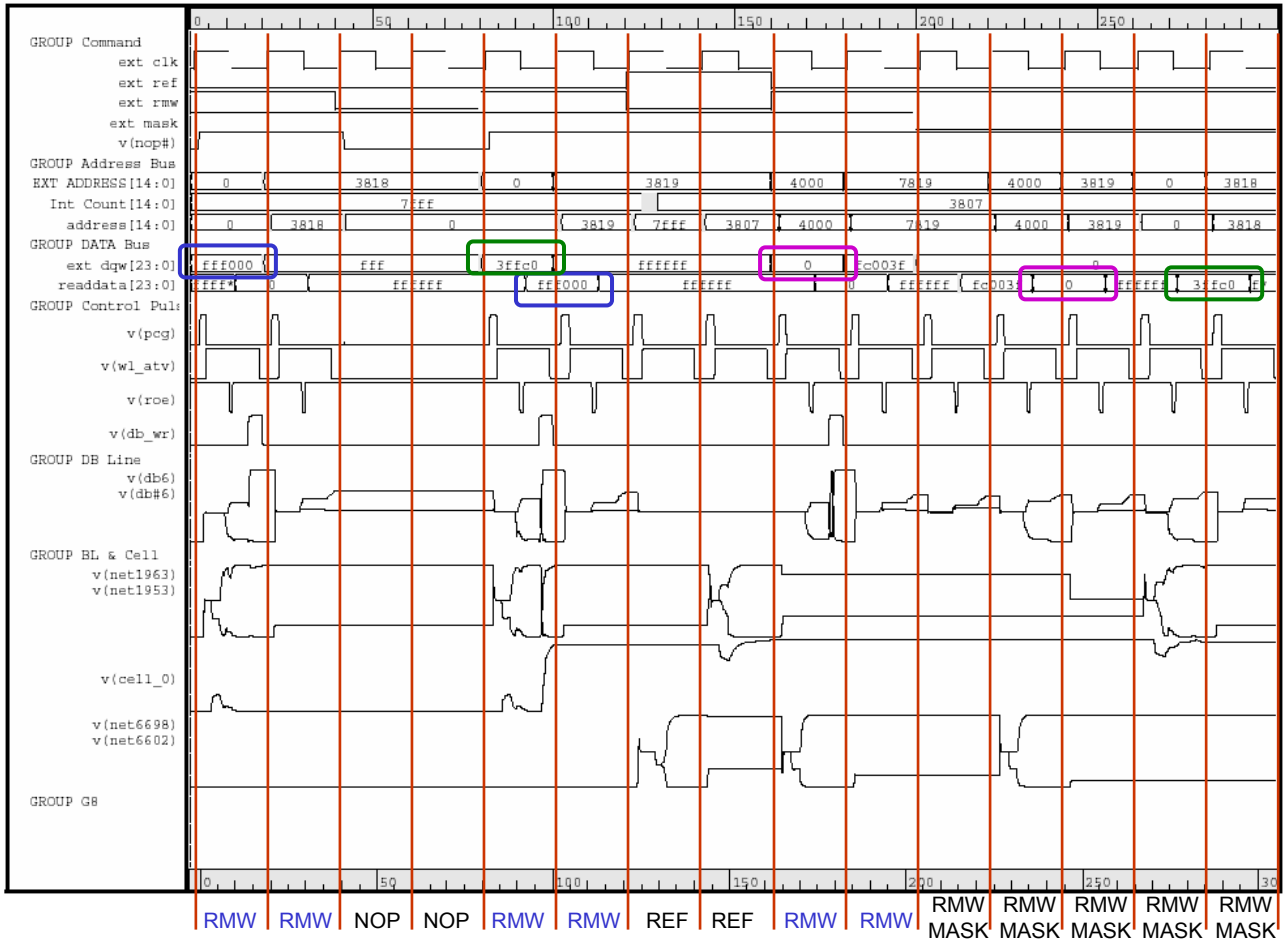
Simulation Results(1/2)

- **Read-Modify-Write Waveform**



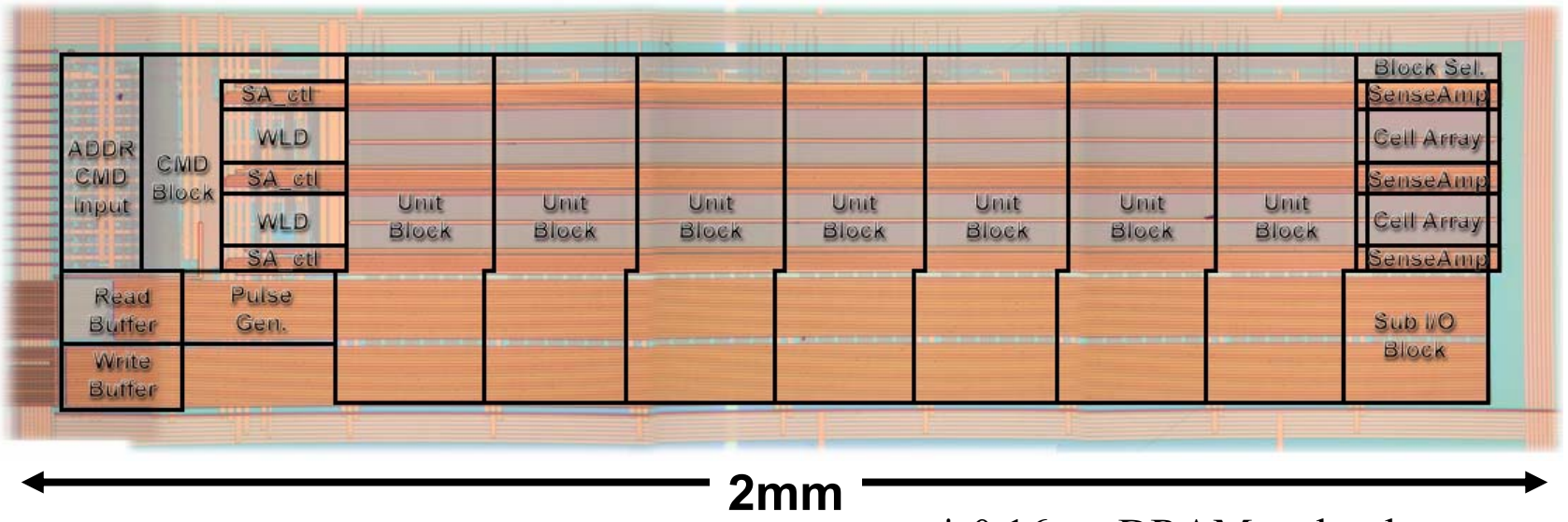
Simulation Results(2/2)

- **Operation Waveform**



Chip Micrograph(1/3)

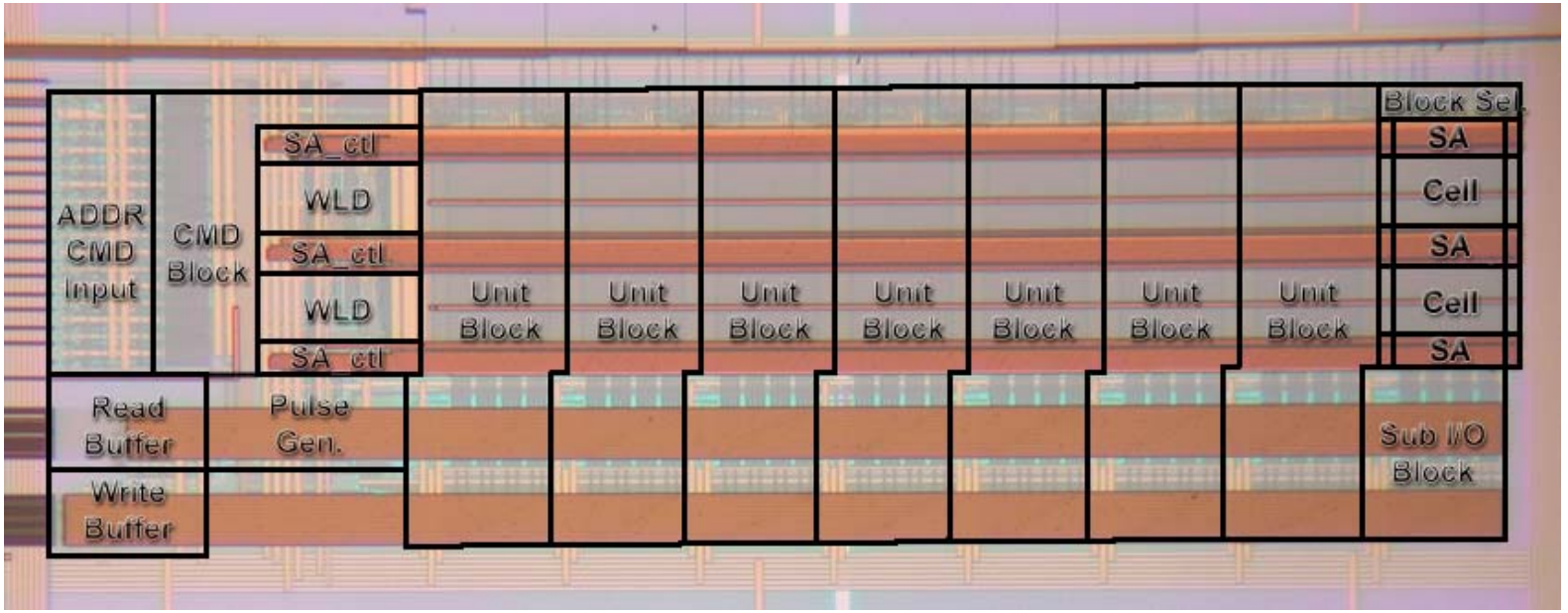
- **Frame-Buffer**



* 0.16um DRAM technology

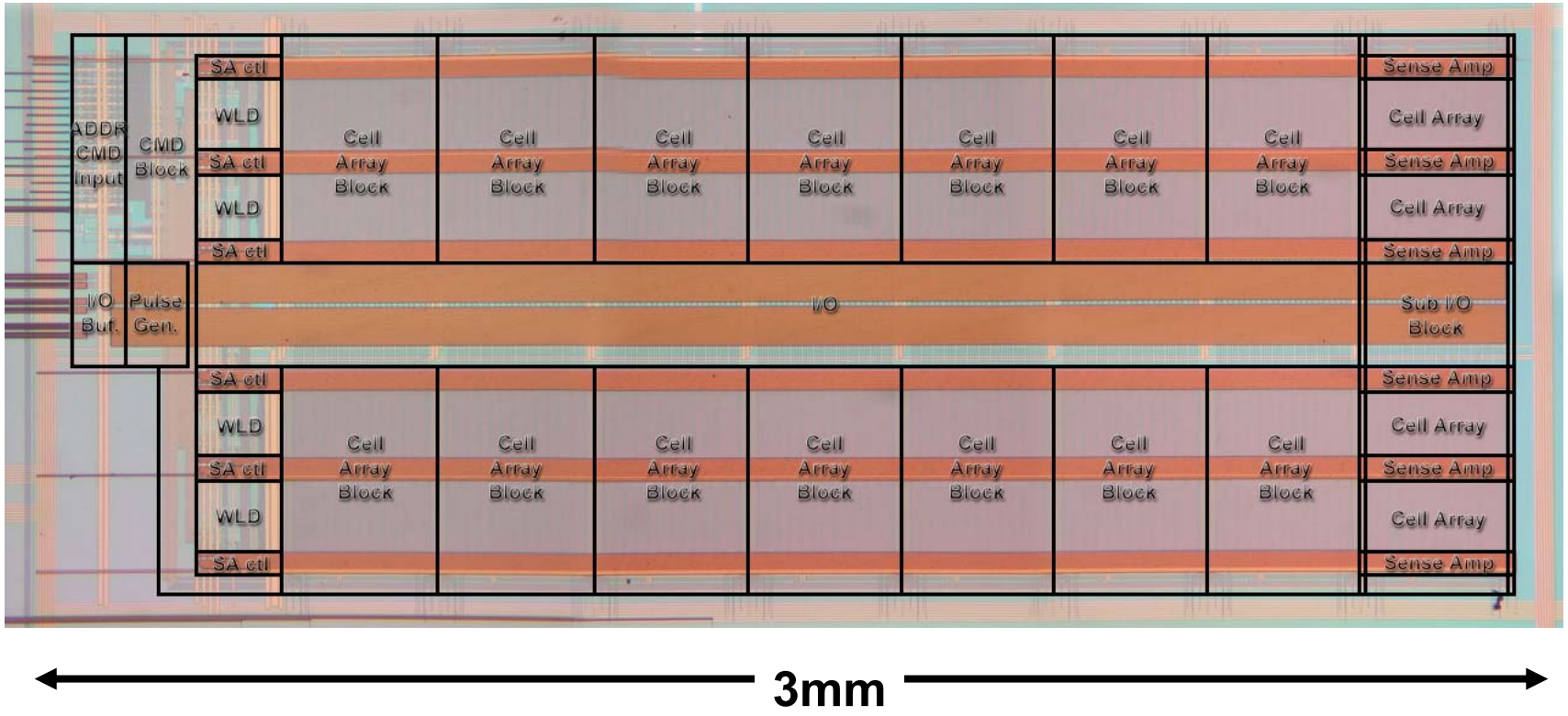
Chip Micrograph(2/3)

- **Z-Buffer**

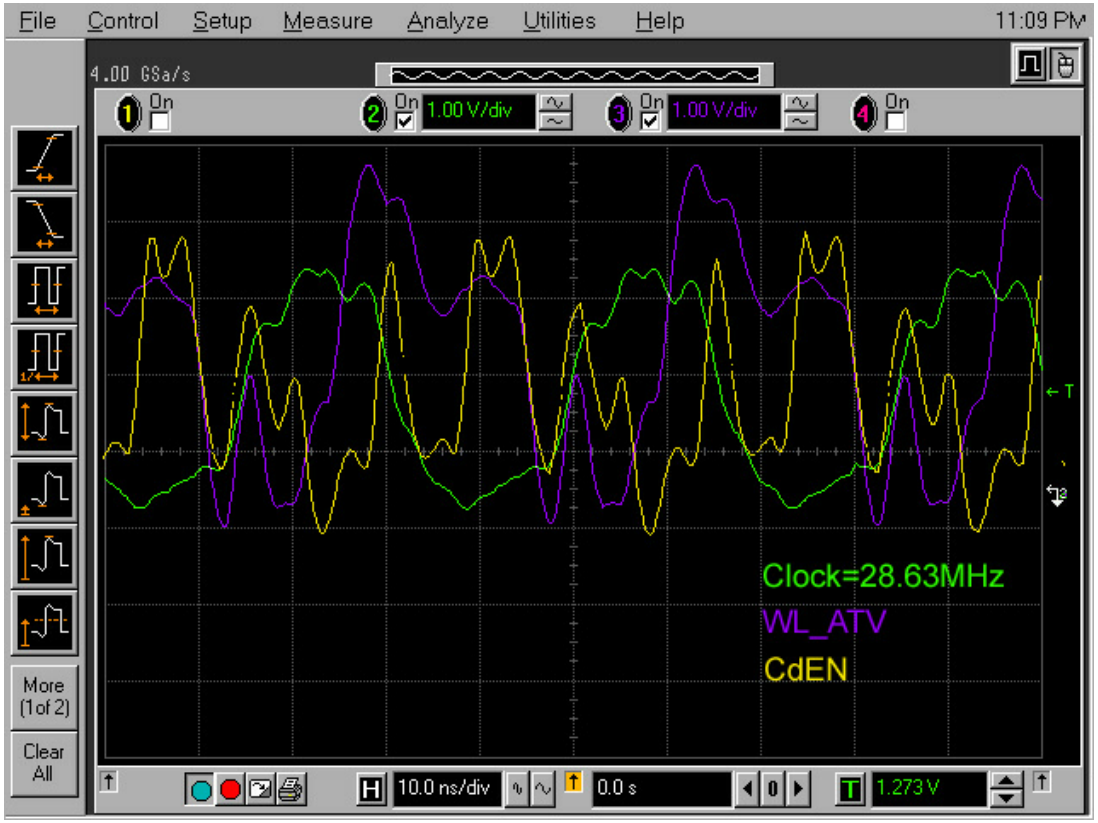


Chip Micrograph(3/3)

- **Texture Memory**



Measurement



Using Probe-Station

Performance Comparison

		Frame-Buffer	Texture Memory	Memory1*	Memory2**
Process		0.16um DRAM	0.16um DRAM	0.40um DRAM	0.25um DRAM
Memory Size		768kb	6Mb	2Mb	2.4M
Clock		50MHz	50MHz	100MHz	160MHz
Bandwidth		2.4Gb/s (RMW)	1.2Gb/s	12.8Gb/s Max.	3.8Gb/s
Latency		1	1	2:Read 1:Write + Activation Period	1
Interface		SRAM-like	SRAM-like	SDRAM-like	SRAM-like
Power	**RMW** **MASK** **Read** **Write**	**20mW** **9mW**	**18mW** **27mW**	160mW	155mW Max. 110mW Typ.
Chip Size		< 1.5mm^2/Mb (2.0mmx0.56mm)	0.6mm^2/Mb (3.0mmx1.2mm)	6.4mm^2/Mb (2.95mmx4.37mm)	2.4mm^2/Mb (2.9mmx2.0mm)
Target		Mobile 3D Graphics		PC 3D Graphics	Embedded DRAM

Low-Power & Optimized Operation for Mobile 3D Graphics

* Takao Watanabe, “A Modular Architecture for a 6.4-Gbyte/s, 8-Mb DRAM-Integrated Media Chip”, 1997

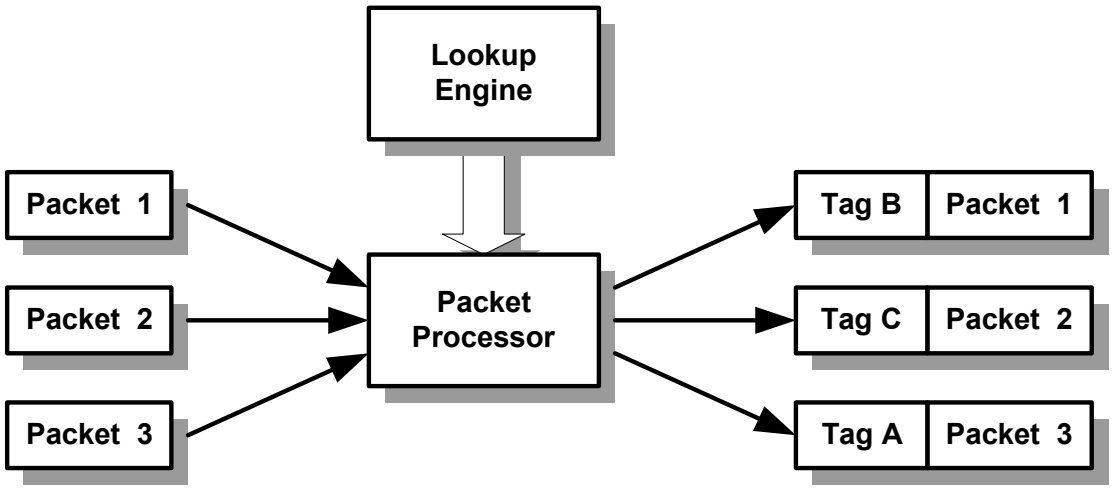
** Paul DeMone, “A 6.25ns Random Access 0.25um Embedded DRAM”, 2001

Outline

- Introduction
- Memories for Mobile 3D Graphics
- **Content-Addressable-Memory (CAM) for Network Memory**
  - Lookup Engine in a Network Processor
  - Problem of Conventional CAM
  - Proposed CAM for Lookup Engine
  - Simulation Results
- Conclusion

Lookup Engine in a Network Processor

- **Lookup Engine**
  - Searches Tag for incoming Packet
  - Has Lookup Memory (Rule Memory)
  - "Search" is main job



Implementing Lookup Engine

- **Conventional Memory Architecture***
  - Requires Frequent Memory Access
  - Consumes Many Clock Cycle ➔ Low Performance

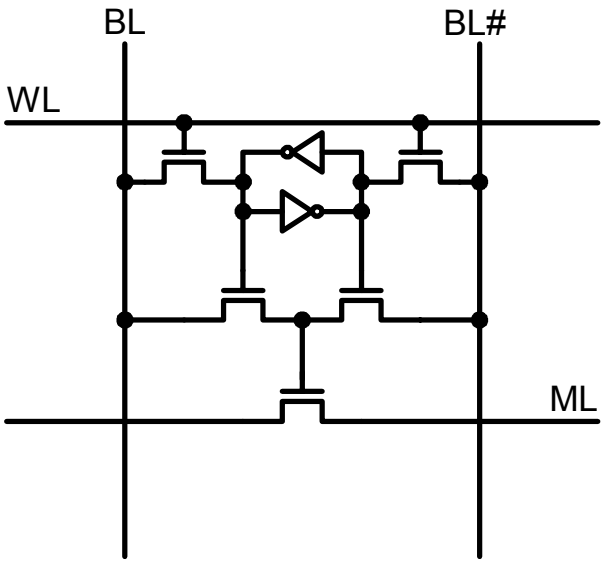
- **CAM Architecture**
  - Enables One-Cycle Search Operation
  - Large Power & Area Consumption

CAM is preferred.

* Miguel A. Ruiz-Sanchez, “Survey and Taxonomy of IP Address Lookup Algorithms”, 2001

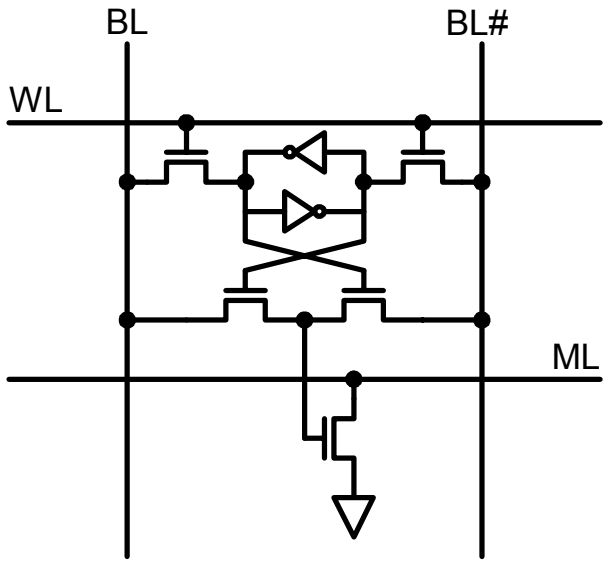
Conventional CAM

- **Structure of Conventional CAM***



Conventional NAND type

☺ **Low Power Consumption**
☹ **Slow Match Line Propagation**
☹ **Cannot handle 'Don't Care'**



Conventional NOR type

☺ **Fast Match Line Propagation**
☹ **Large Power Consumption**
☹ **Cannot handle 'Don't Care'**

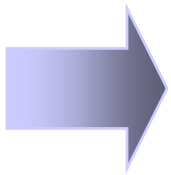
* Ilion Yi-Liang Hsiao, et. al., "Power Modeling and Low-Power Design of Content Addressable Memories", ISCAS 2001

Conventional TCAM(1/3)

- **Need for 'Don't Care'**

Prefix	Next hop
143.248.1.1	L2
143.248.1.2	L2
…	L2
143.248.255.255	L2
203.238.128.56	L5
203.238.128.1	L6
203.238.128.2	L6
…	L6
203.238.128.255	L6

Routing Table without Don't Care



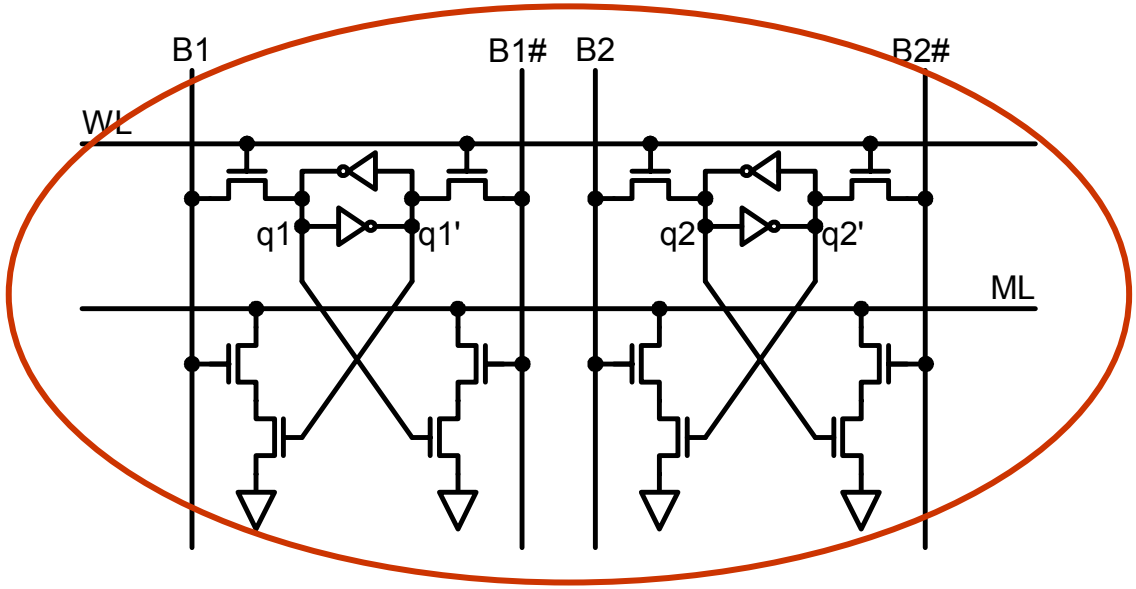
Prefix	Next hop
*	L9
143.248.*.*	L2
203.238.128.56	L5
203.238.128.*	L6
203.*.*.*	L1

Routing Table using Don't Care

'Don't Care' reduces the size of routing table.

Conventional TCAM(2/3)

- **Handling Don't Care**



	q1	q2
0	0	1
1	1	0
*	0	0

2 BCAMs build 1 TCAM

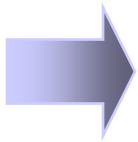
Encode Don't Care using Two CAM Cells*

* Sergio R. Ramirez-Chavez, "Encoding Don't Cares in Static and Dynamic CAMs", 1992

Conventional TCAM(3/3)

- **Dynamic TCAM***
  - Use Gate Cap. as a storage
- **DRAM CAM****
  - 2 DRAM cells for one TCAM
- **Multiple-Valued CAM*****
  - EEPROM technology

Refresh Operation
Core speed
Technology
…



Bottleneck

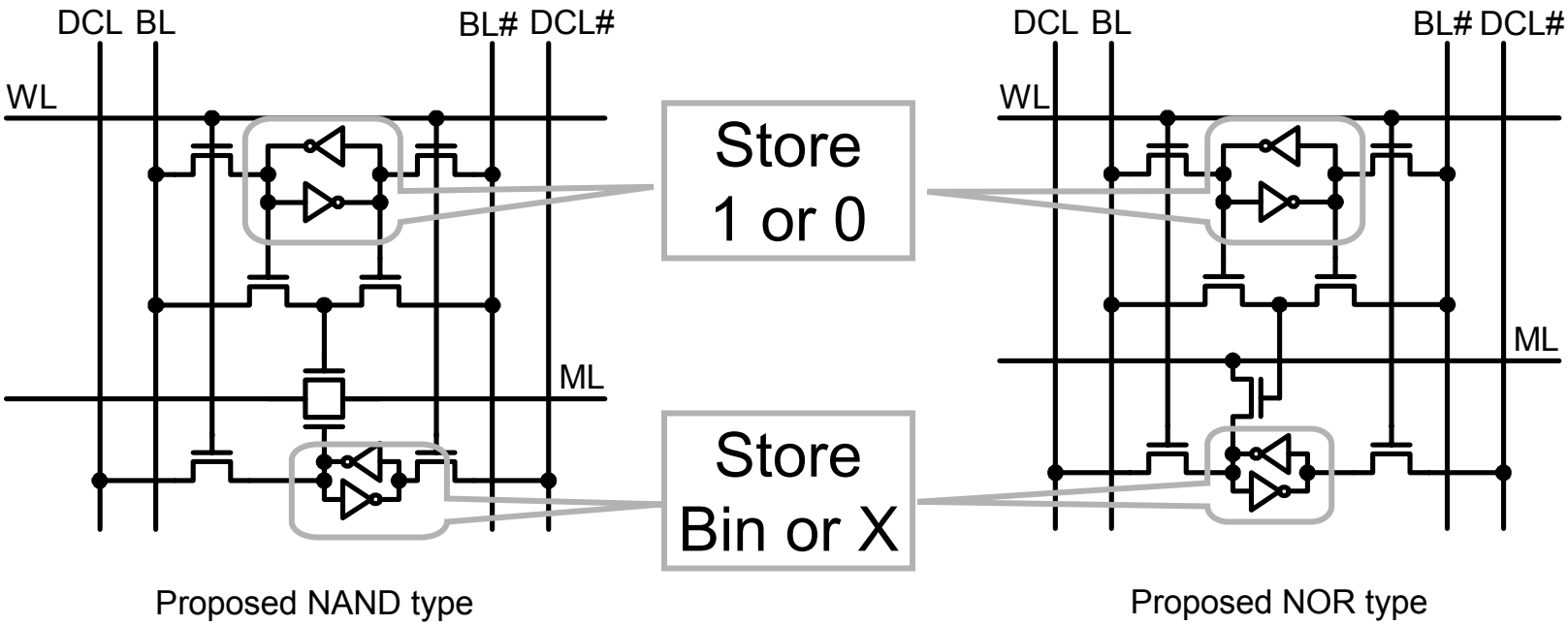
* JON P. WADE, “A Ternary Content Addressable Search Engine”, 1989

** Tadato Yamagata, “A 288kb Fully Parallel Content Addressable Memory Using a Stacked-Capacitor Cell Structure”, 1992

*** Takahiro Hanyu, “Design of a One-Transistor-Cell Multiple-Valued CAM”, 1996

Proposed CAM(1/3)

- **Ternary CAM Cell Structure**

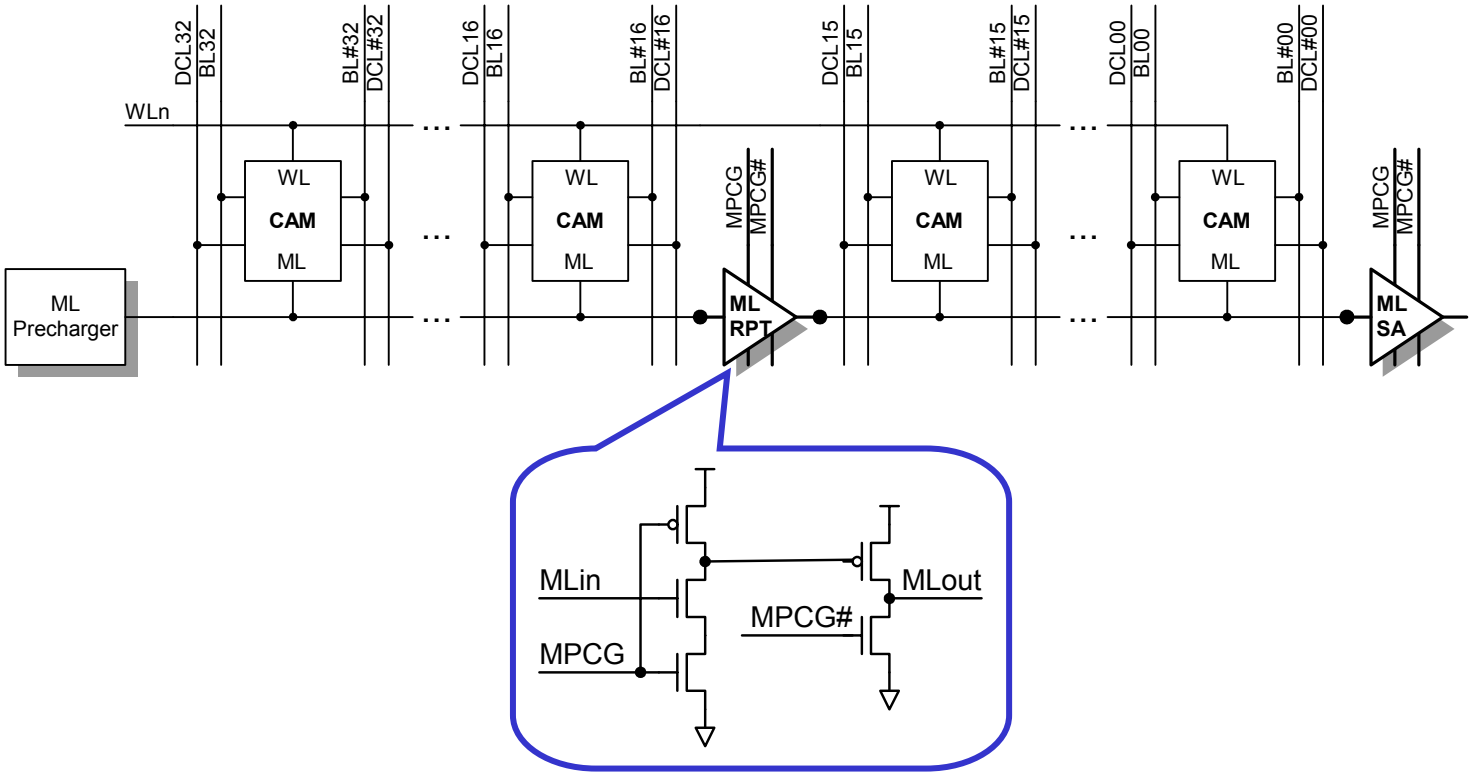


NAND type consumes less power.

Patent Pending 2002/11

Proposed CAM(2/3)

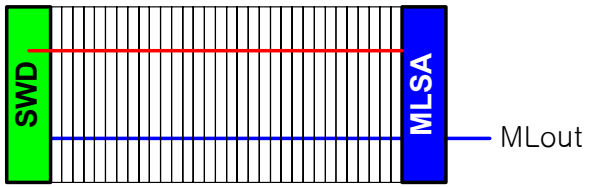
- **Match Line Repeater**

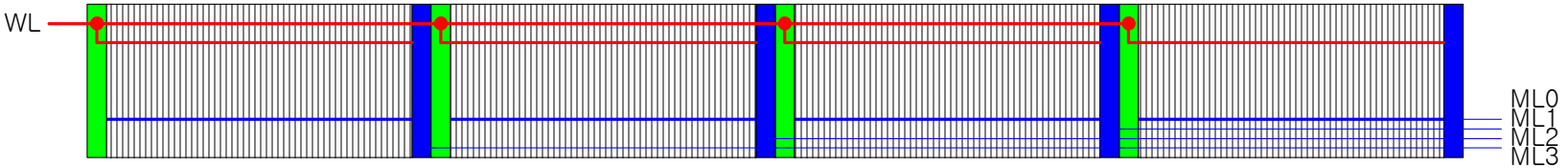


Enhance Match Line speed of NAND type Cell Structure

Proposed CAM(3/3)

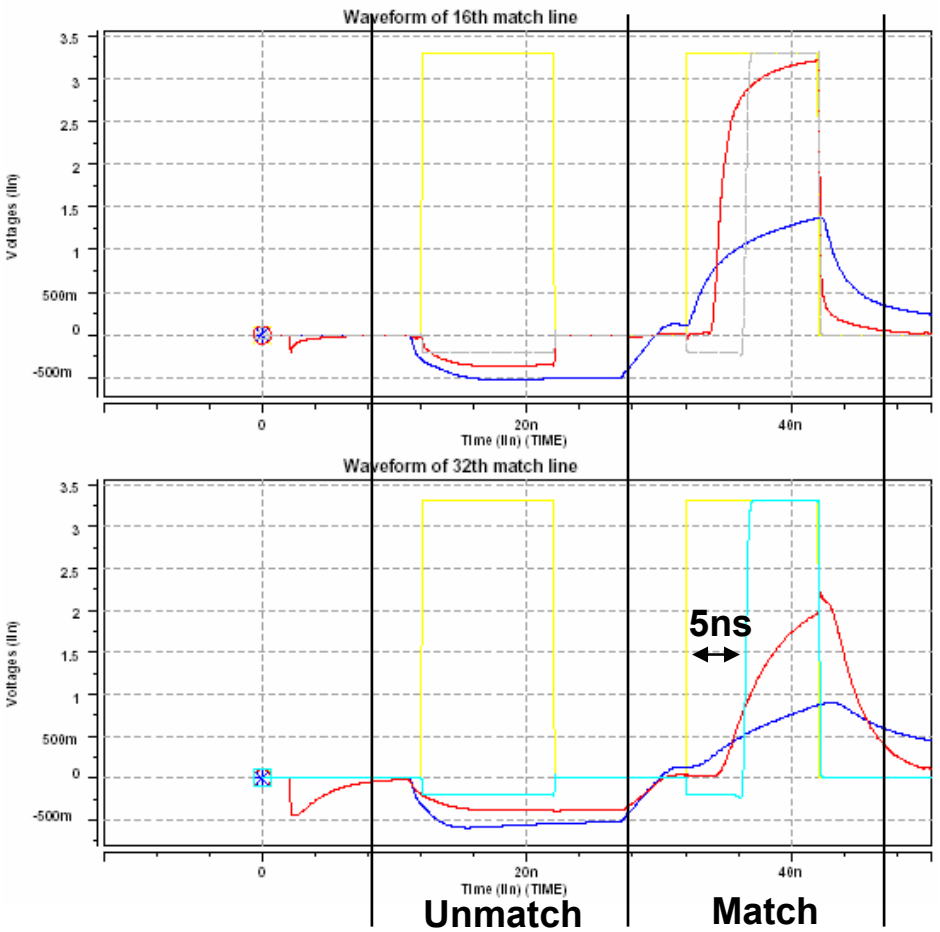
- **2-D Decoding Method**
  - By Long Aspect Ratio of the Proposed Cell





Simulation Results

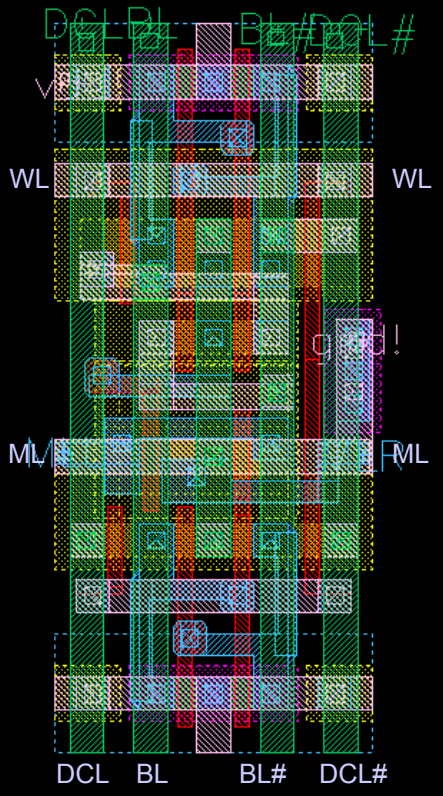
- **Match Line Propagation**



Proposed Scheme
Conventional Scheme
Final Match Out

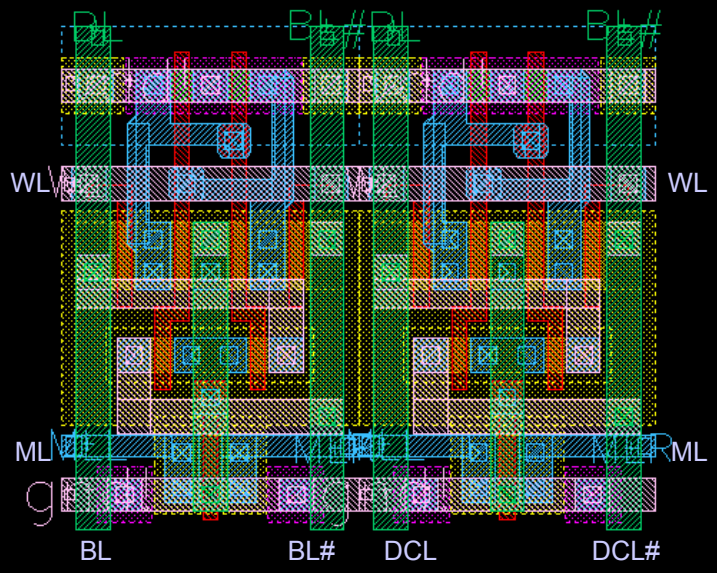
More than x3 Faster than Conventional Scheme

Cell Size



Proposed Cell

25% smaller than conventional TCAM



Conventional Cell

Expected Performance

	Proposed	JSSC 1998*	JSSC 2001**
Process	0.35um CMOS	0.35um 5M1P	0.35um Standard
Clock	100MHz	30MHz	70MHz
Memory Capacity	64k x 32bit **Ternary**	64k x 40bit **Binary**	256 x 54bit **Binary**
Search Power Dissipation	100~120fJ/Bit/Search	83fJ/Bit/Search	131fJ/Bit/Search
Search Speed	5ns for Search Evaluation	26ns for Search Out	7.3ns for Search Evaluation
Match Line Type	NAND	NAND	NOR
Etc	Store Don't Care	Separated Bit-Line and Search-Line	pMOS NOR type

* Farhad Shafai, et al, "Fully Parallel 30-MHz, 2.5-Mb CAM", JSSC November 1998

** Hisatada Miyatake, et al, "A Design for High-Speed Low-Power CMOS Fully Parallel Content-Addressable Memory Macros", JSSC June 2001

Outline

- Introduction
- Memories for Mobile 3D Graphics
- Content-Addressable-Memory (CAM) for Network Memory
- **Conclusion**

Conclusion

- **Application Specific Memory Architectures are Proposed.**
  - Frame-Buffer, Z-Buffer and Texture Memory for Mobile 3D Rendering Engine
  - Ternary-CAM Structure for Network Lookup Engine
- **Mobile 3D Graphic Memories are Implemented.**
  - 0.16um DRAM Technology.
  - 20ns t_{RC} with Read-Modify-Write Operation.
  - Improve the performance of the 3D Graphics System.
- **New Ternary-CAM Structure is Proposed.**
  - 5ns Match Evaluation Time
  - 25% Reduced Cell Size