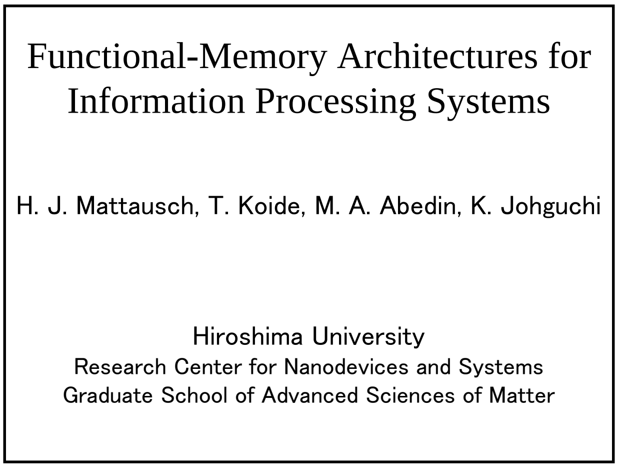
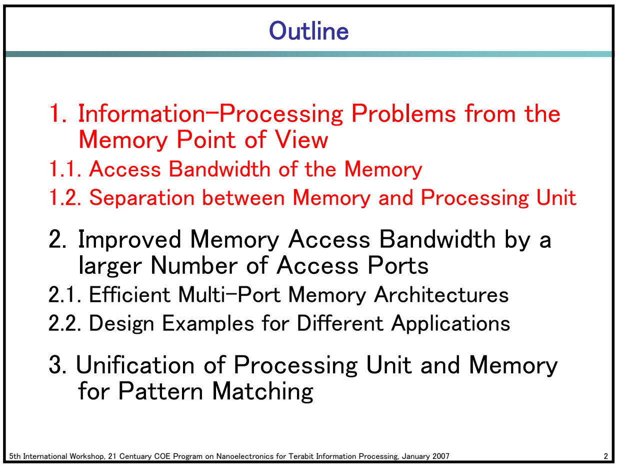


Functional-Memory Architectures for Information Processing Systems

H. J. Mattausch, T. Koide, M. A. Abedin, K. Johguchi

Hiroshima University
Research Center for Nanodevices and Systems
Graduate School of Advanced Sciences of Matter



Outline

1. Information-Processing Problems from the Memory Point of View

1.1. Access Bandwidth of the Memory

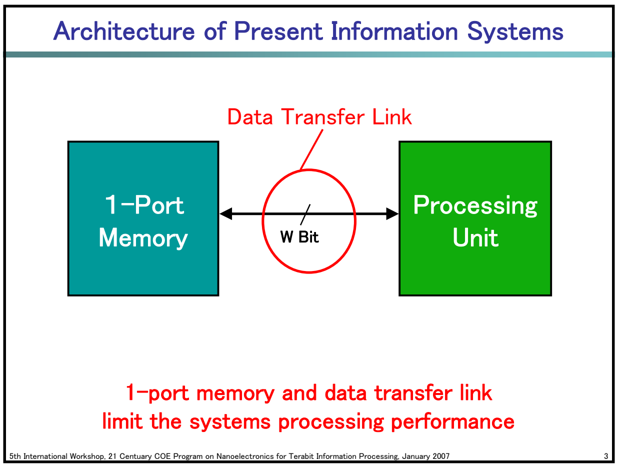
1.2. Separation between Memory and Processing Unit

2. Improved Memory Access Bandwidth by a larger Number of Access Ports

2.1. Efficient Multi-Port Memory Architectures

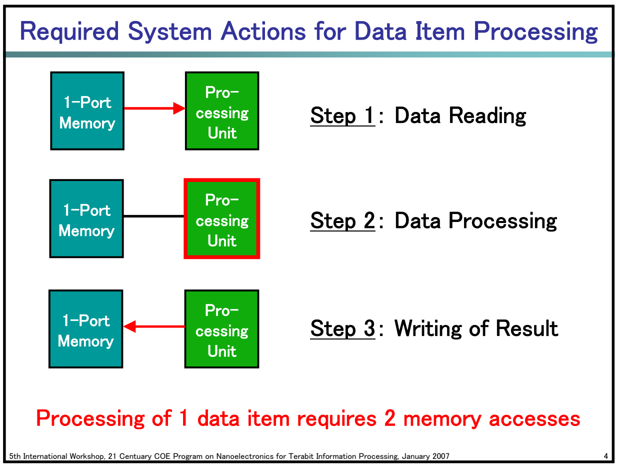
2.2. Design Examples for Different Applications

3. Unification of Processing Unit and Memory for Pattern Matching



Architecture of Present Information Systems

1-port memory and data transfer link limit the systems processing performance



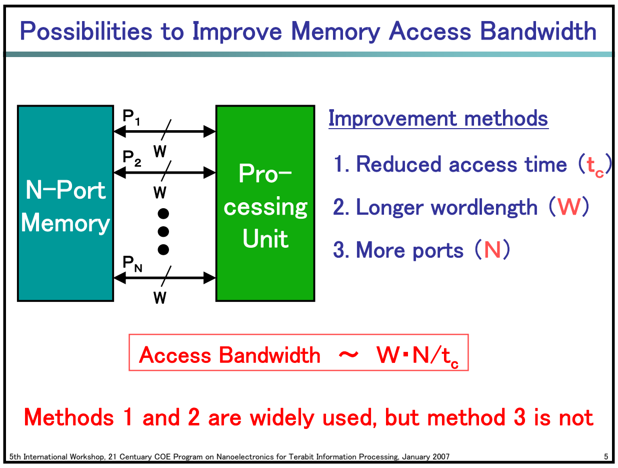
Required System Actions for Data Item Processing

Step 1: Data Reading

Step 2: Data Processing

Step 3: Writing of Result

Processing of 1 data item requires 2 memory accesses



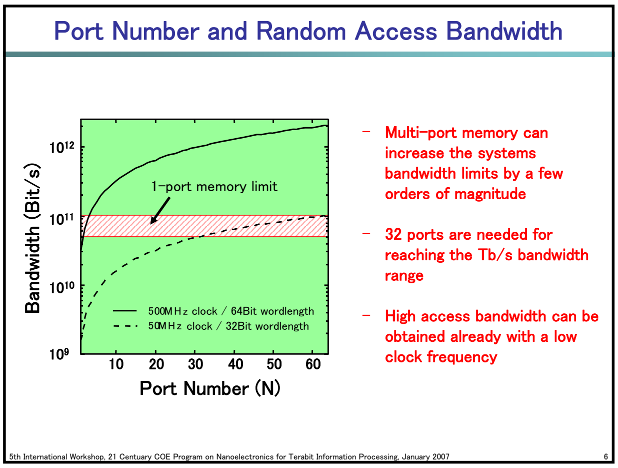
Possibilities to Improve Memory Access Bandwidth

Improvement methods

1. Reduced access time (t_c)
2. Longer wordlength (W)
3. More ports (N)

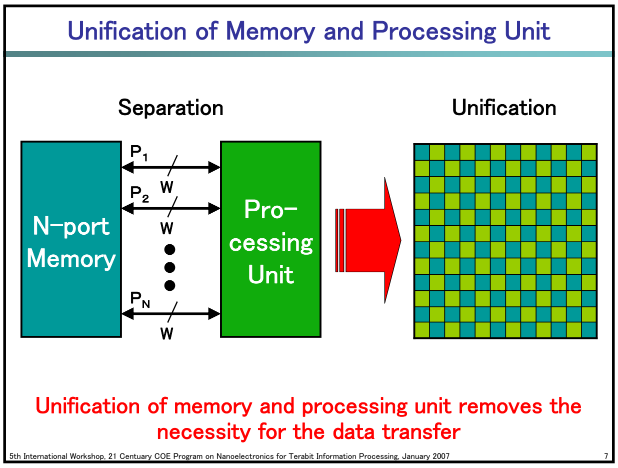
Access Bandwidth ～ $W \cdot N/t_c$

Methods 1 and 2 are widely used, but method 3 is not



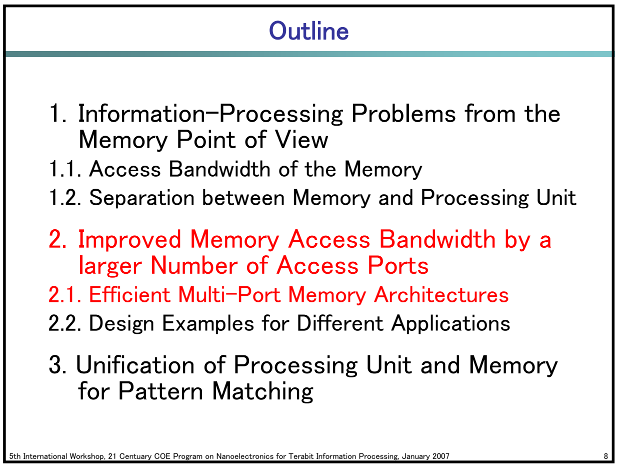
Port Number and Random Access Bandwidth

- Multi-port memory can increase the systems bandwidth limits by a few orders of magnitude
- 32 ports are needed for reaching the Tb/s bandwidth range
- High access bandwidth can be obtained already with a low clock frequency



Unification of Memory and Processing Unit

Unification of memory and processing unit removes the necessity for the data transfer

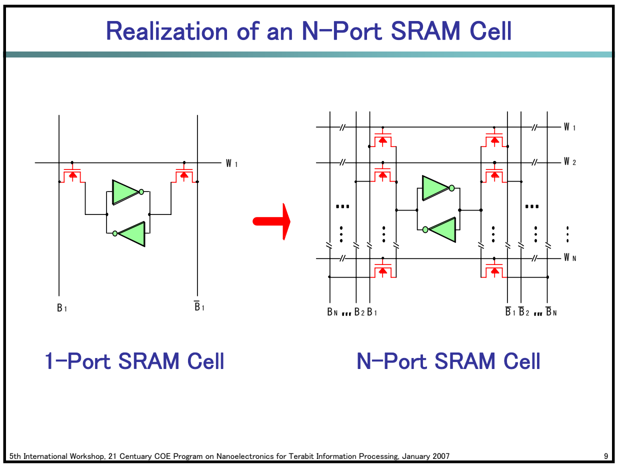


Outline

1. Information-Processing Problems from the Memory Point of View
1.1. Access Bandwidth of the Memory
1.2. Separation between Memory and Processing Unit

2. Improved Memory Access Bandwidth by a larger Number of Access Ports
2.1. Efficient Multi-Port Memory Architectures
2.2. Design Examples for Different Applications

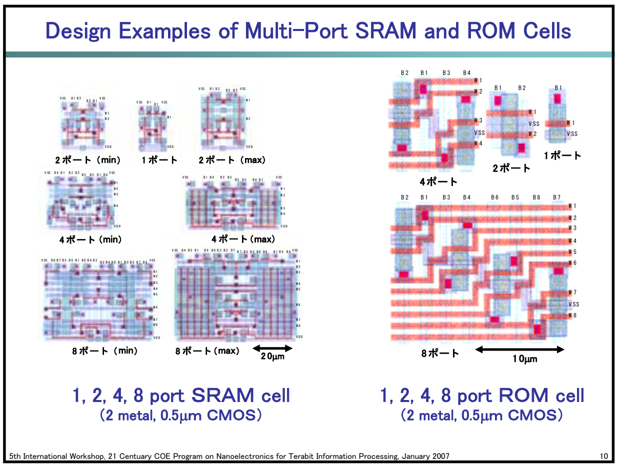
3. Unification of Processing Unit and Memory for Pattern Matching



Realization of an N-Port SRAM Cell

1-Port SRAM Cell

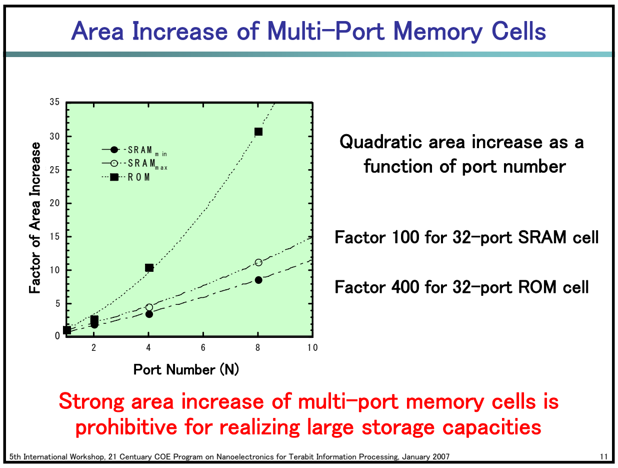
N-Port SRAM Cell



Design Examples of Multi-Port SRAM and ROM Cells

1, 2, 4, 8 port SRAM cell
(2 metal, 0.5μm CMOS)

1, 2, 4, 8 port ROM cell
(2 metal, 0.5μm CMOS)



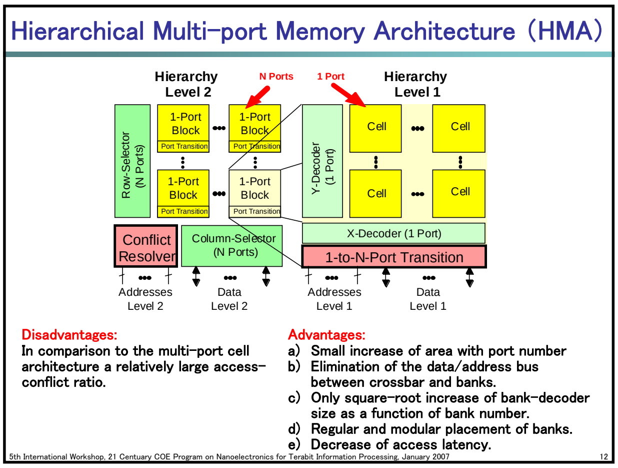
Area Increase of Multi-Port Memory Cells

Quadratic area increase as a function of port number

Factor 100 for 32-port SRAM cell

Factor 400 for 32-port ROM cell

Strong area increase of multi-port memory cells is prohibitive for realizing large storage capacities

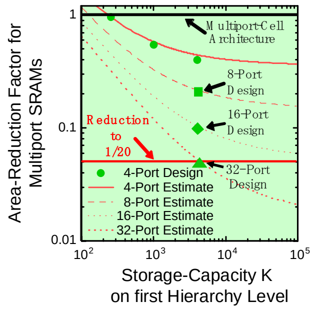


Hierarchical Multi-port Memory Architecture (HMA)

Disadvantages:
In comparison to the multi-port cell architecture a relatively large access-conflict ratio.

Advantages:
a) Small increase of area with port number
b) Elimination of the data/address bus between crossbar and banks.
c) Only square-root increase of bank-decoder size as a function of bank number.
d) Regular and modular placement of banks.
e) Decrease of access latency.

HMA Comparison with Multi–Port Cell Architecture

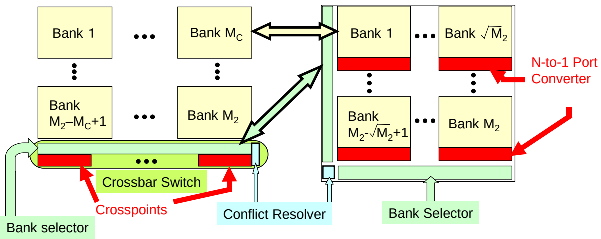


Conclusion

HMA can reduce the area consumption of the multi–port cell architecture by an order of magnitude.

Structural Comparison of HMA and Crossbar

	Crossbar Memory	HMA Memory
N-to-1 port Converter	Located in the Crossbar	Located in the Bank
Placement of Banks	Restricted by Crossbar Size	Free Placement
Transistors of Bank Selector	Proportional to $M_2 \log_2 M_2$	Proportional to $\sqrt{M_2} \log_2 M_2$
Global Signals	Data／Address Lines divided at Crossbar	Data／Address Lines undivided up to Banks



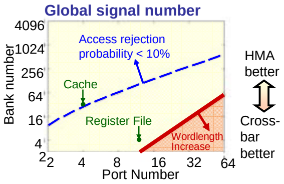
Quantitative Comparison with Crossbar

S. Fukae, T. Inoue, H.J. Mattausch, T. Koide, and T. Hironaka, "Distributed against centralized crossbar function for realizing bank–based multiport memories", IEE Electronics Letters 40, 101–103 (2004)

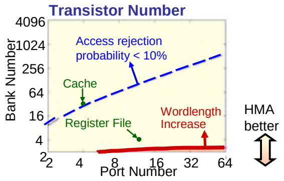
- Comparison of global signal number
  - – HMA removes global signals from crossbar's crosspoints to the banks
  - – Bank-related data/address signals are direct bank inputs in HMA
  - → HMA has less global signals for larger port and bank numbers.
- Comparison of transistor number
  - – Transistor number of bank selectors is smaller for HMA.
  - – HMA does not require signal buffers between crosspoints and banks.
  - → HMA has smaller transistor numbers in nearly all cases

In real-world applications HMA is superior to the centralized crossbar architecture.

Global signal number



Transistor Number



Outline

1. Information–Processing Problems from the Memory Point of View

1.1. Access Bandwidth of the Memory

1.2. Separation between Memory and Processing Unit

2. Improved Memory Access Bandwidth by a larger Number of Access Ports

2.1. Efficient Multi–Port Memory Architectures

2.2. Design Examples for Different Applications

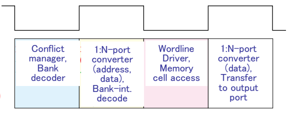
3. Unification of Processing Unit and Memory for Pattern Matching

16–Port HMA SRAM Features

- Introduction of Pipeline Stages

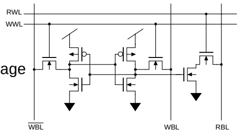
  Access-path division, Overlap of access requests leads to higher clock frequencies

  Increase of throughput (Access Bandwidth)



Conflict manager, Bank decoder
1:N–port converter (address, data), Bank–int. decode
Wordline Driver, Memory cell access
1:N–port converter (data), Transfer to output port

- 2-port SRAM Cell with separate Read/Write ports
  - Read-access is faster than 6T-SRAM cell.
  - Separate read and write access path up to storage cell facilitate pipe-lining.
  - Read access introduces no static noise at storage node, so that read access is fast and reliable.
  - Large dynamic margin, so that low voltage operation is possible.

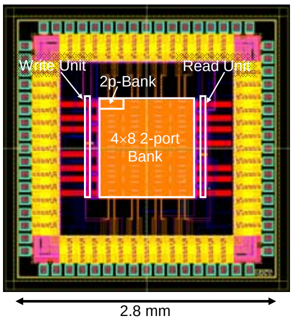


- Large Signal Sensing Technology

  Small node capacities are used to facilitate large signal read sensing.
  → Read-path becomes very digital like (High speed and large noise immunity)

- High Speed or Low-Power Mode selectable by Supply Voltage

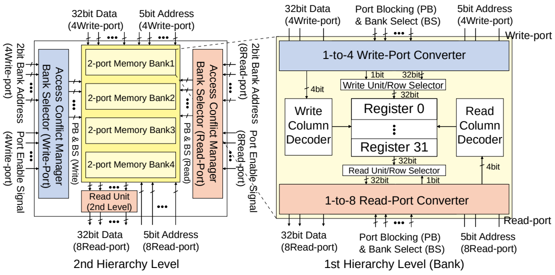
90nm CMOS, 16–port HMA Memory Design



Technology	90nm CMOS
Routing Layers	6 metal (Cu)
Supply Voltage	1.0 V
SRAM Cell	2-Port SRAM Cell (1 Read, 2 Write)
Area Consumption	0.97 mm^2
Port Number	8-Read, 8-Write
Bank Number	32
Bank Capacity	2 Kbit
SRAM Capacity	64 Kbit
Wordlength	32 bit
Clock Cycle Time	0.86 nsec
Power Dissipation	41mW at 1GHz
Access Bandwidth	590 Gbps

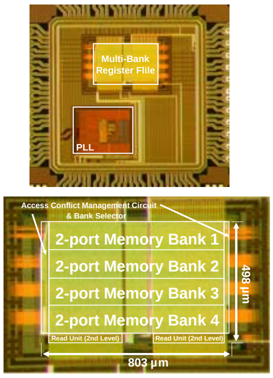
A wordlength of 64 bit or a port–number increase to 32 ports would result in 1.18 Tbps access bandwidth.

12-Port Multi-Bank Register-File Architecture


32bit Data (4Write-port)
5bit Address (4Write-port)
2bit Bank Address (4Write-port)
Port Enable Signal (4Write-port)
Access Conflict Manager Bank Selector (Write-Port)
2-port Memory Bank1
2-port Memory Bank2
2-port Memory Bank3
2-port Memory Bank4
PB & BS (Write)
PB & BS (Read)
Access Conflict Manager Bank Selector (Read-Port)
2bit Bank Address (8Read-port)
Port Enable Signal (8Read-port)
Read Unit (2nd Level)
32bit Data (8Read-port)
5bit Address (8Read-port)
2nd Hierarchy Level
32bit Data (4Write-port)
Port Blocking (PB) & Bank Select (BS)
5bit Address (4Write-port)
Write-port
1-to-4 Write-Port Converter
1bit
32bit
4bit
Write Unit/Row Selector
32bit
Write Column Decoder
Register 0
Register 31
Read Column Decoder
32bit
Read Unit/Row Selector
32bit
1bit
4bit
1-to-8 Read-Port Converter
Read-port
32bit Data (8Read-port)
Port Blocking (PB) & Bank Select (BS)
5bit Address (8Read-port)
1st Hierarchy Level (Bank)

- 2 port SRAM cells and 1-to-8 read, 1-to-4 write port converters in each bank.
  - Reduction of area and access delay time.
- Access conflicts are avoided by,
  - Access scheduler integrated after instruction decoding in the processor.
  - Completely divided read and write path until SRAM cells.

12-Port Register File Design in 180nm CMOS



Technology	0.18µm CMOS
Gate Length	200 nm
Interconnection	5 Layer Al Metal
Power Supply	1.8V
Die Size	2.8mm × 2.8mm
Port Number	12 port (Read 8, Write 4)
Registers	128
Capacity	4 Kbit
Bank Capacity	1 Kbit
Bank Number	4 bank
Wordlength	32 bit
Max. Clock	640 MHz (simulated)
Area	0.39 mm^2

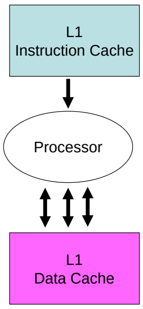
Bank Decoder is placed in the center of each bank.

Comparison with Multi-Port Cell Register Files

	Multi-bank Register File (HMA)	Conventional Multi-port-cell Register File	Multi-bank Register File (HMA), **estimated**	Conventional Multi-port-cell Register File ISSCC2002
Technology	200nm L_{gate} 5 metal CMOS	200nm L_{gate} 5 metal CMOS	110nm L_{gate} 5 metal CMOS	110nm L_{gate} 4 metal CMOS
Supply voltage	1.8 V	1.8 V	1.2 V	1.2 V
Access ports	12 (8r, 4w)	12 (8r, 4w)	16 (10r, 6w)	16 (10r, 6w)
Registers	128	128	34	34
Word length	32 bit	32 bit	64 bit	64 bit
Core area	0.39 mm^2	1.43 mm^2	0.21 mm^2	0.5 mm^2
Max operation frequency	640 MHz (simulated) 417 MHz (measured)	330 MHz (simulated)	1140 MHz (from sim.) 746 MHz (from meas.)	545 MHz (measured)
Power dissipation	210 mW @500 MHz (simulated)	105 mW @330MHz (simulated)	106 mW @500 MHz	220 mW @500 MHz

Conventional Separate Data and Instruction Caches

- ● Processor's increasing parallel instruction execution requires adequate cache port numbers
  - Large cache-area increase
  - Lower maximum clock frequency
  - Higher power consumption
- ● Separation of data and instruction cache leads to suboptimal usage of the combined storage capacity
  - Fragmentation



⇨ Unification of data and instruction cache is desirable

Unification Method of Data and Instruction Cache

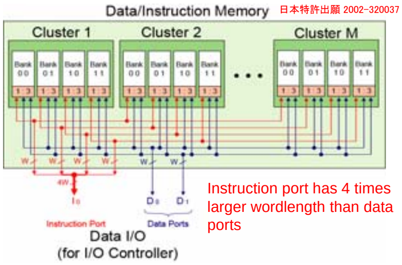
Different access patterns for data and instruction cache

- Instruction cache: Sequential (in order) access is the normal case
  → 1 port with large wordlength is sufficient for access to several instructions
- Data cache : Random access for different data words is normal
  → Increase of the port number is necessary

Assignment between ports and banks must be improved

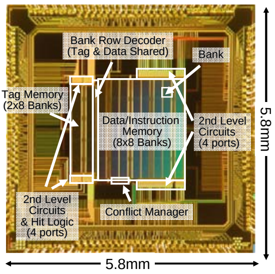
⇩

Realization of different wordlength for each port



Instruction port has 4 times larger wordlength than data ports

4-Port Unified Data/Instruction Cache Design

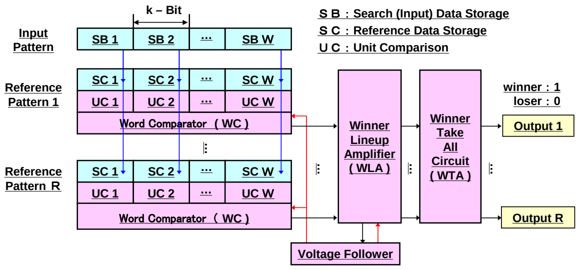

Bank Row Decoder (Tag & Data Shared)
Bank
Tag Memory (2x8 Banks)
Data/Instruction Memory (8x8 Banks)
2nd Level Circuits (4 ports)
2nd Level Circuits & Hit Logic (4 ports)
Conflict Manager
5.8mm
5.8mm

Technology	0.18µm CMOS
Interconnections	5 Metal Layers
Power Supply	1.8 V
Cache Area	6.2 mm^2
Port Number	4 Ports
Total Capacity	128 Kbit
Bank Capacity	2 Kbit
Bank Number	16(Tag) + 64(Data)
Wordlength	16 or 64 bit
Clock cycle	3.4 nsec (Sim.)
Power Dissipation	247mW@250MHz (Sim.)

Service/Instruction ports : 2 ports (64bit); Data ports : 2 port (16bit)

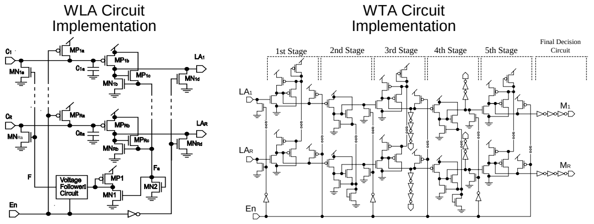
A 1-port cache with the same access bandwidth would need a 340ps clock cycle.

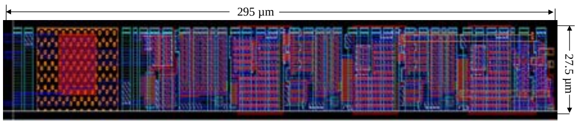
Minimum Distance Search Architecture



Unified architecture of memory and processing unit makes fully parallel minimum distance search possible

Circuits and Layout of Winner-Search Circuits





WLA and WTA Layout in 350nm CMOS

Important Distance Measures

Equation of Application-Relevant Measures:

Distance of reference pattern i : $=\left\{\sum_{j=1}^{W}\left|SB_j-SC_{i,j}\right|^m\right\}^{\frac{1}{m}}$

Bit number k of SB/SC and power m of |SB-SC| :

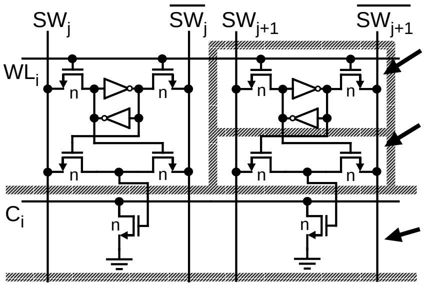
Hamming Distance : k = 1, m = 1

Manhattan Distance : k > 1, m = 1

Euclidean Distance : k > 1, m = 2

Unified Architecture of memory and processing unit allows the realization of important distance measures

Memory-Field Construction for Hamming Distance



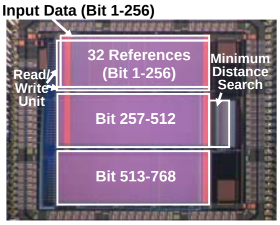
$SC_{i,j+1}$ (Storage Cell)

$BC_{i,j+1}$ (Bit Comparator)

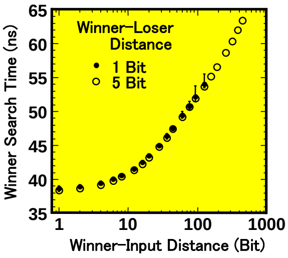
WC_i (Word Comparator)

32 pattern Hamming-Distance Search Example

Input Data (Bit 1-256)



3. 5mm
(3層配線0.6μm CMOS技術)

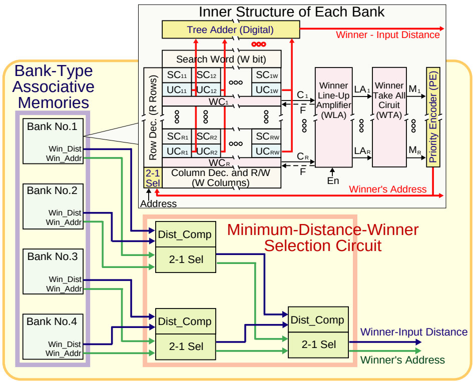


Power Dissipation: ~4.4 mW/mm^2

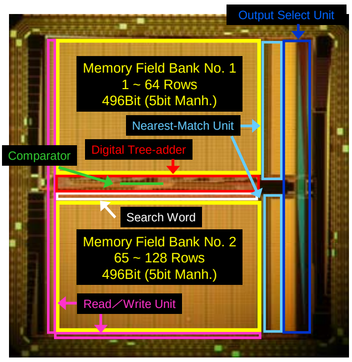
Performance: ~100 GOPS/mm^2

Reliability: < 1.13% Winner-Distance Error

Bank Concept for Large Reference-Pattern Number



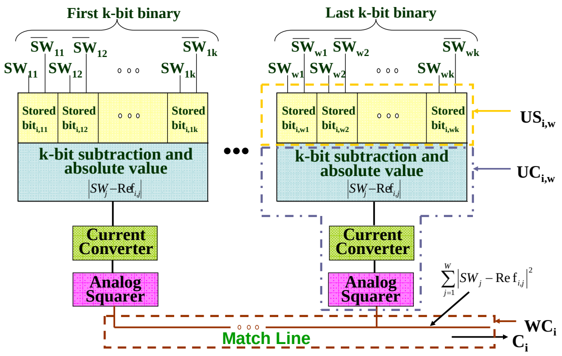
128 Pattern Manhattan–Distance Search Example



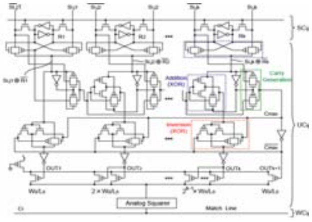
2 Bank Design

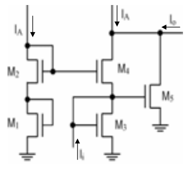
Technology	0.35μm CMOS
Metal Layers	3
Die Size	4.9mm • 4.9mm
Pin Number	144
Supply Voltage	3.3 V
Design Area	14.1$_{mm}^2$ (4.0$_{mm}$ × 3.6$_{mm}$)
Bank Number	2
Reference Pattern	64 × 2=128
Distance	Manhattan Distance
Transitors	516211
Search Time	< 136nsec
Power Dissipation	< 157mW

Circuit Concept for the Euclidean Distance

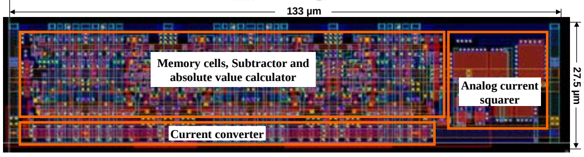

First k-bit binary
Last k-bit binary
$\overline{SW}_{11}$ $\overline{SW}_{12}$ $\overline{SW}_{1k}$
SW_{11} SW_{12} ○ ○ ○ SW_{1k}
$\overline{SW}_{w1}$ $\overline{SW}_{w2}$ $\overline{SW}_{wk}$
SW_{w1} SW_{w2} ○ ○ ○ SW_{wk}
Stored bit$_{i,11}$
Stored bit$_{i,12}$
Stored bit$_{i,1k}$
Stored bit$_{i,w1}$
Stored bit$_{i,w2}$
Stored bit$_{i,wk}$
k-bit subtraction and absolute value
$|SW_j - Ref_{i,j}|$
$US_{i,w}$
$UC_{i,w}$
Current Converter
Analog Squarer
$\sum_{j=1}^{W} |SW_j - Ref_{i,j}|^2$
Match Line
WC_i
C_i

Memory–Field Construction for Euclidean Distance



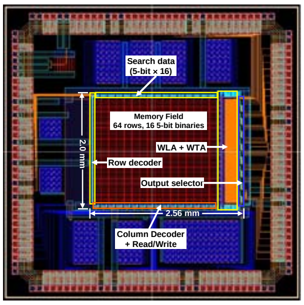


Analog current squarer



Layout of 5-bit Euclidean Distance Memory Block in 350nm CMOS

64 Pattern Euclidean Distance Search Example



Distance Measure	Euclidean-Distance
Reference Patterns	64 Patterns (16 binaries each 5-bit long)
Design Area	5.12 mm^2 (2.56mm x 2mm)
Nearest Match Unit Area	0.53mm^2 = 11.1% of design area
Nearest Match Times (simulation)	< 157 nsec
Power Dissipation (simulation)	< 195 mW
Chip size	4.9 mm × 4.9 mm
Chip pin	144
No. of Transistors	1,86,648
Technology	0.35 μm, 2-poly, 3-metal CMOS
Supply Voltage	3.3V

Associative Memory's Computational Performance

C-Code

```
for (i = 0); i < 64; i++) {
    for (j = 0; j < 16; j++) {
        S[j] = abs( In[j] − Ref[i][j] );
        D[i] += S[j]*S[j]; }
    if ( min > D[i] ) min = D[i];
}
```

Compile

Assembler Program

```
main:
    pushl  %ebp
    movl   %esp, %ebp
    andl   $−16, %esp
    addl   $15, %eax
    ■
    ■
L5:
    cmpl   $15, −4496(%ebp)
    jg     L6
    movl   −4496(%ebp), %ecx
    movl   −4492(%ebp), %eax
    sall   $6, %eax
    ■
    ■
```

Count Operations

Necessary Operations:	22,159
Associative Memory's Computation Time:	158ns

Equivalent Performance = Operations/Time = 140 GOPS

An 8-bank associative memory with 64 patterns per bank achieves a performance of 1.12 TOPS

Conclusion

- ✦ Data transmission between memory and processing unit limits the performance improvements of integrated systems.
- ✦ Two methods for mitigating this problem have been proposed:
  - – Bank-based Multi-porting of the memory
  - – Unification of memory and processing unit
- ✦ Applications of these two methods lead to key technologies for terabit information processing, enabling in particular:
  - – Tera-bit-per-second (Tbps) memory-access bandwidth
  - – Tera-operation-per-second (TOPS) processing power for the pattern-matching function