# Memory Hierarchies in Intelligent Memories: Energy/Performance Design

Wei Huang, Jose Renau, Seung-Moon Yoo and Josep Torrellas

University of Illinois at Urbana-Champaign

# Motivation

- Advances in technology:
  - Processor and Memory integration
  - Many processors on a chip
- How to design for high performance
- Energy consumption is a big concern
- Problems in cooling system



# Goals of this work

- Evaluate trade-offs in memory hierarchy
  - Energy consumption
  - Performance
  - Area requirements
- Detailed energy consumption analysis

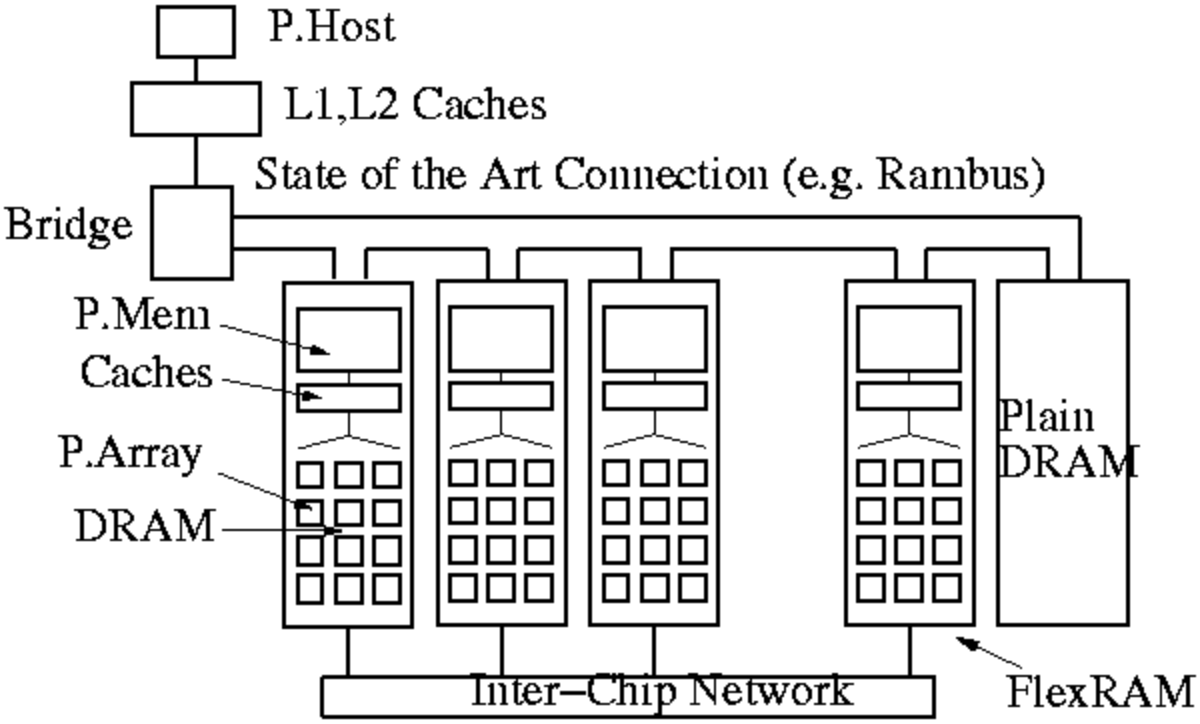


# Findings of the Work

- Modest cache size is necessary
- Easy modifications in memory reduce energy consumption



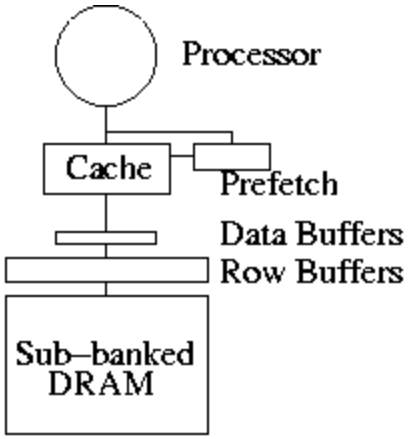
# The FlexRAM Architecture

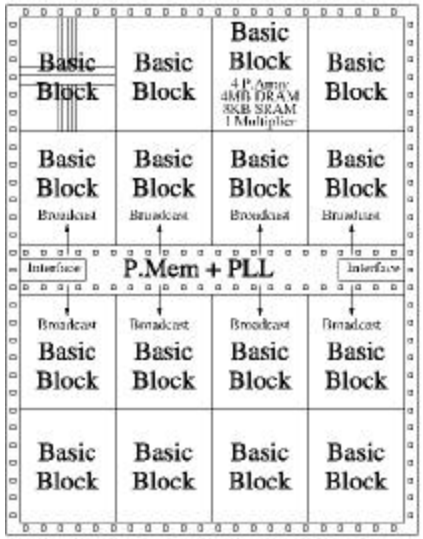



Yi Kang, Wei Huang, Seung-Moon Yoo, Diana Keen, Zhenzhou Ge, Vinh Lam, Pratap Pattanaik and Josep Torrellas - ICCD99

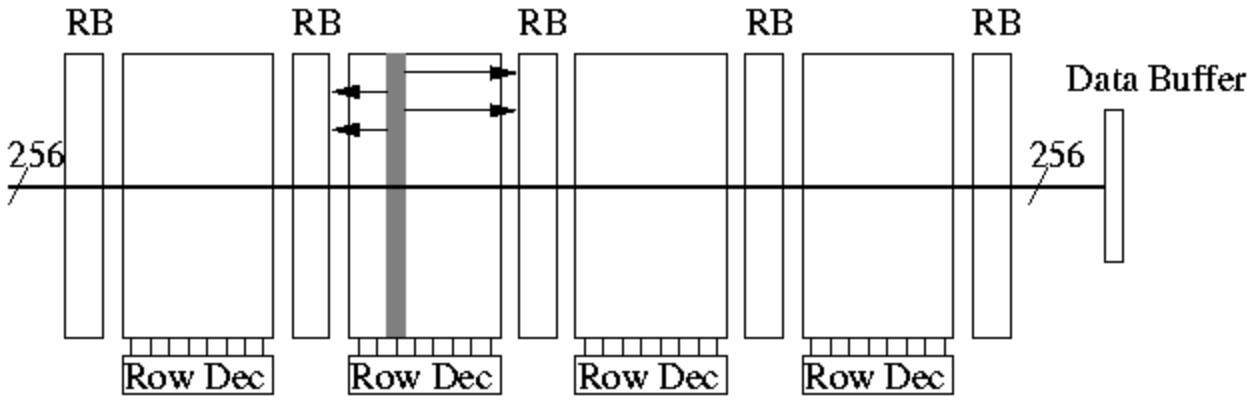
# Chip Architecture

- 64 nodes, each one includes:
  - 2-issue processor @800Mhz
  - 1MByte DRAM (12 clk)
  - Row Buffers (6 clk)
  - Cache (1 clk)



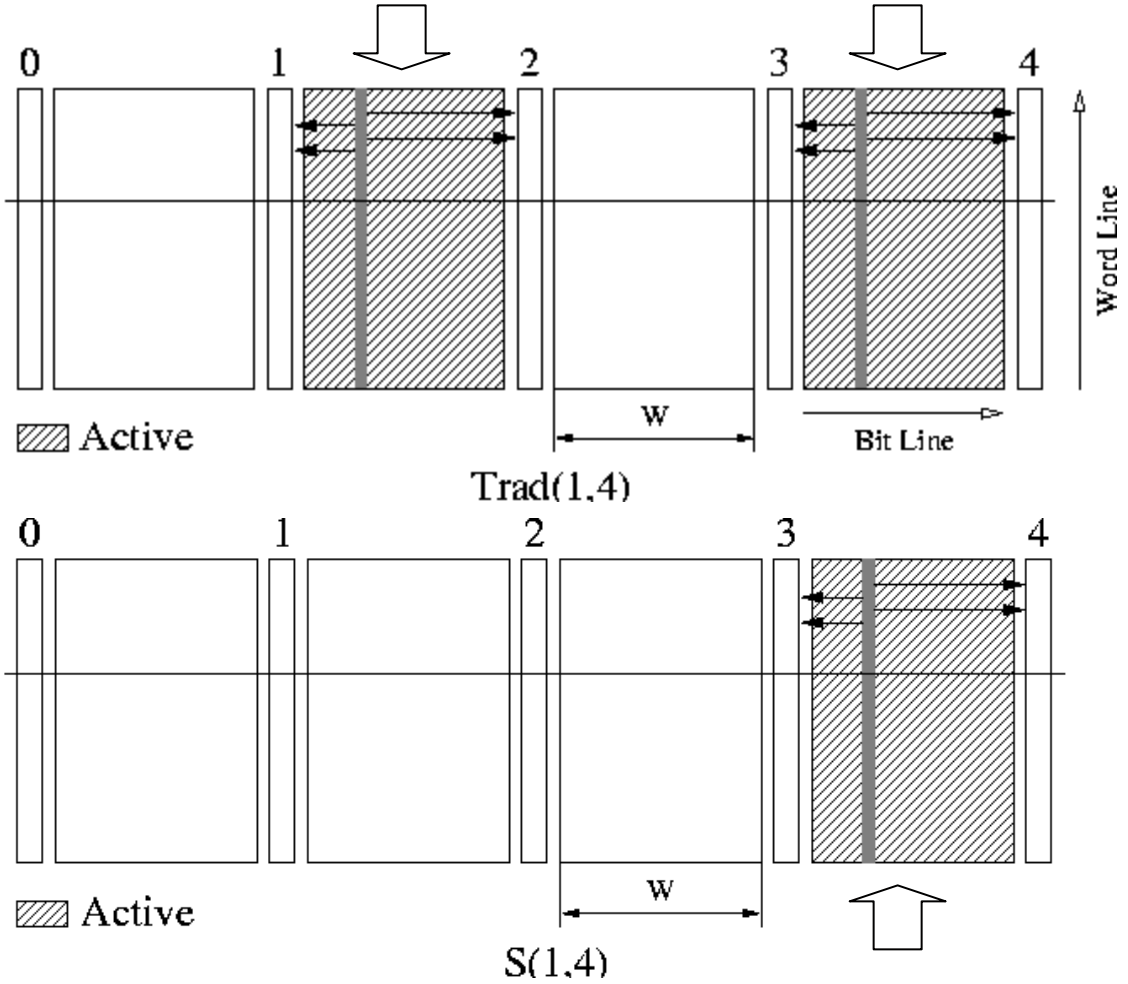



# How a Memory Bank Works




- 4 Memory sub-banks, each 256KBytes
- 5 Row Buffers, each 1KByte
- 1 Data Buffer 256bits

# Small Area Memory Banks

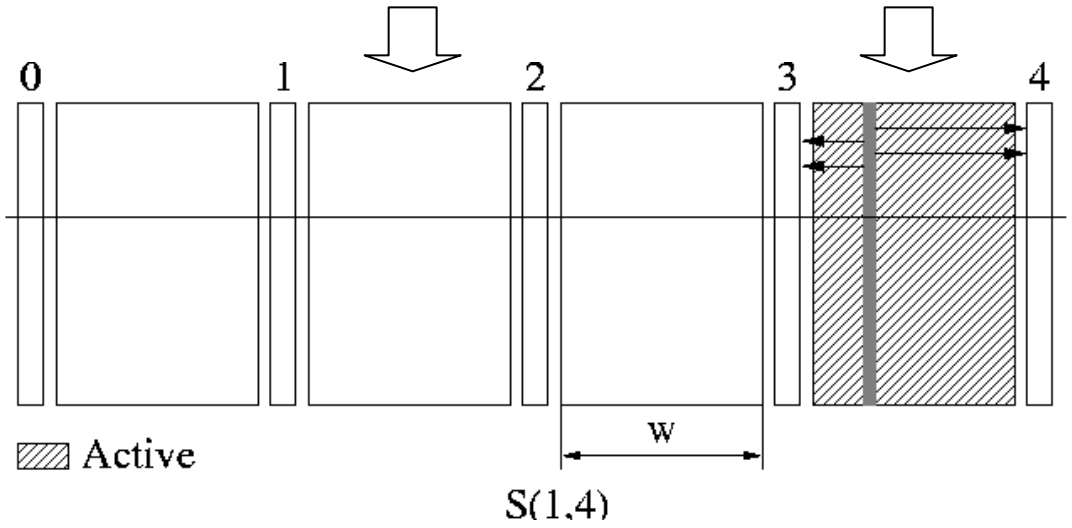



+More Energy

+Less Spatial Locality

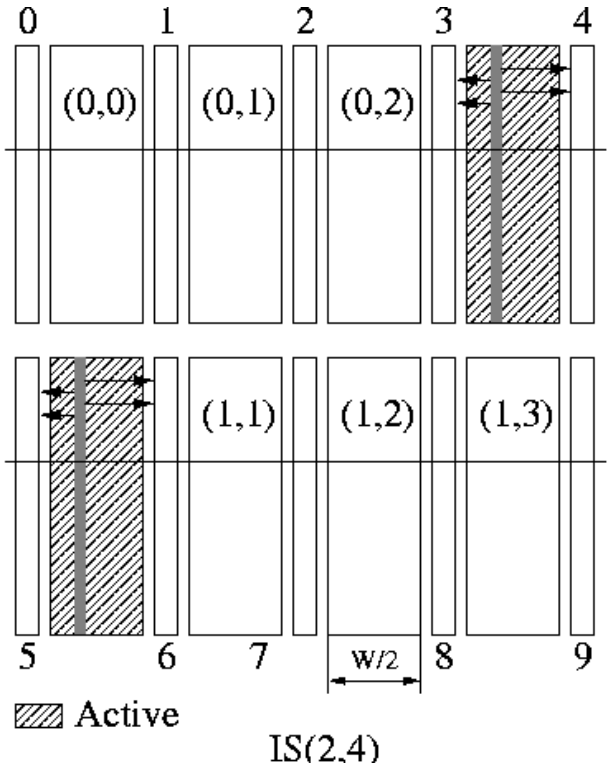
+More Localities

# Pipelining the requests




Faster memory system without increased energy consumption

# Advanced Memory Banks I

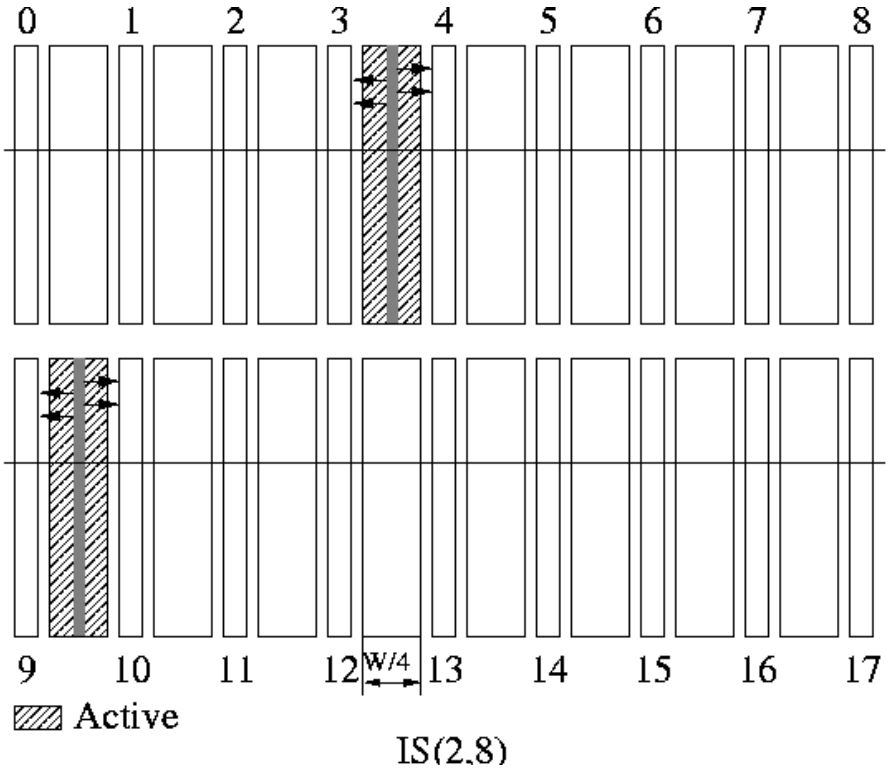


IS(2,4)

-Less Energy and Contention

+More Area

# Advanced Memory Banks II




-Less Energy and Contention

+More area

# Terminology for Memory Systems

- Trad(i,j): Traditional
- S(i,j): Segmented
- IS(i,j): Interleaved Segmented
- ISP(i,j): Interleaved Segmented Pipelined

- i : Degree of interleaving
- j : Number of sub-banks per interleaving way



# Energy and Area Issues

| Access Type | Trad(1,4) | S(1,4) | IS(2,4) | IS(2,8) |
|---|---|---|---|---|
| Cache hit (8KB) | 191pj | 191pj | 191pj | 191pj |
| RB Hit | 468pj | 468pj | 506pj | 517pj |
| Bank Access | 6999pj | 3729pj | 2287pj | 1556pj |

| | Trad(1,4) | S(1,4) | IS(2,4) | IS(2,8) |
|---|---|---|---|---|
| Area (.18µm) | 4.25mm$^2$ | 4.25mm$^2$ | 4.83mm$^2$ | 5.23mm$^2$ |

- More advanced configurations:
  - More Area
  - Less Energy



# Evaluation Environment

- Fixed parameters:
  - 2-issue processor @800MHz
  - Prefetch
  - Cache, RB, Bank latencies (1,6,12 cycles)
- Variable parameters:
  - Cache sizes (256B,1KB,8KB,16KB)
  - Memory Banks:

    Trad(1,4),S(1,4),SP(1,4)

    IS(2,4),ISP(2,4),IS(2,8),ISP(2,8)

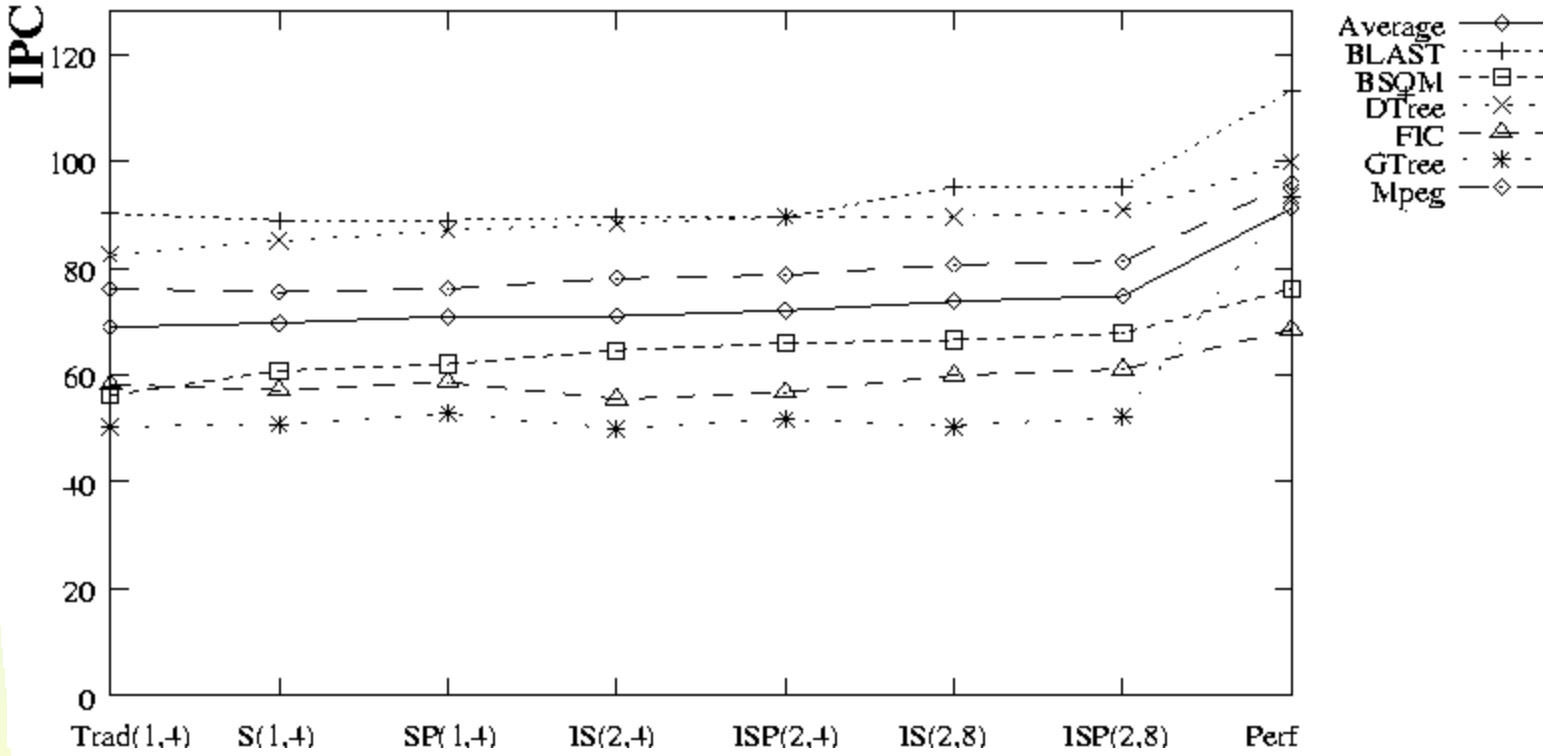


# Applications

| Applic | What It Does | Cache Hit Rate (%) |
|---|---|---|
| GTree | DM Tree Generation | 50.7 |
| DTree | DM Tree Deployment | 98.6 |
| BSOM | BSOM Neural Network | 94.7 |
| BLAST | Protein Matching | 96.9 |
| Mpeg | Mpeg-2 Motion Estimation | 99.9 |
| FIC | Fractal Image Compressor | 97.8 |

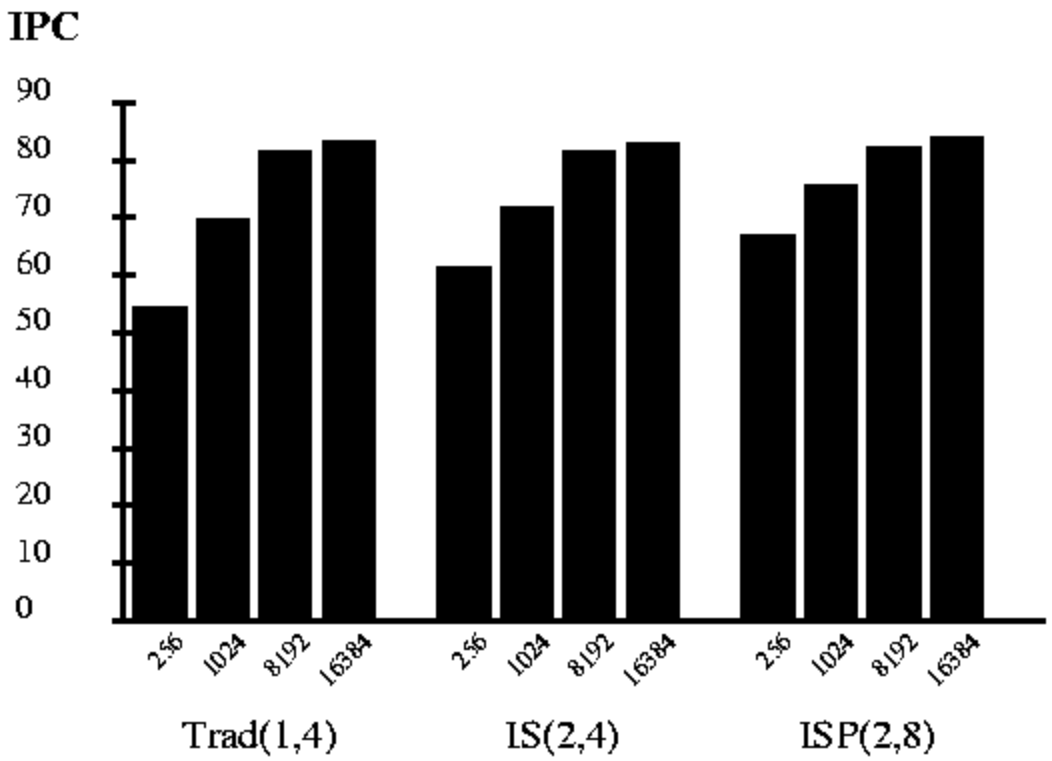


# Performance: Memory Banks



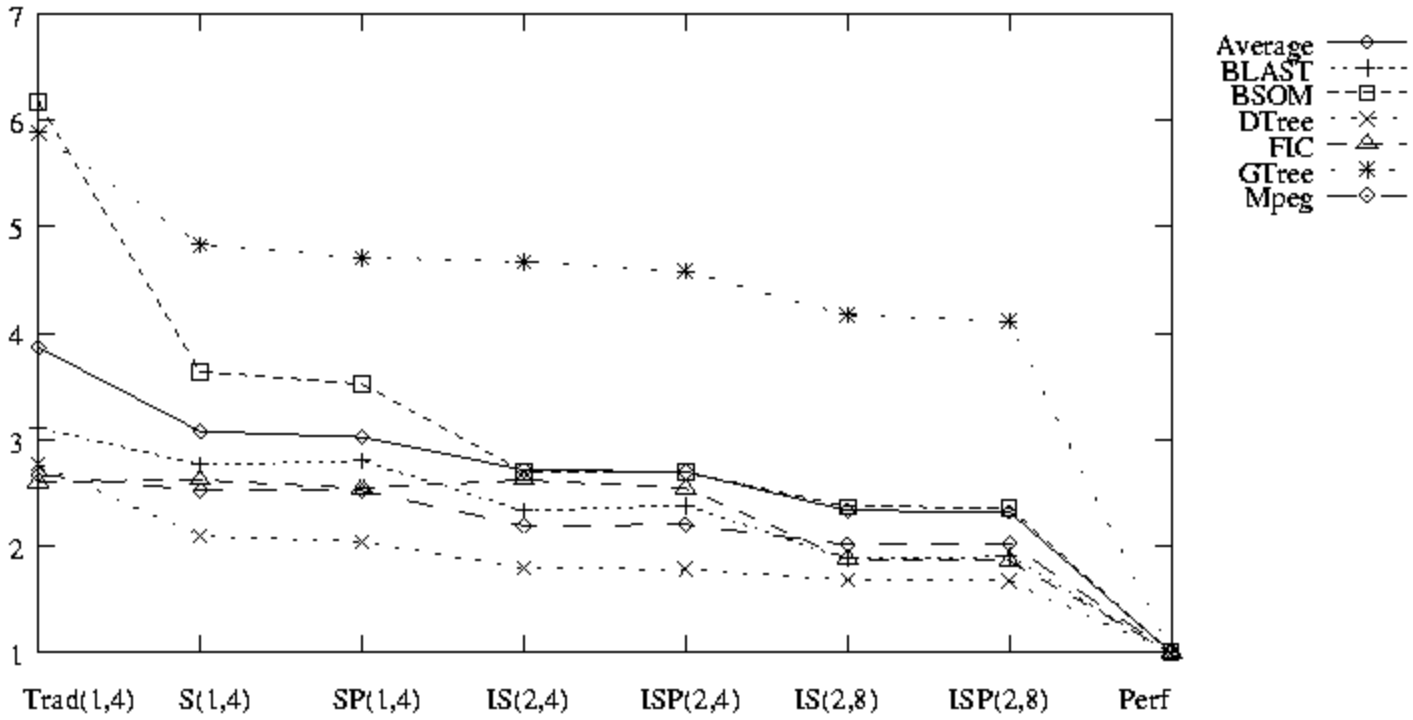

- Small performance improvement in advanced configurations with 1KByte cache

# Performance: Cache Effect



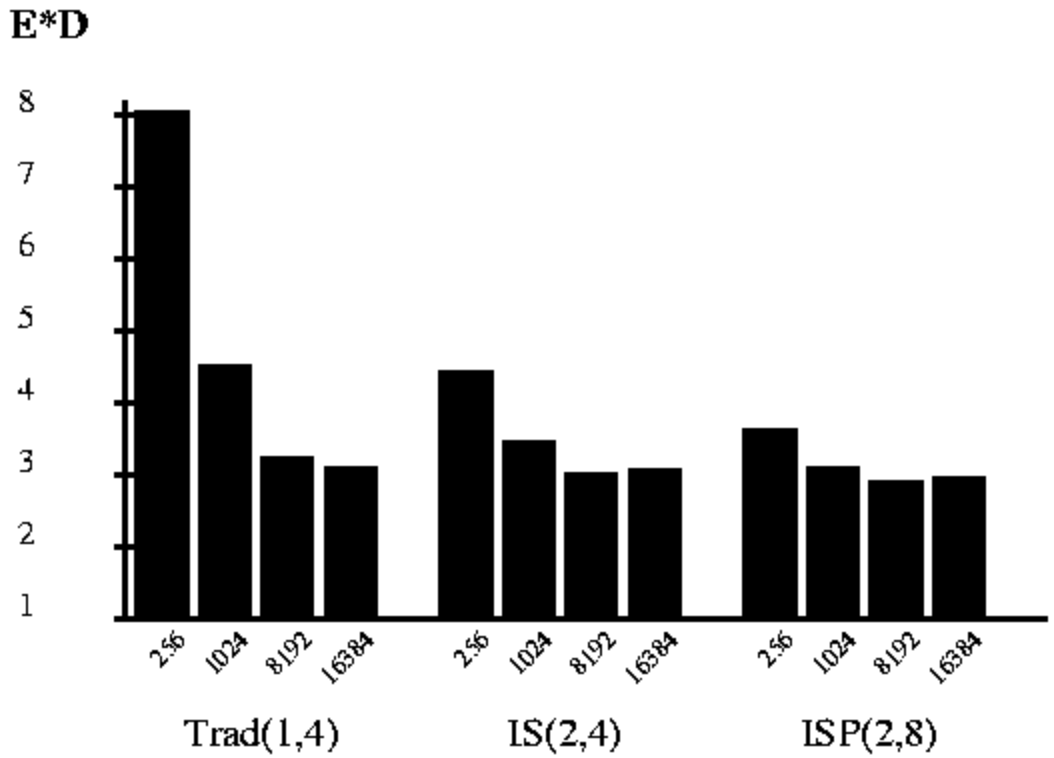

- Modest cache size is required for performance

# Energy-Delay Product




- Big improvement in energy-delay product with more advanced memory configurations

# Energy-Delay Product: Cache




- 8KBytes have the best energy-delay product

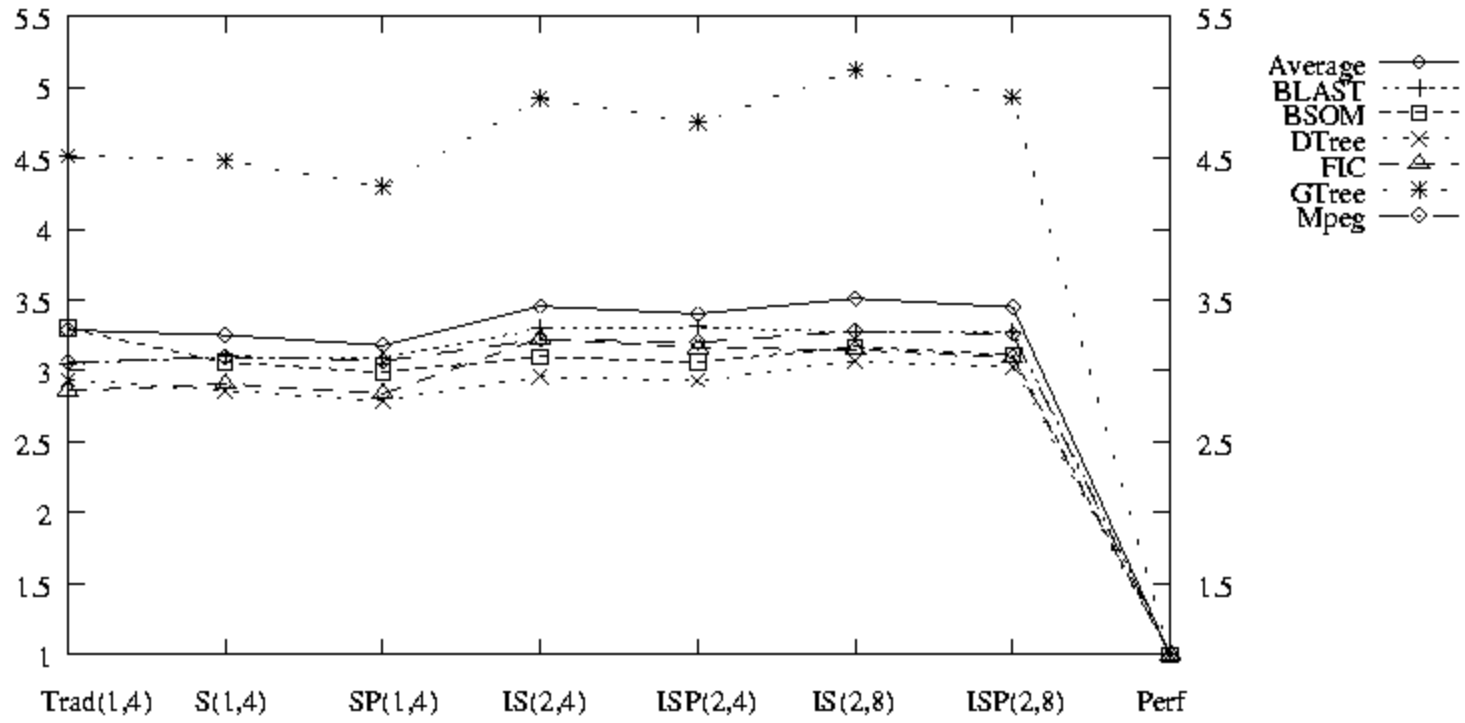
# Conclusions

- Modest size cache is enough (8KBytes)
  - Improves performance
  - Reduces energy consumption
- Segmentation S(1,4)
  - Reduces energy consumption
- When area is available: use interleaving
  - IS(2,4) increases by 14% the area



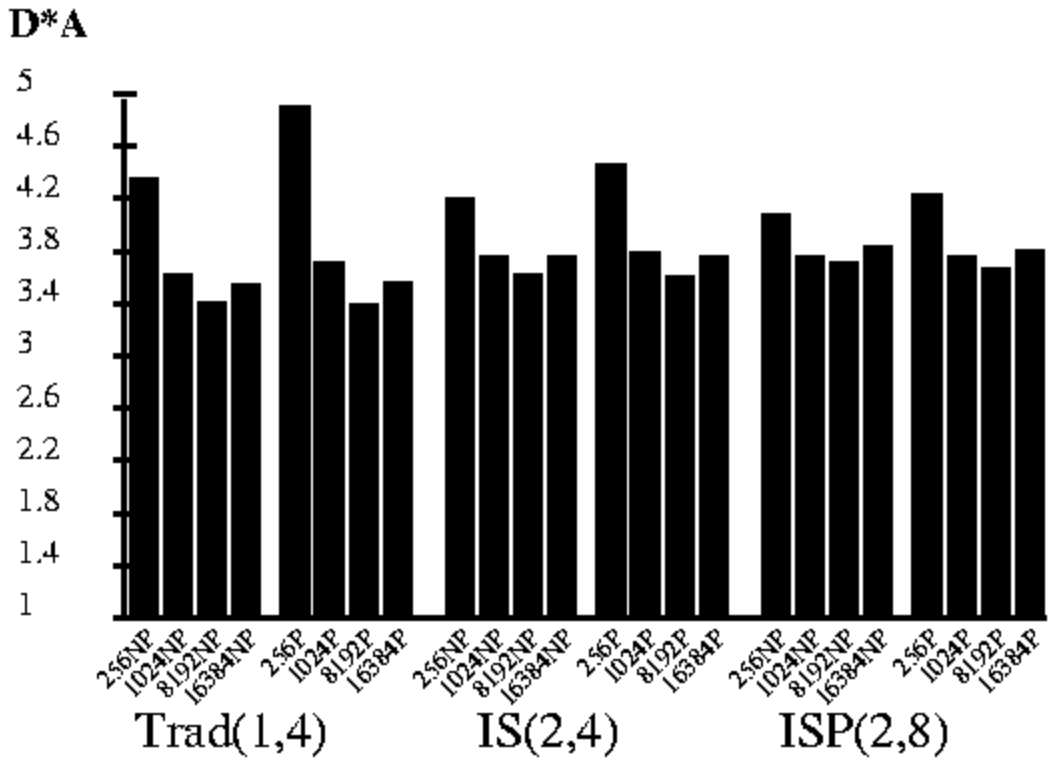
# Backup Slides

# Area-Delay Product: MB



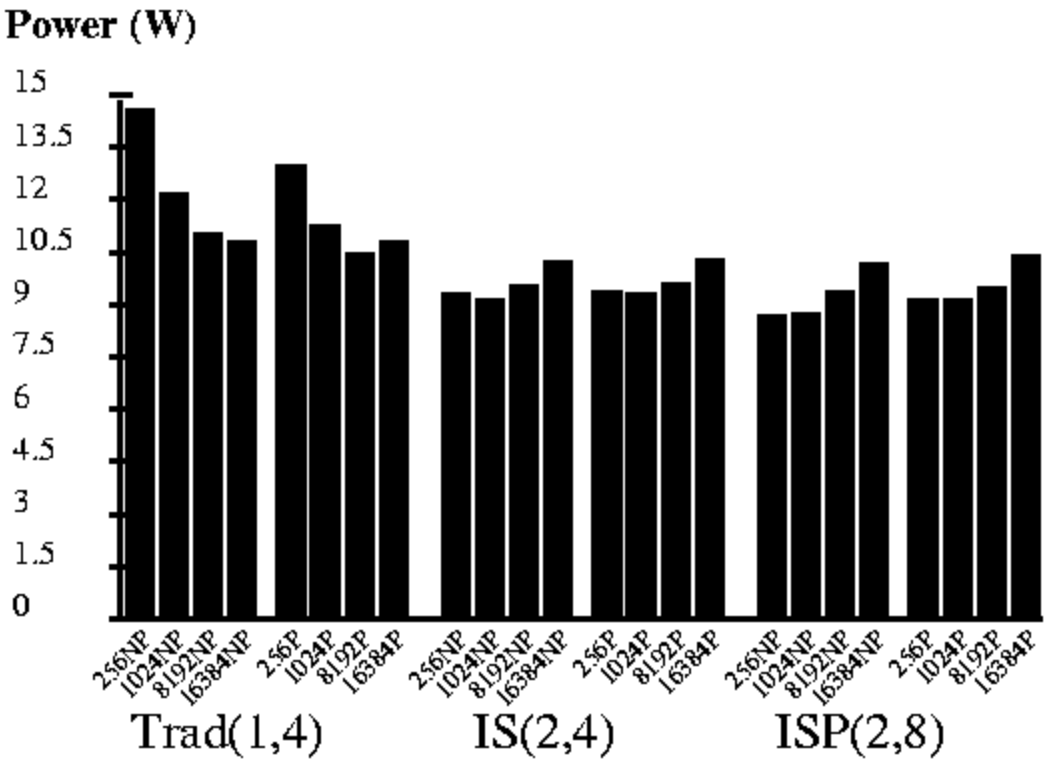

- SP(1,4) best are utilization

# Area-Delay Product: Cache




- 8KBytes is a sweet point for area-delay product

# Power Consumption: Cache




- Power is a bad metric, only useful as a constraint

# Memory Access Timing

| Operation | Cycles |
|---|---|
| X-address buffer | 1 |
| X-address decoder | 1 |
| Wordline enabling | 2 |
| Charge sharing | 2 |
| Bit line sensing | 2 |
| DRAM Data buffer | 2 |
| L1 Cache | 1 |
| **Total** | **11** |



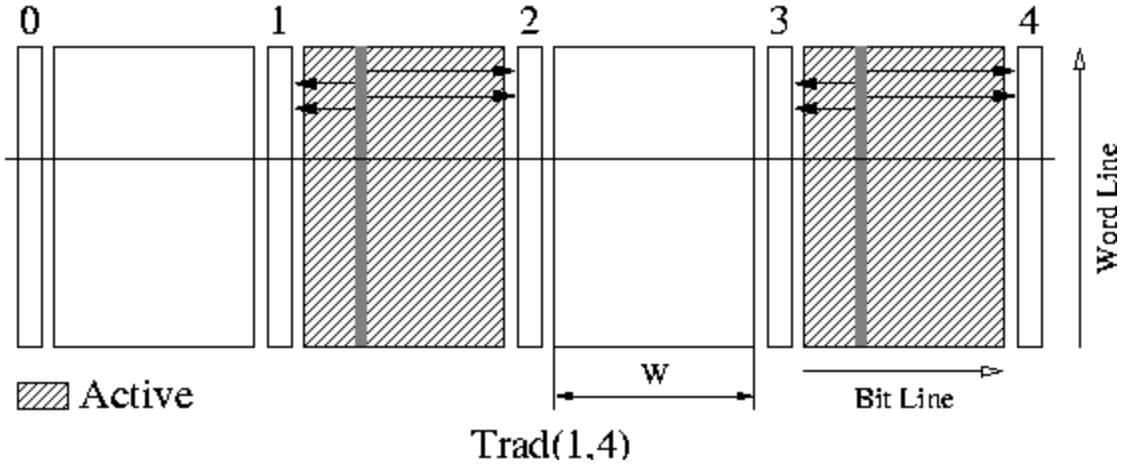
# Area Requirements

| Cache size | Area |
|---|---|
| 256B | 0.07 |
| 1K | 0.16 |
| 8K | 0.60 |
| 16K | 1.15 |

| Bank size | Area |
|---|---|
| (1,4) | 4.25 |
| (2,4) | 4.83 |
| (2,8) | 5.23 |

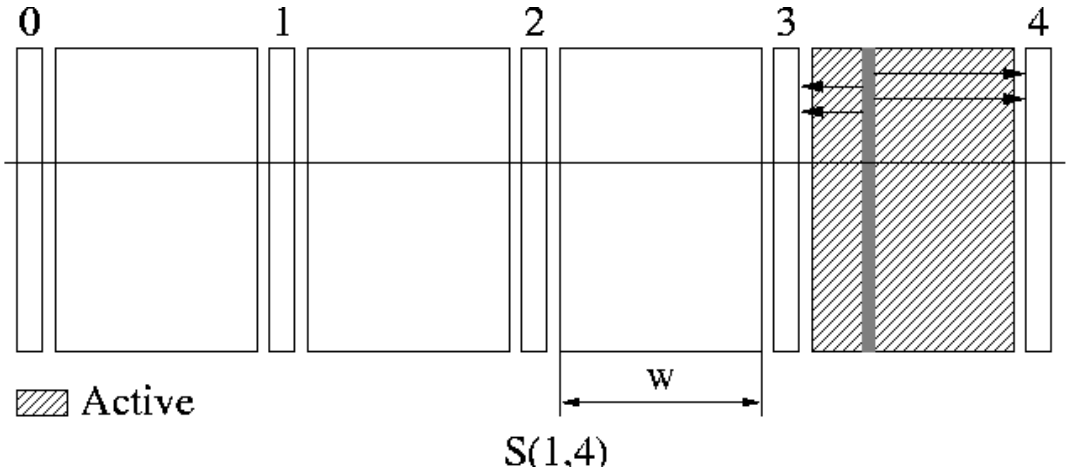


# Small Area Memory Banks



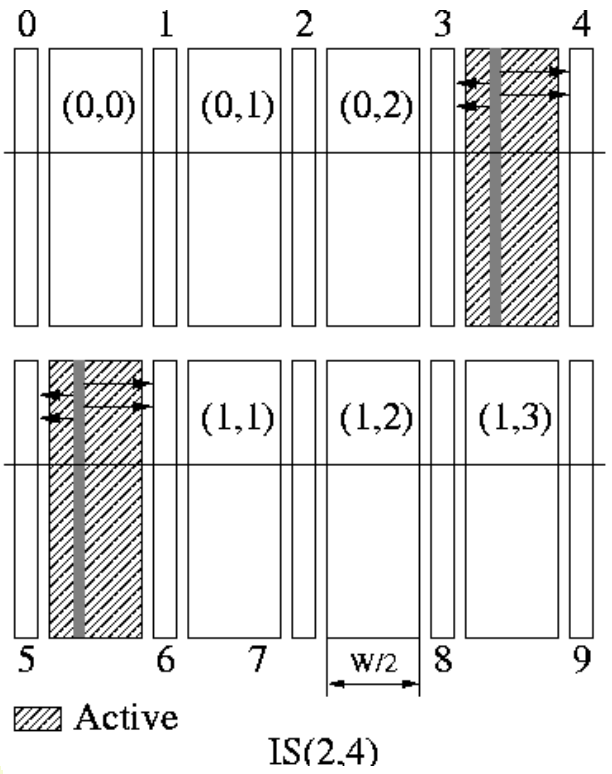

+Energy

+Spatial Locality



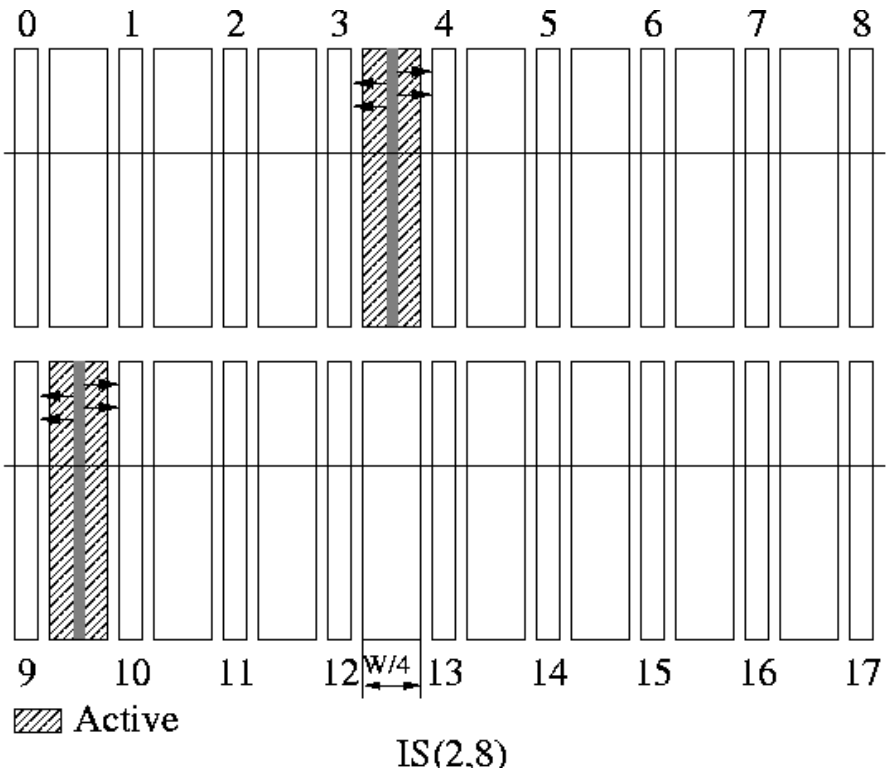

+Localities

# Advanced Memory Banks




Less Energy and Contention

More area




Less Energy and Contention

Even more area