FlexRAM: Toward an Advanced Intelligent Memory System

A Retrospective Paper

Josep Torrellas

University of Illinois

http://iacoma.cs.uiuc.edu

Original paper by Y. Kang, W. Huang, S. Yoo, D. Keen, Z. Ge, V. Lam, P. Pattnaik, and J. Torrellas appeared in ICCD-1999

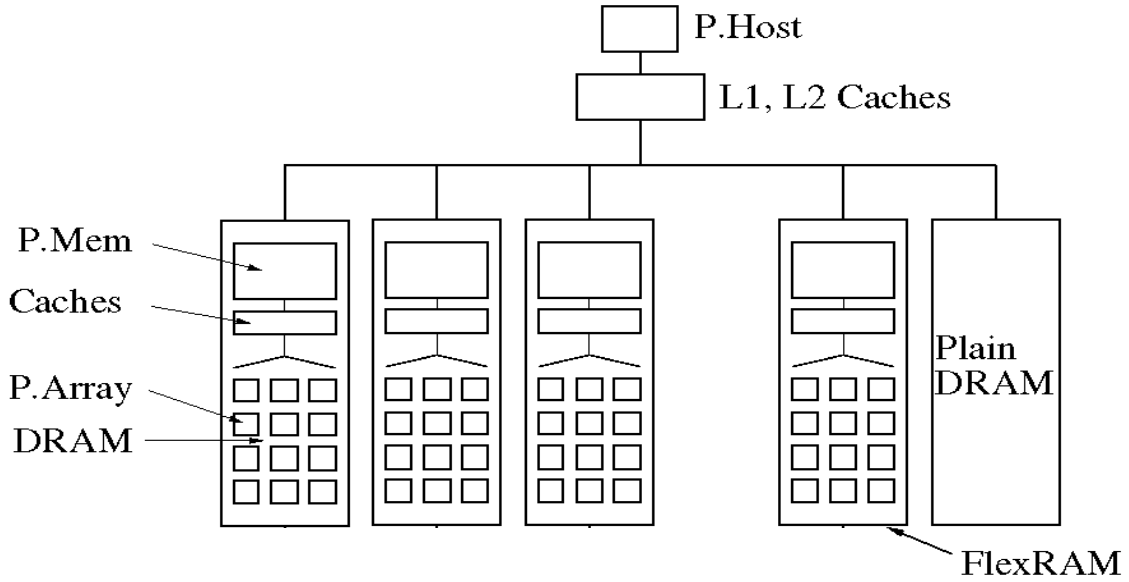
i-acoma group

ILLINOIS

Context of the Work

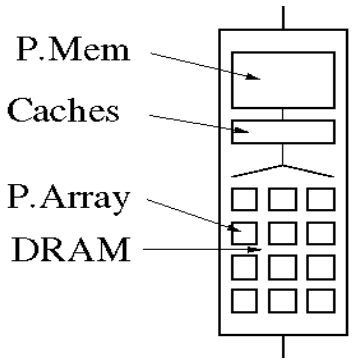
- Work started in 1996
- Processing In Memory (PIM) or Intelligent Memory:
  - Integration compute + memory for improved latency/bandwidth/power
- Great interest in the community
  - Peter Kogge: EXECUBE (1994)
  - Berkeley IRAM, workshop
  - Bob Lucas at DARPA
  - IBM 7LD, Mitsubishi ERAM
- Our focus at UIUC and sabbatical at IBM Watson → applications

What the Paper Was About



- Unmodified machine + many cores in the memory system
- Cores are tiny and numerous
- General purpose rather than trying to pattern-match an application

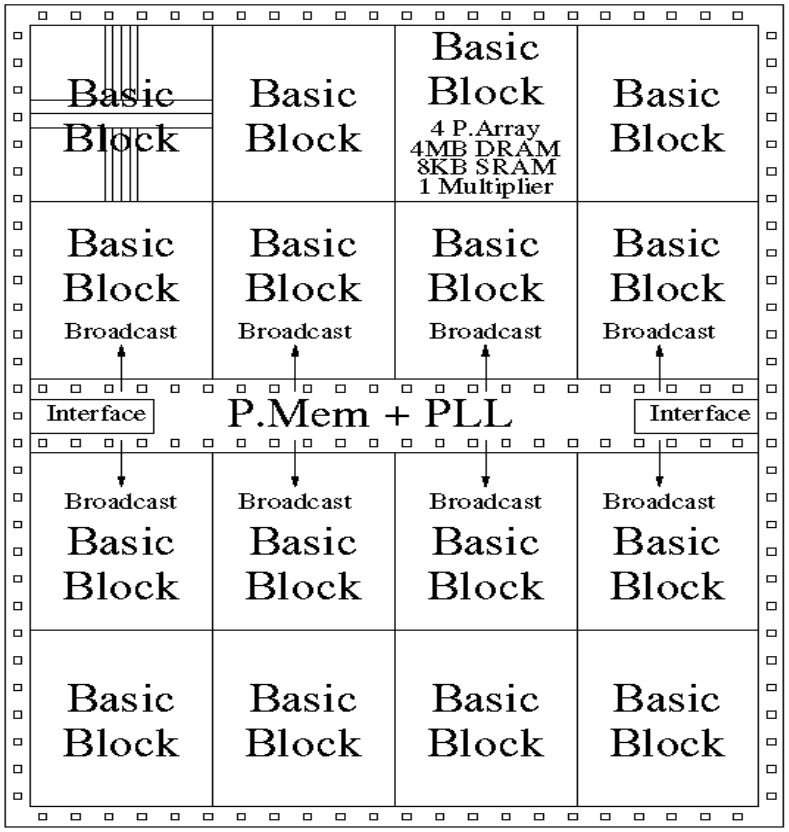
The FlexRAM Chip

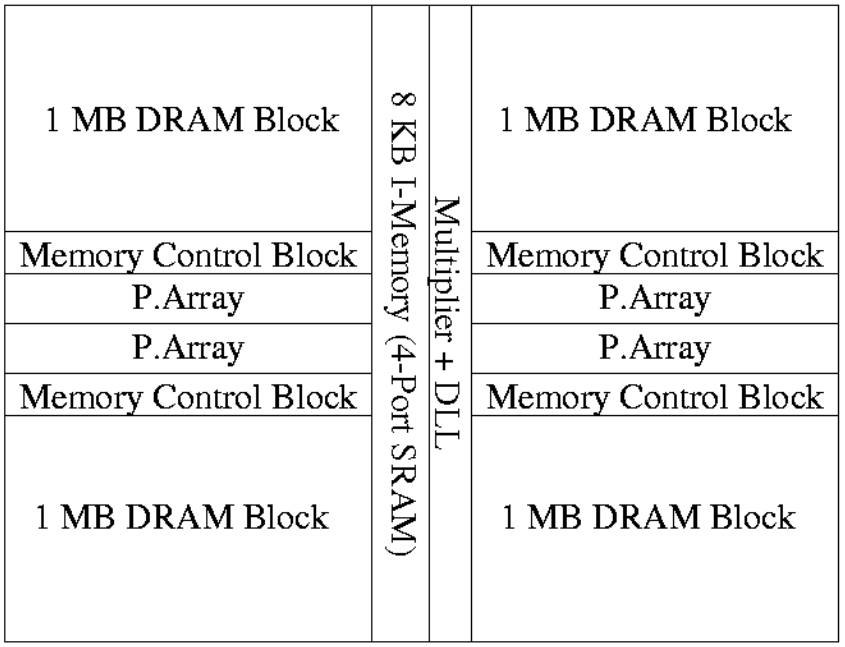


64 Mbytes
of DRAM

- P.Arrays:
  - 64 single-issue in-order cores in SPMD
  - Each sees 1MB of DRAM and can communicate with 2 neighbors
- P.Mem:
  - A 2-issue in order core for broadcast and reduction

The FlexRAM Chip

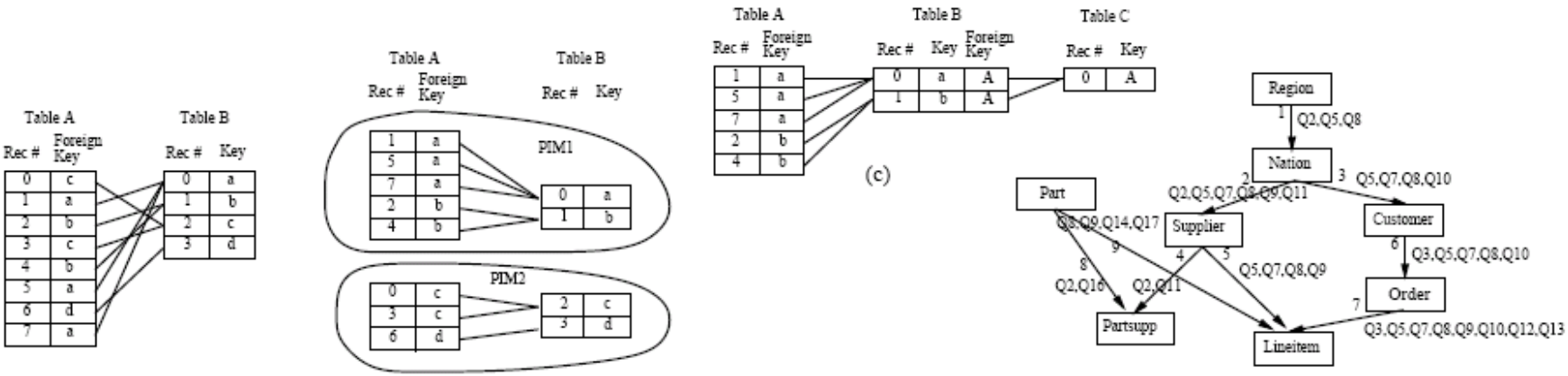




What the Paper Was About (II)

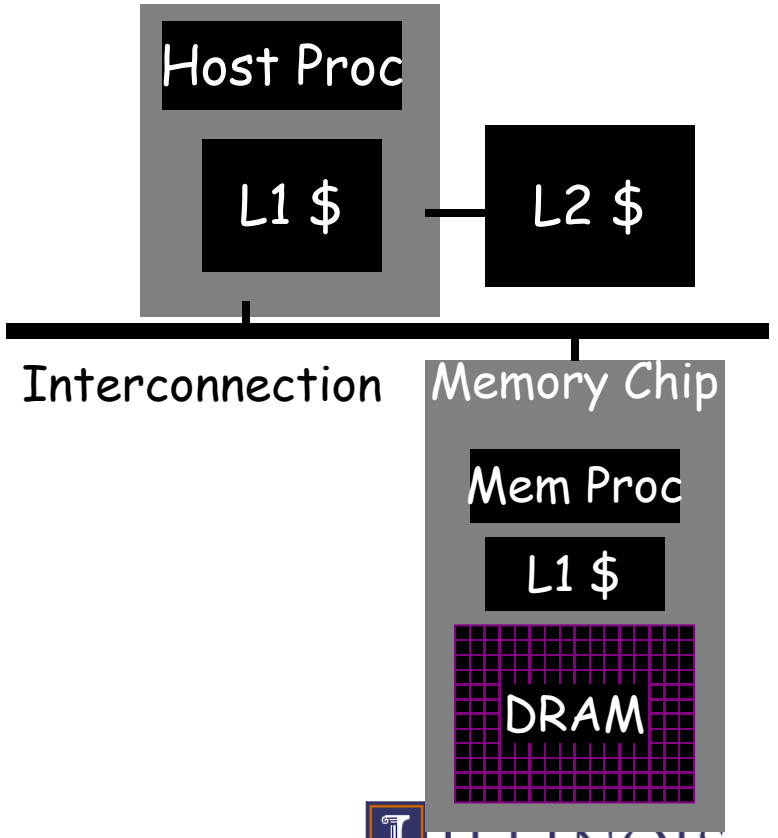
Mapping applications to the intelligent memory

- Data mining (decision trees and neural networks)
- Computational biology (protein sequence matching)
- Multimedia
- Decision support systems (TPC-D)
- Speech recognition



The FlexRAM Project

- Work of about 10 grad students
- Collaborated with other faculty (Padua, Chien, Reed, Huang) and scholars (Fraguela, Lee, Feautrier, Pattnaik, Ekanadham, Lim)
- Published about 15 papers
  - Programming environment
  - Mapping code to heterogeneous platforms
  - Memory-side prefetching



Programming Environment

- CFlex: OpenMP-like directives

```
int TreeAdd (register tree_t *t) {
   if (t == NULL) return 0;
   else {
     int leftval, rightval;

#pragma FlexRAM phost sync
     {
#pragma FlexRAM parray async on_home(*(t->tleft)) if (lcl(t->left))
       leftval = TreeAdd(t->left);

#pragma FlexRAM parray async on_home(*(t->tright)) if (lcl(t->right))
       rightval = TreeAdd(t->right);
     }

     return leftval + rightval + t->val;
   }
}
```

Programming Environment

- IMOs: Library of Intelligent Memory Operations

Array IMOs	Example
Element-by-element (with optional mask): *add, mpy, and, or*	x(i) = (y(i)+z(i)) mask w(i)
Reduction: *sum, product, or, and, minvalue, maxvalue*	x = or(y(i))
Recurrence: *linear first order, linear second order*	x(i)=a(i)*x(i-1)+b(i)
Manipulation: *rotate, transpose, tile*	x(i,j)=transpose(y(i,j))
Set IMOs	**Example**
Set to set: *union, intersection, difference, equal*	S3 = union(S1,S2)
Element to set: *in, from, include,remove,exists,forall*	Bool = in(x,S1)

Intelligent Memory: What has happened?

- Excitement in the field lasted for a few years past 2000
  - Other projects: UC-Davis Active Pages, ISI DIVA
  - DARPA started a program (Stanford Smart Memories)
  - Company products: Mitsubishi M32Rx/D microcontroller
  - Micron working on Yukon Active Memory (256 procs in memory chip)

- Excitement fizzled soon:
  - Economies of building specialized DRAM with suboptimal integration and non-standard interfaces
  - Less disruptive technologies like Multi-Chip Module (MCM):
    - IBM POWER5 in 2004: MCM with 4 proc dies + 4 cache dies

Intelligent Memory: What has happened?

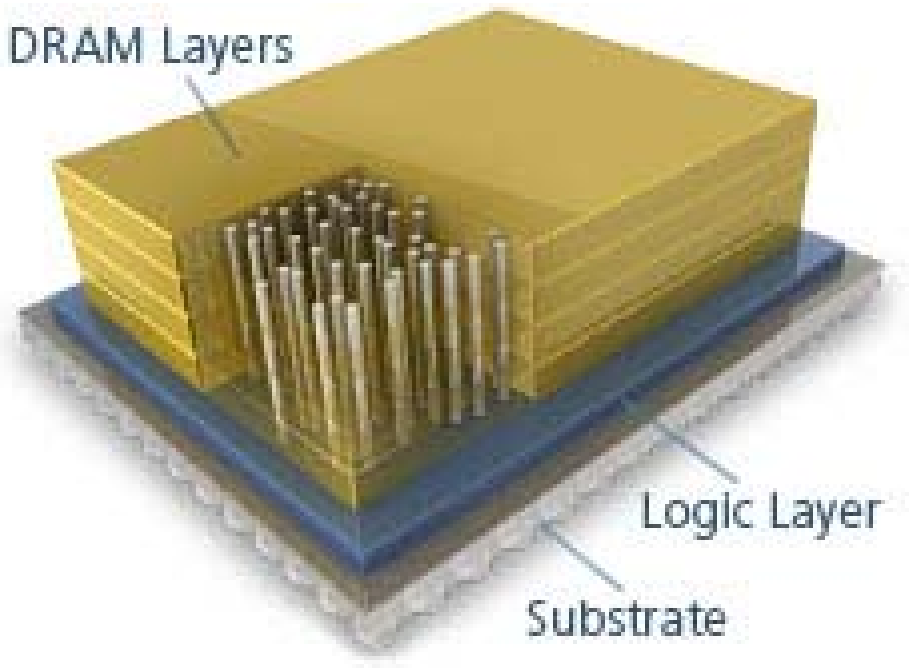
- Within supercomputing R&D, always interest in PIM
  - More integration → more costly to go out to memory
  - Moving data in/out consumes energy: best to compute “in place”
  - Many positive research results

- Even in this market segment, commercial considerations prevented PIM
  - DARPA HPCS systems: from many sophisticated PIM to no PIM

Intelligent Memory: The Future

- Technical rationale for PIM is compelling → likely used in future
- We may be watching it take shape:
  - 3D ICs that stack multiple memory dies over a logic die (Samsung, Micron)

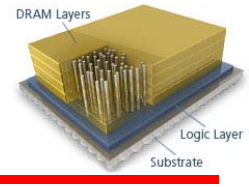
Micron's Hybrid Memory Cube (HMC)



- August 2011: Announced by Micron
- September 2011: Intel announced it is working with Micron on HMC

- Memory chip with 4 or 8 DRAM dies over 1 logic die
- Can be placed in an MCM with processor dies
- DRAM dies only store data while logic die handles DRAM control

Micron's Hybrid Memory Cube (HMC)



- Current logic die:
  - Advanced memory controller functions + self test
  - Improves bandwidth, latency, and energy
- Future logic die:
  - Support for Intelligent Memory Operations?
    - Preprocessing data as it is read from memory
    - Performing processor commands "in place"

- Still uncertainties: Who will design and test the logic? A memory company?

Conclusion

- Commercial considerations repeatedly prevented the emergence of PIM
- We may be finally seeing the beginning of PIM

FlexRAM: Toward an Advanced Intelligent Memory System

A Retrospective Paper

Josep Torrellas

University of Illinois

http://iacoma.cs.uiuc.edu

Original paper by Y. Kang, W. Huang, S. Yoo, D. Keen, Z. Ge, V. Lam, P. Pattnaik, and J. Torrellas appeared in ICCD-1999

i-acoma group

ILLINOIS