CS533: Processor in memory (PIM)

Josep Torrellas

University of Illinois in Urbana-Champaign

Today's Paper

A Case For Intelligent RAM

D. Patterson, T. Anderson, N. Cardwell, R. Fromm, K. Keeton, C. Kozyrakis, R. Thomas, K. Yelick
University of California Berkeley

Main Idea

- memory system is the greatest inhibitor of performance (low bandwidth, high latency)
- therefore: integrate a high-performance processor & DRAM main memory on a chip
  - low latency = 20-30ns instead of 300ns
  - high bandwidth = 128 bytes instead of 16 bytes

Tradeoffs

Advantages

- high bandwidth to memory
- low latency to memory
- energy consumption in memory system decreases several times
  - reduction of off-chip accesses (high capacitance)
- fewer pins necessary in chip (currently, most pins used for wide mem. interface)
  - smaller packages thanks to fewer pins
  - regular chp layout, more dense

Tradeoffs (Continued)

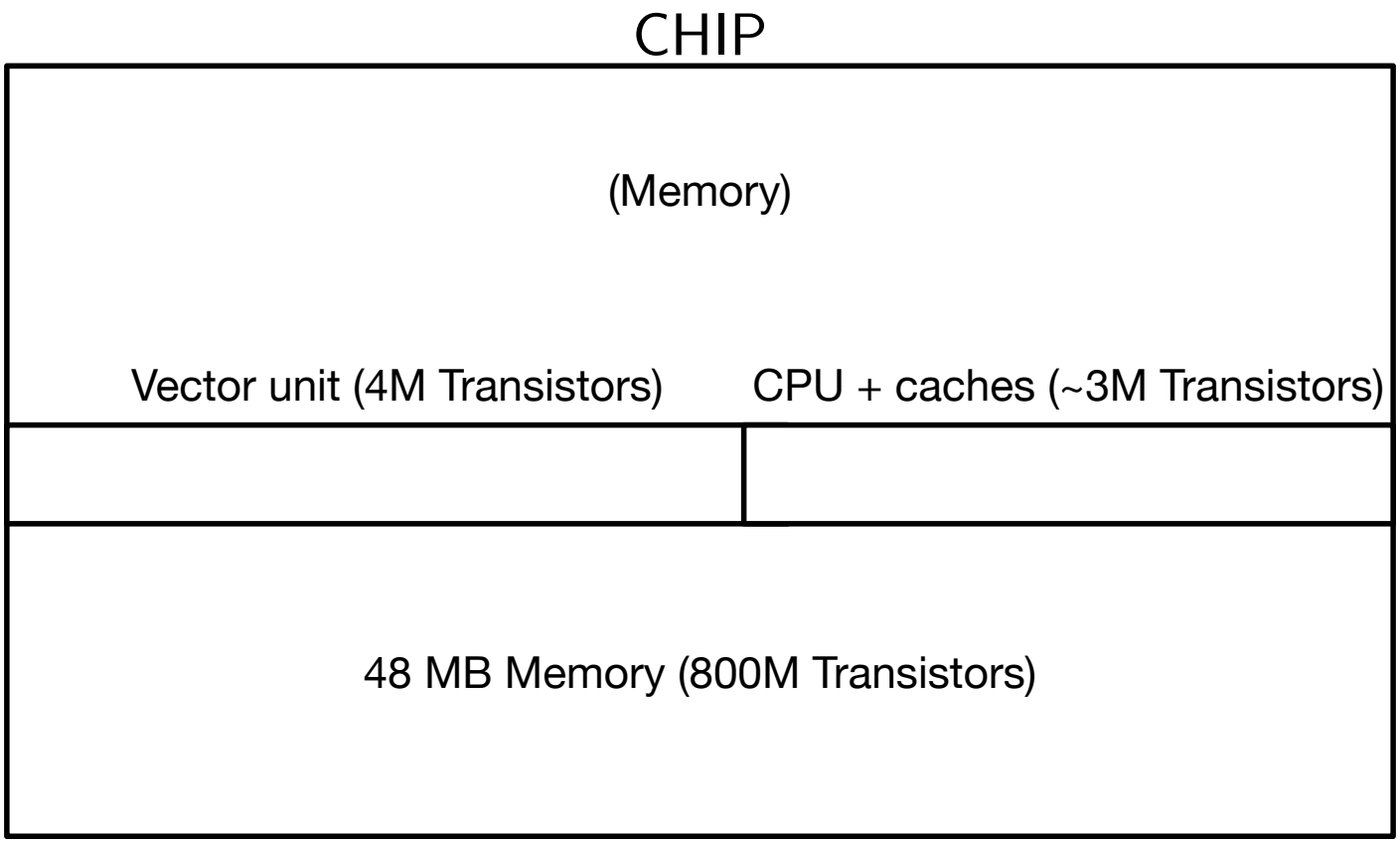
Advantage

- can be combined with any processor organization

Disadvantages

- conventional processor design gains little from IRAM because they were designed with assumption of slow memory system
  - $\rightarrow$ need to open up the memory
- if the programming model is too revolutionary, old apps will not run

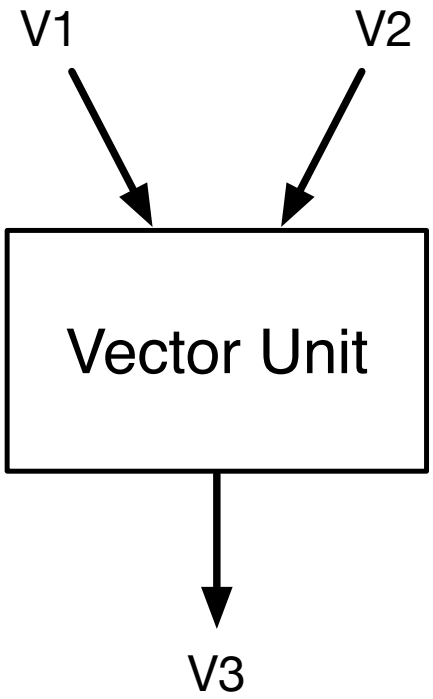
Layout



- DRAM technology (memory) is much more dense than SRAM technology (caches)
  - 16 to 32 times more $\Longrightarrow$ more storage on chip

Vector Processors

- ↑ capacity ↑ bandwidth $\Longrightarrow$ vector processing would work well



MULTV V1,V2,V3

– Need explicit compilation into vector code

\+ it is claimed that many multimedia apps will be vectorizable

- e.g. MMX can be considered modest vector unit

Advantages of IRAM

- Higher bandwidth
- Lower latency
- Energy efficiency
- Memory size and width $\rightarrow$ free organization
- Board space (small)

Disadvantages of IRAM

- Area and speed of logic in a DRAM process technology
  - Area: 30% - 70% Larger, Speed: 30% - 70% Less
- Area and power impact of increasing bandwidth to DRAM core
- Retention time of DRAM when operating at high Temperature
  - retention rate halved for every 10° C
  - refresh rates must increase for high temperature
- Scaling system beyond single IRAM
- Matching IRAM to commodity focus of DRAM industry
-

Challenges to IRAM

1. Fabrication process is tough
   - DRAM fabrication technology is different than logic (microprocessor) technology
     - you get slower microprocessors
     - you need to complicate design to avoid noise of switching logic on memory array
     - refresh rates increase as temperature increases
2. Bounded amount of DRAM: soon 96 MBytes
   - OK for portable computers
   - not OK for workstations
     - what about multiple IRAMs?