High-Performance Architectures for Embedded Memory Systems

Christoforos E. Kozyrakis

Computer Science Division
University of California, Berkeley

kozyraki@cs.berkeley.edu
http://iram.cs.berkeley.edu/
Embedded DRAM systems roadmap

Previous talks

- Network Computer
- Video Games
- Graphics/Disks
- Camera/Phone

Desktop
Laptop

32 MB

8 MB

2 MB

This presentation
Outline

• Overview of general-purpose processors today
• Future processor applications & requirements
• Advantages and challenges of processor-DRAM integration
• Future microprocessor architectures
  – characteristics and features
  – compatibility and interaction with embedded DRAM technology
• Comparisons and conclusions
Current state-of-the-art processors (1)

- High performance processors
  - 64-bit operands, wide instruction issue (3-4)
  - dynamic scheduling, out-of-order execution, speculation
  - large multi-level caches
  - support for parallel systems
  - optimized for technical and/or commercial workloads
  - SIMD multimedia extensions (VIS, MAX, MMX, MDMX, Altivec)
  - 200 to 600 MHz, 20 to 80 Watts, 200 to 300 sq. mm
  - e.g. MIPS R10K, Pentium II, Alpha 21264, Sparc III
Current state-of-the-art processors (2)

- Embedded processors
  - 32/64-bit, single/dual issue, in-order execution
  - single-level (small) caches
  - code density improvements (Thump/MIPS16)
  - DSP/SIMD support
  - integrated I/O and memory controllers
  - some on-chip DRAM (up to 4MB)
  - optimized for low power, price/performance, MIPS/Watt
  - 50 to 250MHz, 0.3 to 4 Watts, 10 to 100 sq. mm
  - e.g. M32R, ARM-9, StrongARM, MIPS R5K, SH-4
Current microprocessor applications

- Desktop: technical workloads (e.g. CAD), office productivity tools
- Servers: file system workloads, transaction processing, decision support
- Embedded: variety of workloads, from printers to digital cameras
- Benchmarks:
  - desktop: SPEC95 (Int/FP)
  - servers: TPC C/D
  - embedded: Dhrystone
Future microprocessor applications

**Personal mobile computing**

- A single device is: PDA, video game, cell phone, pager, GPS, tape recorder, radio, TV remote…
- Basic interfaces: voice (speech recognition) and image (image/video processing)
- Small size, battery operated devices
- Media processing functions are the basic workload
Requirements on microprocessors

• High performance for multimedia:
  – real-time performance guarantees
  – support for continuous media data-types
  – fine-grain parallelism
  – coarse-grain parallelism
  – high-instruction reference locality, code density
  – high memory bandwidth

• Low power/energy consumption

• Low design/verification complexity, scalable design

• Small size/chip count
Embedded DRAM advantages (1)

• **High memory bandwidth**
  – make internal DRAM bandwidth available to processor
  – wide memory interfaces, custom organizations
  – multiple independent banks interconnected with processor through crossbar
  – how does it translate in performance?

• **Low memory latency**
  – no off-chip memory controller
  – no off-chip bus to arbitrate/drive
  – latency equal to inherent DRAM latency plus on-chip interconnect; still longer than SRAM
Embedded DRAM advantages (2)

• **Energy/power efficiency**
  – elimination of off-chip accesses through high capacitance bus
  – potential for lower power via on-demand memory module activation

• **System size benefits**
  – system-on-a-chip
  – no need for additional cache, external DRAM chips
  – potential for low pin count
Embedded DRAM challenges (1)

- The biggest worry: eDRAM cost
  - wafer cost
    - process steps compared to pure DRAM or logic processes
  - cost per DRAM bit
    - density of eDRAM compared to pure DRAM
  - yield
    - yield of DRAM part of die
    - yield of logic part of die
  - cost of testing
Embedded DRAM challenges (2)

- **Performance of logic**
  - traditional DRAM processes have slow logic transistors
  - potential solutions for eDRAM processes:
    - 2 types of transistors: fast for logic, high Vt for DRAM
    - additional layers of metal
    - cost of process steps?
  - still logic transistor speed may be lower than that of pure logic processes
    - deep pipelined designs
    - use architectures that do rely only on clock frequency for performance; utilize forms of parallelism
Embedded DRAM challenges (3)

- **Power consumption of logic**
  - directly affects temperature, refresh rate and DRAM yield
  - low power logic design
  - intelligent power management in hardware (e.g. clock gating) and software (dynamic voltage scaling)
  - dynamic control of refresh rate

- **Yield of logic component**
  - logic has lower yield than DRAM
  - employ redundancy in processor design?
  - already done for some cache designs
Embedded DRAM challenges (4)

- **Cost/complexity of testing**
  - manufacture testing of a chip with a processor and tens of Mbytes DRAM is expensive
  - processor can be used as build-in-self-test (BIST) engine
Embedded DRAM challenges (5)

- **Organization of on-chip DRAM**
  - width of interface: cache line or datapath width
  - hierarchical structures: multiple independent banks, organized in sub-bank sharing common bus
    - high random bandwidth
    - selective activation for lower power
  - memory crossbar instead of bus
  - caching for latency reduction
    - row buffers or virtual channels to keep more pages open
    - interaction with software?
  - optimum design point?
  - benefits vs. area overhead?
Trends in high-performance architecture

- Advanced superscalar processors
- VLIW: Very long instruction word processors (IA-64/EPIC)
- Single chip multiprocessors
- Reconfigurable processors
- Vector microprocessors (Vector IRAM)
Advanced superscalar processors

• Scale up current designs to issue more instructions (16-32)

• Major features:
  – dynamic instruction scheduling in hardware, out-of order execution
  – branch/dependence/stride/data/trace prediction buffers
  – large multibank caches
Advanced superscalar processors (2)

• Advantages
  – dynamic scheduling exploits run-time info
  – software compatibility
  – high-performance for current desktop applications

• Disadvantages
  – relies on high-speed logic and fast, large caches
  – unpredictable performance (high misprediction cost)
  – limited media processing support (MMX-like units)
  – high design/verification complexity
  – high power consumption due to extensive speculation

• eDRAM perspective
  – cannot fully utilize available eDRAM bandwidth
  – DRAM “unfriendly” environment (power, complexity, size)
  – DRAM for second-level cache?
VLIW processors (IA-64/EPIC)

- Very long instruction word scheme

- Major features
  - Instruction scheduling by compiler (dependence analysis, register renaming etc)
  - Template specifies if instructions can be executed in parallel
  - Software speculation and predicated (conditional) execution
  - Large number of registers
  - Multiple functional units
  - Cache based designs
VLIW processors (IA-64/EPIC) (2)

- Advantages
  - simpler hardware
  - highly scalable

- Disadvantages
  - code size (loop unrolling, software pipelining)
  - compiler performance?
  - software compatibility
  - limited media processing support (MMX-like units)

- eDRAM perspective
  - cannot fully utilize available eDRAM bandwidth
  - requires high-speed logic to make up for run-time information
  - DRAM for second-level cache?
Single chip multiprocessors

- Place multiple processors on a single chip

- Major features
  - symmetric multiprocessor system (shared memory system)
  - shared second-level cache
  - 4 to 8 uniprocessors, similar to current out-of-order designs
Single chip multiprocessors (2)

• Advantages
  – modular design
  – coarse-grained parallelism

• Disadvantages
  – difficulty of efficient parallel programming
  – limited media processing support
  – high power consumption
  – complexity of shared-memory protocols

• eDRAM perspective
  – can utilize bandwidth of multi-bank eDRAM
  – inherent redundancy
  – multiprocessors require large amount of memory
Reconfigurable Processors

- Use reconfigurable (programmable) logic, e.g. look-up tables

- Major features
  - meshes or hierarchical arrays of look-up tables
  - multiple configurations stored within the array
  - multiprocessor organizations with reconfigurable interconnects (RAW)
Reconfigurable Processors (2)

- Advantages
  - programmable/flexible functional unit(s)
  - any data width/function can be supported
  - fine and coarse grain parallelism

- Disadvantages
  - software path complexity (mapping high-level-languages to arrays, run-time environment)
  - power consumption and array size

- eDRAM perspective
  - on-chip DRAM for high-bandwidth data and configuration storage
  - array can be used as high-performance BIST engine
  - DRAM latency complicated programming/software tools
Vector microprocessors

- Vector instructions
  - \((v3[i]=v1[i]+v2[i], \text{for } i=1 \text{ to } N)\)
- Major features
  - vector coprocessor unit
  - instructions define operations on vectors (arrays) of data
  - vector register file
  - strided and indexed memory accesses
  - support for multiple data widths
  - support for DSP/fixed-point
  - conditional/speculative execution support through flag registers
Vector microprocessors (2)

• Advantages
  – predictable performance: in-order model, no caches
  – high performance for media processing
  – low power/energy consumption
  – performance through parallel pipelines, not just clock frequency
  – scalable
  – simple design: no complex issue/speculation logic
  – small code size: single instruction loops

• Disadvantages
  – cannot utilize random instruction-level or thread-level parallelism; just fine-grain parallelism
  – poor performance for many current desktop applications
  – requires high-bandwidth memory system
Vector processors and eDRAM

- Vector processors require multi-bank, high-bandwidth memory system:
  - multiple wide DRAM banks, crossbar interconnect
- Vector processors can tolerate DRAM latency
  - delayed vector pipelines
- eDRAM friendly environment
  - low power, low complexity, modest clock frequencies
- eDRAM testing
  - use vector processor as BIST engine; 10x faster than scalar processors
- Logic redundancy
  - use a redundant vector pipeline
Vector IRAM-1

- **Scalar core**
  - 2-way superscalar MIPS
  - 16KByte I/D caches
- **Vector coprocessor**
  - 64b, 32b, 16b data types
  - maximum vector length 32 @64b, 64@ 32b or 128 @ 16b
  - 2 arithmetic, 2 load/store, 2 flag processing units
  - 4 64bit pipelines per functional unit
  - separate multi-ported TLB

- **Memory system**
  - 16Mbytes DRAM
  - 8 independent banks
  - 256b synchronous interface
  - crossbar interconnect for 12.8GB/sec aggregate bandwidth per direction

- **I/O**
  - 4 serial lines, 1Gb/s per direction
  - fast messaging though network interface connected to memory system
Vector IRAM-1 Block Diagram
VIRAM-1 Technology Summary

Technology: 0.25 micron embedded DRAM-logic process
Memory: 16 MBytes
Die size: 350-400 mm²
Vector pipelines: 4 64-bit (or 8 32-bit or 16 16-bit)
Clock Frequency: 200MHz scalar, vector, DRAM
Serial I/O: 4 lines @ 1 Gbit/s
Power: ~2 W
Performance: 1.6 GFLOPS₆₄ – 6.4 GOPS₁₆
VIRAM-1 Floorplan
Comparison: current desktop domain

<table>
<thead>
<tr>
<th></th>
<th>SS</th>
<th>VLIW</th>
<th>CMP</th>
<th>RC</th>
<th>VIRAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPEC Int</td>
<td>+</td>
<td>+</td>
<td>=</td>
<td>=</td>
<td>–</td>
</tr>
<tr>
<td>SPEC FP</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>=</td>
<td>+</td>
</tr>
<tr>
<td>TPC (DB)</td>
<td>=</td>
<td>=</td>
<td>+</td>
<td>–</td>
<td>=</td>
</tr>
<tr>
<td>SW Effort</td>
<td>+</td>
<td>=</td>
<td>=</td>
<td>–</td>
<td>=</td>
</tr>
<tr>
<td>Design Scalability</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>=</td>
</tr>
<tr>
<td>Design Complexity</td>
<td>–</td>
<td>=</td>
<td>=</td>
<td>+</td>
<td>=</td>
</tr>
</tbody>
</table>

Legend: + positive, = neutral, – negative
## Comparison: personal mobile computing domain

<table>
<thead>
<tr>
<th></th>
<th>SS</th>
<th>VLIW CMP</th>
<th>RC</th>
<th>VIRAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real-time Perf.</td>
<td>–</td>
<td>=</td>
<td>=</td>
<td></td>
</tr>
<tr>
<td>Cont. Data Support</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td></td>
</tr>
<tr>
<td>Energy/power</td>
<td>–</td>
<td>=</td>
<td>=</td>
<td>–</td>
</tr>
<tr>
<td>Code Size</td>
<td>=</td>
<td>–</td>
<td>=</td>
<td></td>
</tr>
<tr>
<td>Fine-grain parall.</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>+</td>
</tr>
<tr>
<td>Coarse-grain parall.</td>
<td>=</td>
<td>=</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Memory BW</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td></td>
</tr>
<tr>
<td>Design Scalability</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td></td>
</tr>
<tr>
<td>Design Complexity</td>
<td>–</td>
<td>=</td>
<td>=</td>
<td>+</td>
</tr>
</tbody>
</table>
### Comparison: eDRAM perspective

<table>
<thead>
<tr>
<th></th>
<th>SS</th>
<th>VLIW CMP</th>
<th>RC</th>
<th>VIRAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>BW Utilization</td>
<td>=</td>
<td>=</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Latency Tolerance</td>
<td>−</td>
<td>=</td>
<td>=</td>
<td>−</td>
</tr>
<tr>
<td>Power Consumption</td>
<td>−</td>
<td>=</td>
<td>=</td>
<td>−</td>
</tr>
<tr>
<td>Need for Fast Logic</td>
<td>−</td>
<td>−</td>
<td>=</td>
<td>=</td>
</tr>
<tr>
<td>DRAM Testing</td>
<td>−</td>
<td>−</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Logic Redundancy</td>
<td>−</td>
<td>−</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Design Scalability</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>=</td>
</tr>
<tr>
<td>Design Complexity</td>
<td>−</td>
<td>=</td>
<td>=</td>
<td>+</td>
</tr>
</tbody>
</table>
Conclusions

- Unlikely that eDRAM will make it in the desktop high-performance microprocessors (at least for a while)
- Yet, microprocessor applications shifting from desktop domain to personal mobile computing domain
- eDRAM can be of significant benefit to future processor architectures for this environment
  - high bandwidth memory system
  - system-on-a-chip benefits
- Challenges of eDRAM environment can be met by architectures developed for the new computing model
- Cost of eDRAM based processors remains to be seen...
References (1/6)


Computer Architecture

High-performance Processors
• M. Choudhury et.al, “A 300MHz CMOS Microprocessor with Multi-Media Extensions”, Digest of Technical Papers, ISSCC, February 1997
References (2/6)

**Embedded Processors**

- J. Choquette, “Genesis microprocessor”, Hot Chips Conference Record, August 1998

**Embedded DRAM**

- D. Patterson et.al, “Intelligent RAMs”, Digest of Technical Papers, ISSCC, February 1997
References (3/6)

- K. Murakami et.al, “Parallel Processing RAM Chip with 256Mb DRAM and Quad Processors”, Digest of Technical Papers, ISSCC, February 1997
References (4/6)

- N. Bowman et.al, "Evaluation of Existing Architectures in IRAM Systems", Workshop on Mixing Logic and DRAM: Chips that Compute and Remember at ISCA '97, June 1997
- NEC Corp., “irtual Channel Memory Technology”, http://www.nec.com

Microprocessor Architecture Trends

- J. Crawford, J. Huck, “Motivations and Design Approach for the IA-64 64-Bit Instruction Set Architecture”, In the Proceedings of the Microprocessor Forum, October 1997
- Y.N. Patt et.al.,” One Billion Transistors, One Uniprocessor, One Chip”, IEEE Computer, 30(9):51-57, September 1997
References (5/6)


General-purpose architectures and media-processing