The Future of Microprocessors Embedded in Memory

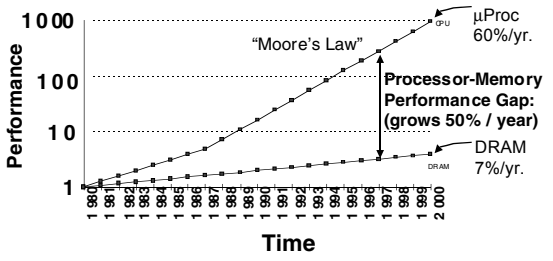
David A. Patterson

patterson@cs.berkeley.edu
http://cs.berkeley.edu/~patterson/talks
EECS, University of California
Berkeley, CA 94720-1776

Outline

- Desktop/Server Microprocessor State of the Art
- A New Technology: Embedded DRAM
- Mobile Multimedia Computing as New Direction
- A New Architecture for Mobile Multimedia Computing
- Berkeley's Mobile Multimedia Microprocessor
- Historical Perspective
- Challenges & Potential Industrial Impact

Processor-DRAM Gap (latency)



Processor-Memory Performance Gap "Tax"

Processor	% Area (≈cost)	%Transistors (≈power)
Alpha 21164	37%	77%
StrongArm SA110	61%	94%
Pentium Pro	64%	88%

- 2 dies per package: Proc/I$/D$ + L2$
- Caches have no inherent value, only try to close performance gap

Today's Situation: Microprocessor

MIPS MPUs	R5000	R10000	10k/5k
Clock Rate	200 MHz	195 MHz	1.0x
On-Chip Caches	32K/32K	32K/32K	1.0x
Instructions/Cycle	1(+ FP)	4	4.0x
Pipe stages	5	5-7	1.2x
Model	In-order	Out-of-order	---
Die Size (mm^2)	84	298	3.5x
– without cache, TLB	32	205	6.3x
Development (man yr.)	60	300	5.0x
SPECint_base95	5.7	8.8	1.6x

Desktop/Server State of the Art

- Processor performance doubling / 18 months
- Microprocessor-DRAM performance gap & tax
- Cost fixed at ≈$500/chip, power whatever can cool
  - 10X cost, 10X power => 2X performance?
- Desktop apps slow at rate processors speedup?
  - Word struggles to keep up with typing?
- Consolidation of industry?

PA-RISC PowerPC MIPS Alpha SPARC IA-64

Outline

- Desktop/Server Microprocessor State of the Art
- A New Technology: Embedded DRAM
- Mobile Multimedia Computing as New Direction
- A New Architecture for Mobile Multimedia Computing
- Berkeley's Mobile Multimedia Microprocessor
- Historical Perspective
- Challenges & Potential Industrial Impact

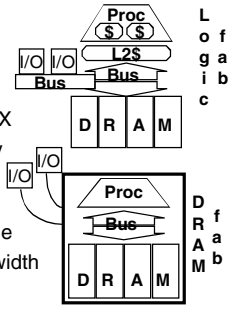
A More Revolutionary Approach: Processor Embedded in DRAM

- Faster logic in DRAM process
  - DRAM vendors offer faster transistors + same number metal layers as good logic process? @ ≈ 20% higher cost per wafer?
  - As die cost ≈ f(die area^4), 4% die shrink ⇒ equal cost
- Called Intelligent RAM ("IRAM") since most of transistors will be DRAM

IRAM Vision Statement

Microprocessor & DRAM on a single chip:

- on-chip memory latency 5-10X, bandwidth 50-100X
- improve energy efficiency 2X-4X (no off-chip bus)
- serial I/O 5-10X v. buses
- smaller board area/volume
- adjustable memory size/width



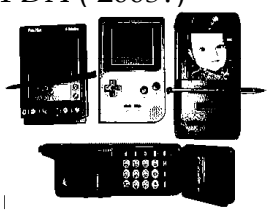
Outline

- Desktop/Server Microprocessor State of the Art
- A New Technology: Embedded DRAM
- Mobile Multimedia Computing as New Direction
- A New Architecture for Mobile Multimedia Computing
- Berkeley's Mobile Multimedia Microprocessor
- Historical Perspective
- Challenges & Potential Industrial Impact

Intelligent PDA (2003?)

Pilot PDA (todo,calendar, calculator, addresses,...)

+ Gameboy (Tetris, ...)
+ Nikon Coolpix (camera)
+ Cell Phone, Pager, GPS, tape recorder, TV remote, am/fm radio, garage door opener, ...
+ Wireless data (WWW)
+ Speech, vision recog.
+ Voice output for conversations



– Speech control of all devices
– Vision to see surroundings, scan documents, read bar codes, measure room

New Architecture Directions

- "…media processing will become the dominant force in computer arch. & microprocessor design."
- "... new media-rich applications... involve significant real-time processing of continuous media streams, and make heavy use of vectors of packed 8-, 16-, and 32-bit integer and Fl. Pt."
- Needs include high memory BW, high network BW, continuous media data types, real-time response, fine grain parallelism
  - "How Multimedia Workloads Will Change Processor Design", Diefendorff & Dubey, *IEEE Computer* (9/97)

Outline

- Desktop/Server Microprocessor State of the Art
- A New Technology: Embedded DRAM
- Mobile Multimedia Computing as New Direction
- A New Architecture for Mobile Multimedia Computing
- Berkeley's Mobile Multimedia Microprocessor
- Historical Perspective
- Challenges & Potential Industrial Impact

Potential IRAM Architecture

- "New" model: VSIW=Very Short Instruction Word!
  - Compact: Describe N operations with 1 short instruct.
  - Predictable (real-time) perf. vs. statistical perf. (cache)
  - Multimedia ready: choose N*64b, 2N*32b, 4N*16b
  - Easy to get high performance; N operations:
    - » are independent
    - » use same functional unit
    - » access disjoint registers
    - » access registers in same order as previous instructions
    - » access contiguous memory words or known pattern
    - » hides memory latency (and any other latency)
  - Compiler technology already developed, for sale!

Operation & Instruction Count: RISC v. "VSIW" Processor

(from F. Quintana, U. Barcelona.)

Spec92fp	Operations (M)			Instructions (M)		
Program	RISC	VSIW	R / V	RISC	VSIW	R / V
swim256	115	95	1.1x	115	0.8	142x
hydro2d	58	40	1.4x	58	0.8	71x
nasa7	69	41	1.7x	69	2.2	31x
su2cor	51	35	1.4x	51	1.8	29x
tomcatv	15	10	1.4x	15	1.3	11x
wave5	27	25	1.1x	27	7.2	4x
mdljdp2	32	52	0.6x	32	15.8	2x

VSIW reduces ops by 1.2X, instructions by 20X!

Revive Vector (= VSIW) Architecture!

- Cost: ≈ $1M each?
- Low latency, high BW memory system?
- Code density?
- Compilers?
- Vector Performance?
- Power/Energy?
- Scalar performance?
- Real-time?
- Limited to scientific applications?
- Single-chip CMOS MPU/IRAM
- IRAM = low latency, high bandwidth memory
- Much smaller than VLIW/EPIC
- For sale, mature (>20 years)
- Easy scale speed with technology
- Parallel to save energy, keep perf
- Include modern, modest CPU ⇒ OK scalar (MIPS 5K v. 10k)
- No caches, no speculation ⇒ repeatable speed as vary input
- Multimedia apps vectorizable too: N*64b, 2N*32b, 4N*16b

Software Technology Trends Affecting New Direction?

- any CPU + vector coprocessor/memory
  - scalar/vector interactions are limited, simple
  - Example architecture based on ARM 9, MIPS IV
- Vectorizing compilers built for 25 years
  - can buy one for new machine from The Portland Group
- Can change HW without changing programs
  - More arithmetic units, registers OK
- Lowers software effort
  - Media Processors/DSP, > 50% staff are programmers

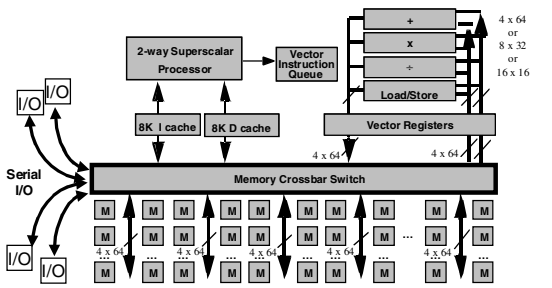
Software Technology Trends Affecting New Direction?

- IRAM has limited memory => Desktop OS wrong
- Real-Time Operating Systems
  - "Microkernel" - exclude modules not used by application
  - Small footprint even with all features (0.5 to 2 MB)
  - Examples: WRS VxWorks, QNX, Windows CE
- Applications not distributed in binary
  - WWW software distribution (Java, Javascript, TCL)
  - Open Source movement (Linix, Netscape, PERL)

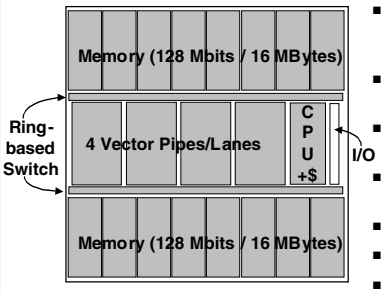
Outline

- Desktop/Server Microprocessor State of the Art
- A New Technology: Embedded DRAM
- Mobile Multimedia Computing as New Direction
- A New Architecture for Mobile Multimedia Computing
- Berkeley's Mobile Multimedia Microprocessor
- Historical Perspective
- Challenges & Potential Industrial Impact

V-IRAM1: 0.18 µm, Fast Logic, 200 MHz 1.6 GFLOPS(64b)/6.4 GOPS(16b)/32MB

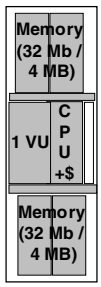


Tentative VIRAM-1 Floorplan



- 0.18 µm DRAM 32 MB in 16 banks x 256b, 128 subbanks
- 0.18 µm, 5 Metal Logic
- ≈ 200 MHz MIPS, 16K I$, 16K D$
- ≈ 4 200 MHz FP/int. vector units
- die: ≈ 16x16 mm
- xtors: ≈ 270M
- power: ≈2 Watts

Tentative VIRAM-"0.25" Floorplan



- Demonstrate scalability via 2nd layout (automatic from 1st)
- 8 MB in 4 banks x 256b, 32 subbanks
- ≈ 200 MHz CPU, 16K I$, 16K D$
- 1 ≈ 200 MHz FP/int. vector units
- die: ≈ 5 x 16 mm
- xtors: ≈ 70M
- power: ≈0.5 Watts

V-IRAM-1 Tentative Plan

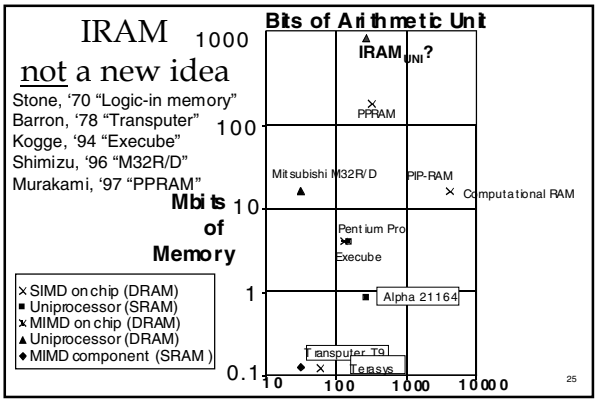
- Phase I: Feasibility stage (≈H1'98)
  - Test chip, CAD agreement, architecture defined
- Phase 2: Design & Layout Stage (≈H2'98)
  - Test chip, Simulated design and layout
- Phase 3: Verification (≈H2'99)
  - Tape-out
- Phase 4: Fabrication,Testing, and Demonstration (≈H1'00)
  - Functional integrated circuit
- First microprocessor ≥ 0.15-.25B transistors?

Outline

- Desktop/Server Microprocessor State of the Art
- A New Technology: Embedded DRAM
- Mobile Multimedia Computing as New Direction
- A New Architecture for Mobile Multimedia Computing
- Berkeley's Mobile Multimedia Microprocessor
- Historical Perspective
- Challenges & Potential Industrial Impact

IRAM not a new idea

Stone, '70 "Logic-in memory"
Barron, '78 "Transputer"
Kogge, '94 "Execube"
Shimizu, '96 "M32R/D"
Murakami, '97 "PPRAM"



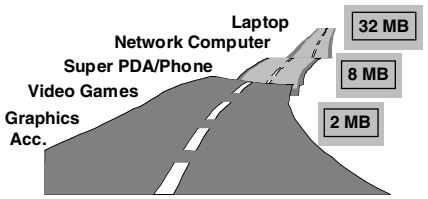
Why IRAM now? Lower risk than before

- DRAM manufacturers now willing to listen
  - Before not interested, so early IRAM = SRAM
- Faster Logic + DRAM available soon?
- Past efforts memory limited ⇒ multiple chips ⇒ 1st solve the unsolved (parallel processing)
  - Gigabit DRAM ⇒ ≈100 MB; OK for many apps?
- Systems headed to 2 chips: CPU + memory
- Embedded apps leverage energy efficiency, adjustable mem. capacity, smaller board area ⇒ "Post-PC Era": PDA >> desktop

IRAM Challenges

- Chip
  - Good performance and reasonable power?
  - Cost for Embedded DRAM process?
    - » Cost of 16 Mbit DRAM $1.78 in June 1998: How compete?
  - Speed in Embedded DRAM? (time behind logic fab?)
  - Testing time of IRAM vs DRAM vs microprocessor?
  - BW/Latency oriented DRAM tradeoffs?
- Architecture
  - How to turn high memory bandwidth into performance for real applications?
  - Extensible IRAM: Large program/data solution?

Commercial IRAM highway is governed by memory per IRAM?



Words to Remember

"...a strategic inflection point is a time in the life of a business when its fundamentals are about to change. ... Let's not mince words: A strategic inflection point can be deadly when unattended to. Companies that begin a decline as a result of its changes rarely recover their previous greatness."

– *Only the Paranoid Survive*, Andrew S. Grove, 1996

Conclusion

- Apps/metrics of future to design computer of future
- Mobile Multimedia >> "Stationary" Computers?
- IRAM potential in mem/IO BW, energy, board area; challenges in cost, power/performance, testing
- 10X-100X improvements based on technology shipping for 20 years (not JJ, photons, MEMS, ...)
- Potential: shift semiconductor balance of power?
  Who ships the most memory? Most microprocessors?

Interested in Participating?

- Looking for ideas of IRAM enabled apps, partners to transfer technology
- Contact us if you're interested:
  `http://iram.cs.berkeley.edu/`
  `email: patterson@cs.berkeley.edu`
- Thanks for advice/support: DARPA, California MICRO, Hitachi, Intel, LG Semicon, Microsoft, Neomagic, Sandcraft, SGI/Cray, Sun Microsystems, TI, TSMC

Backup Slides

(The following slides are used to help answer questions)

VIRAM-1 Specs/Goals

Technology	**0.18-0.20 micron, 5-6 metal layers, fast xtor**
Memory	**16-32 MB**
Die size	**≈ 200-300 mm²**
Vector pipes/lanes	**4 64-bit (or 8 32-bit or 16 16-bit)**
Serial I/O	**4 lines @ 1 Gbit/s**
$Power_{university}$	**≈2 w @ 1-1.5 volt logic**
$Clock_{university}$	