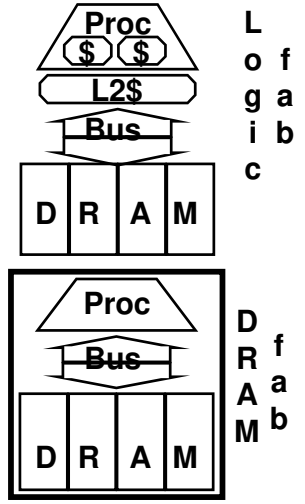
Intelligent RAM (IRAM): Chips that remember and compute

David Patterson, Thomas Anderson, Krste Asanovic,
Ben Gribstad, Neal Cardwell, Richard Fromm,
Jason Golbus, Kimberly Keeton,
Christoforos Kozyrakis, Stelianos Perissakis,
Randi Thomas, Noah Treuhaft,
John Wawrzynek, and Katherine Yelick
patterson@cs.berkeley.edu
http://iram.cs.berkeley.edu/
EECS, University of California
Berkeley, CA 94720-1776

IRAM Vision Statement

Microprocessor & DRAM on a single chip:

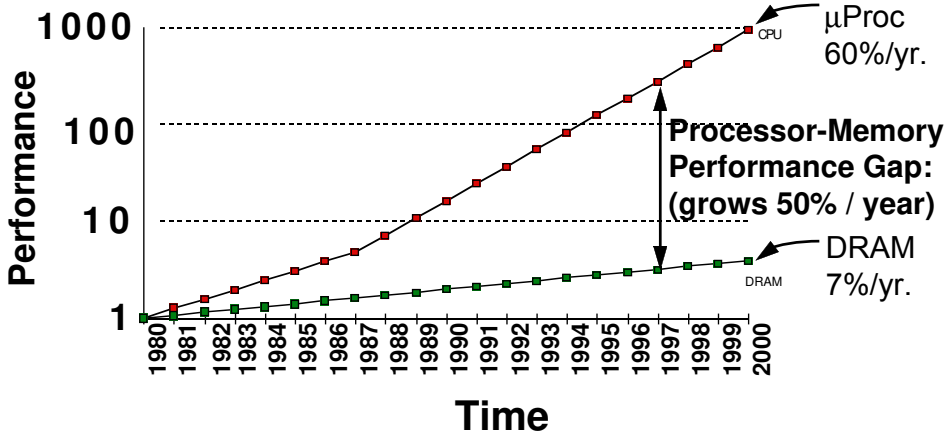
- bridge processor-memory performance gap via on-chip latency 5-10X,bandwidth 100X
- improve energy efficiency 2X-4X (no DRAM bus)
- adjustable memory size/width (designer picks any amount)
- smaller board area/volume



Outline

- Today's Situation: Microprocessor
- Today's Situation: DRAM
- IRAM Opportunities
- IRAM Architecture Options
- Applications of IRAM
- IRAM 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

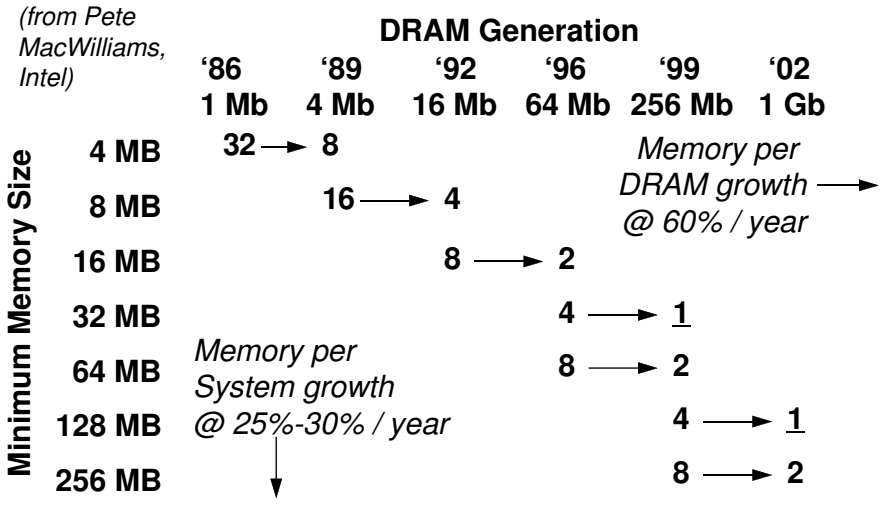
Today's Situation: Microprocessor

- Microprocessor-DRAM performance gap
  - time of a full cache miss in instructions executed
    1st Alpha (7000): 340 ns/5.0 ns = 68 clks x 2 or 136
    2nd Alpha (8400): 266 ns/3.3 ns = 80 clks x 4 or 320
    3rd Alpha (t.b.d.): 180 ns/1.7 ns =108 clks x 6 or 648
  - 1/2X latency x 3X clock rate x 3X Instr/clock ⇒ ≈5X
- Power limits performance (battery, cooling)
- Rely on caches to bridge gap

Today's Situation: DRAM

- Commodity, second source industry ⇒ high volume, low profit, conservative
  - Little organization innovation (vs. processors) in 20 years: page mode, EDO, Synch DRAM
- DRAM industry at a crossroads:
  - Price drop: 16Mb@$50 1/96 ⇒16Mb@$10 12/96
  - Fewer DRAMs per computer over time
    - » Growth bits/chip DRAM : 50%-60%/yr
    - » Nathan Myrvold M/S: mature software growth (33%/yr for NT) ≈ growth MB/$ of DRAM (25%-30%/yr)
  - Starting to question buying larger DRAMs?

Fewer DRAMs/System over Time

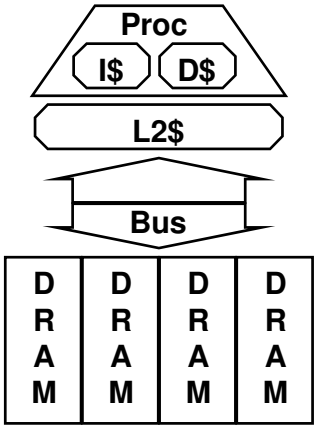


Reluctance for New DRAMs: Proc. v. DRAM BW, Min. Mem. size

- Processor DRAM bus BW = width x clock rate
  - Pentium Pro = 64b x 66 MHz ≈ 500 MB/sec
  - RISC = 256b x 66 MHz ≈ 2000 MB/sec
- DRAM bus BW = width x "clock rate"
  - EDO DRAM, 8b wide x 40 MHz = 40 MB/sec
  - Synch DRAM, 16b wide x 125 MHz = 250 MB/sec
- CPU BW / DRAM BW = 8 -16 chips minimum
  - 64Mb ⇒ 64-128 MB min. memory; 256Mb/Gb?
- Use old generation to lower total memory cost?

Reluctance for New DRAMs: DRAM BW ≠ App BW

- More App Bandwidth (BW)
  ⇒ Cache misses
  ⇒ DRAM RAS/CAS
- Application BW
  ⇒ Lower DRAM latency
- RAMBUS, Synch DRAM good BW but higher latency
- EDO DRAM, Synch DRAM < 5% performance in PCs
- New generation little benefit?



Multiple Motivations for IRAM

- Some apps: energy, board area, memory size
- Gap means performance limit is memory
- DRAM companies at crossroads?
  - Dramatic price drop since January 1996
  - Dwindling interest in future DRAM?
    - » Too much memory per chip?
    - » Need low latency as well as high bandwidth?
- Alternatives to IRAM: packaging breakthrough, more out-of-order CPU, fix capacity but shrink DRAM die, ...

Potential IRAM Latency: 5 - 10X

- No parallel DRAMs, memory controller, bus to turn around, SIMM module, pins…
- New focus: Latency oriented DRAM?
  - Dominant delay = RC of the word lines
  - keep wire length short & block sizes small?
- << 30 ns for 1024b IRAM "RAS/CAS"?
- AlphaSta. 600: 180 ns=128b, 270 ns= 512b Next generation (21264): 180 ns for 512b?

Potential IRAM Bandwidth: 100X

- 1024 1Mbit modules(1Gb), each 256b wide
  - 25% @ 25 ns RAS/CAS = 320 GBytes/sec
- If cross bar switch delivers 1/3 to 2/3 of BW of 25% of modules
  ⇒ 100 - 200 GBytes/sec
- FYI: AlphaServer 8400 = 1.2 GBytes/sec
  - 75 MHz, 256-bit memory bus, 4 banks

Potential Energy Efficiency: 2X-4X

- Case study of StrongARM memory hierarchy vs. IRAM memory hierarchy
  - cell size advantages ⇒ much larger cache
    ⇒ fewer off-chip references
    ⇒ up to 2X-4X energy efficiency for memory
  - less energy per bit access for DRAM
- Memory cell area ratio/process: P6, α '164,SArm
  cache/logic : SRAM/SRAM : DRAM/DRAM
  20-50 : 8-11 : 1

Potential Innovation in Standard DRAM Interfaces

- Optimizations when chip is a system vs. chip is a memory component
  - Improve yield with variable refresh rate?
  - "Map out" bad memory modules to improve yield?
  - Reduce test cases/testing time during manufacturing?
  - Lower power via on-demand memory module activation?
- IRAM advantages even greater if innovate inside DRAM memory interface?

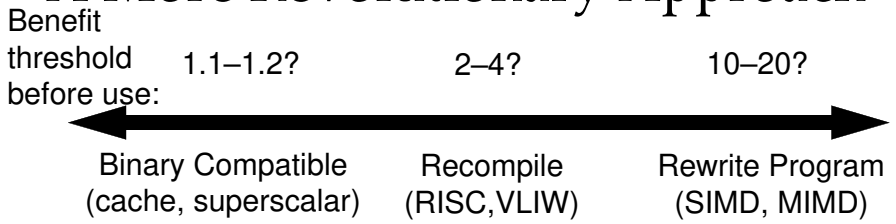
"Vanilla" Approach to IRAM

- Estimate performance IRAM version of Alpha (same caches, benchmarks, standard DRAM)
  - Used optimistic and pessimistic factors for logic (1.3-2.0 slower), SRAM (1.1-1.3 slower), DRAM speed (5X-10X faster) for standard DRAM
  - SPEC92 benchmark ⇒ 1.2 to 1.8 times slower
  - Database ⇒ 1.1 times slower to 1.1 times faster
  - Sparse matrix ⇒ 1.2 to 1.8 times faster
- Conventional architecture/benchmarks/DRAM not exciting performance; energy,board area only

A More Revolutionary Approach

- Faster logic in DRAM process
  - DRAM vendors offer same fast 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

A More Revolutionary Approach



- Find an architecture to exploit IRAM yet simple programming model so can deliver exciting cost/performance for many applications
  - Evolve software while changing underlying hardware
  - Simple ⇒ sequential (not parallel) program; large memory; uniform memory access time
  - More innovative than "Let's build a larger cache!"

Example IRAM Architecture Options

- (Massively) Parallel Processors (MPP) in IRAM
  - Hardware: best potential performance / transistor, but less memory per processor
  - Software: few successes in 30 years: databases, file servers, dense matrix computations, ... delivered MPP performance often disappoints
  - Will potential speedup justify rewriting programs?

Example IRAM Architecture Options

- "New" model: VSIW=Very Short Instruction Word!
  - Compact: Describe N operations with 1 short instruct.
  - Scalable: Binary compatible yet scale no. of registers
  - Easy to get high perforamce; N operations are:
    - » indepedent
    - » 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
  - Compiler technology already developed, for sale!

Isn't Vector (= VSIW) dead?

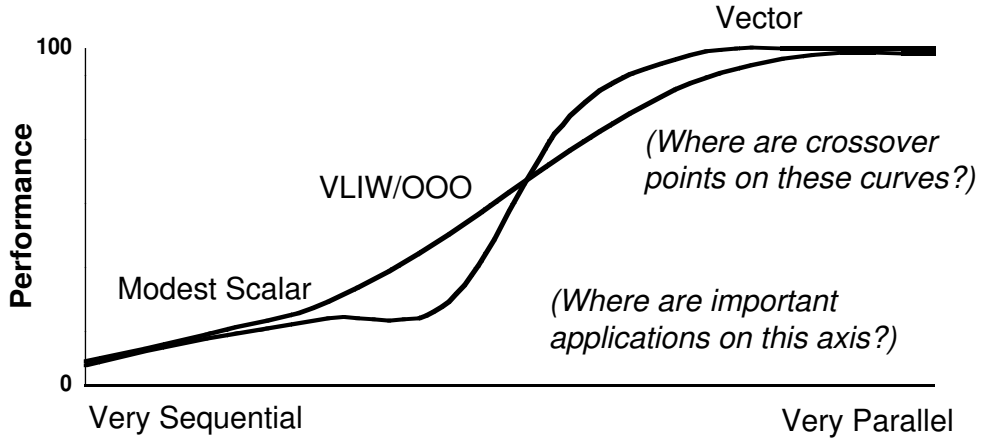
- High cost: ≈ $1M / processor?
- ≈5-10M transistors for vector processor?
- Low latency, high BW memory system?
- Energy?
- Poor scalar performance?
- Limited to scientific applications (2d fft)?

- Single-chip CMOS microprocessor/IRAM
- Small % in future + scales to 10B transistors
- IRAM = low latency, high bandwidth memory
- Fewer instructions/explicit control v. VLIW/OOO ⇒ power lower
- Include modern, modest CPU ⇒ scalar performs OK-good
- Multimedia apps (MMX) are vectorizable too

Simple v. Complex Case Study: MIPS R5000 v. R10000

	R5000	R10000	10k/5k
Clock Rate	180 MHz	180 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	---
SPECint_base95	4.8	7.9	1.6x
Die Size (mm^2)	84	298	3.5x

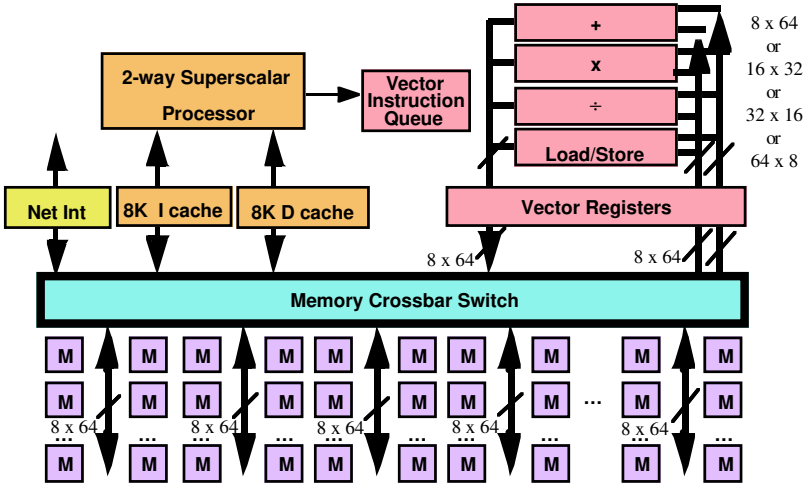
VLIW/OOO vs. Modest Scalar+Vector



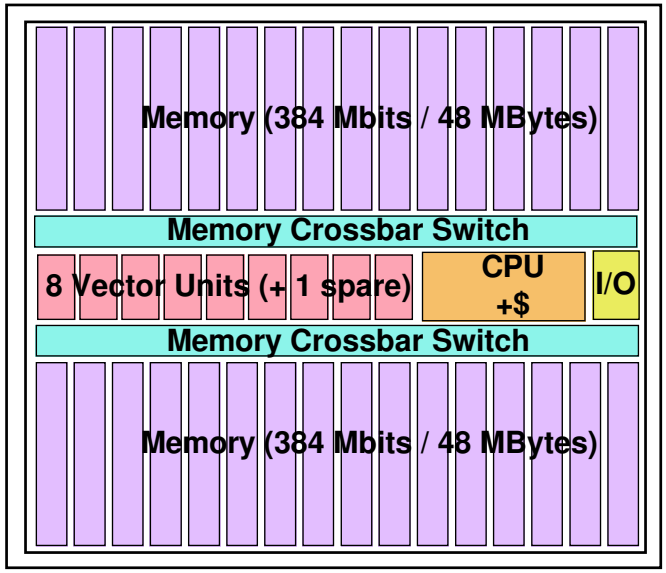
Software Technology Trends Affecting V-IRAM?

- V-IRAM: <u>any</u> CPU + vector coprocessor/memory
  - scalar/vector interactions are limited, simple
- Vectorizing compilers built for 25 years
  - can buy one for new machine from The Portland Group
- Microsoft "Win CE" for non-x86 platforms
- Library solutions for novel CPUs; retarget packages (e.g., MMX, Chromatics)
- Software distribution model is evolving?
  - New Model: Java byte codes over network?
    + Just-In-Time compiler to tailor program to machine?

V-IRAM-2: 0.18 µm, Fast Logic, 1GHz
16 GFLOPS(64b)/128 GOPS(8b)/96MB



V-IRAM-2 Floorplan



- 0.18 µm, 1 Gbit DRAM
- Die size = DRAM die
- 1B Xtors: 80% Memory, 4% Vector, 3% CPU ⇒ regular design
- Spare VU & Memory ⇒ ≈85% die repairable

How difficult to build and sell 1B transistor chip?

- Microprocessor only: ≈600 people, new CAD tools, what to build? (≈100% cache?)
- DRAM only: What is proper architecture/ interface? 1 Gbit with 16b RAMBUS interface? 1 Gbit with new package, new 512b interface?
- IRAM: highly regular design, target is not hard, can be done by a half-dozen Berkeley grad students?

Goal for Vector IRAM Generations

V-IRAM-1 (≈1999)	V-IRAM-2 (≈2002)
256 Mbit generation (0.25)	1 Gbit generation (0.18)
Die size = 256 Mb DRAM die	Die size = 1 Gb DRAM die
1.5 - 2.0 v logic, 0.5-2.0 watts	1.0 - 1.5 v logic, 0.5-2.0 w
300 - 500 MHz	500 - 1000 MHz
4 64-bit pipes/lanes	8 64-bit pipes/lanes
4 GFLOPS(64b)/32GOPS(8b)	16 GFLOPS/128GOPS
30 - 50 GB/sec Mem. BW	100 - 200 GB/sec Mem. BW
24 MB capacity + DRAM bus	96 MB cap. + DRAM bus
PCI bus/ FC-AL (serial SCSI)	Many Gbit Ethernet/FC-AL

IRAM Applications

- "Set-top Supercomputer" (<$200)
  - 4-chip Nintendo ⇒ 1-chip: 3D graphics, sound, fun!
- "Supercomputer on a AA battery" (<$300)
  - Super PDA/Smart Phone: speech I/O + "voice" email...
- Intelligent SIMM ("Smart SIMM")
  - 16 IRAMs + serial network + serial I/O into SIMM & put in standard memory system ⇒ Cluster/Network of IRAMs
  - Read/compare/write all memory in 1 ms
  - 4 to 8 Smart SIMMs in Gbit generation ≈ 1 TeraFLOPS
- Intelligent Disk ("IDISK") 2.5" disk + IRAM + net.
  - "shared nothing" Decision Support, Information Retreval

IRAM Cost

- Fallacy: IRAM must cost ≥ Intel chip in PC (≈ $250 to $750)
  - Lower cost package for IRAM:
    - » IRAM: 1 chip with ≈ 30-40 pins, 1-3 watts
    - » Intel Pentium II module (242 pins): 1 chip with ≈ 400 pins, + 512KB cache, graphics/memory controller = 43 watts
  - Cost of whole IRAM applications < $300
  - Mitsubishi M32R with 2MB memory < 2-3X memory
- Smaller footprint, lower power ⇒ IRAM cluster cost ≈ "DRAM cluster" (SIMM)

ISIMM/IDISK Example: Sort

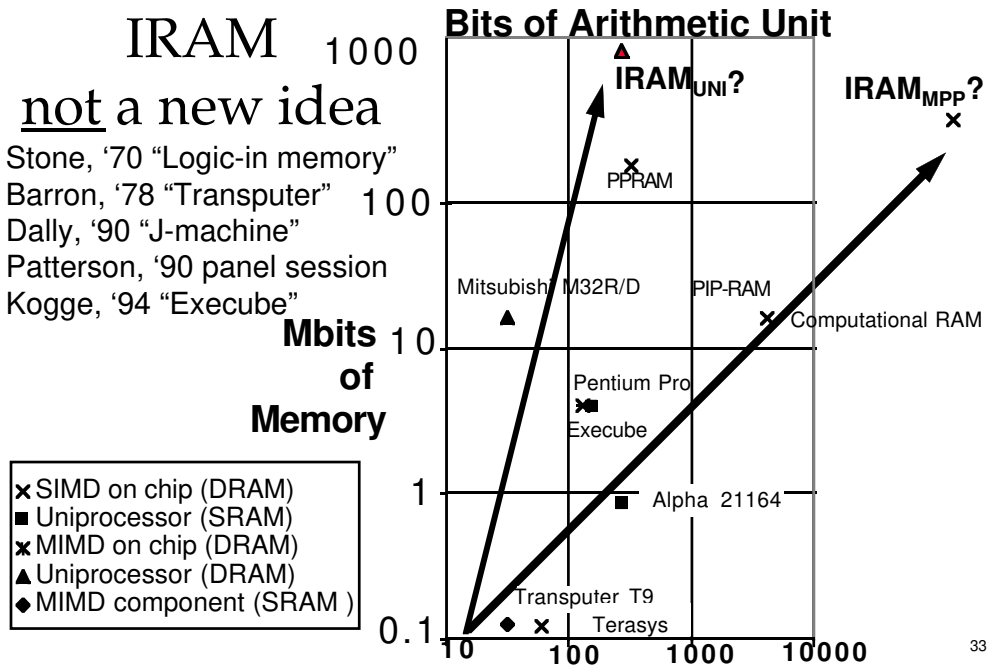
- Berkeley NOW cluster has world record sort: 8.6GB disk-to-disk using 95 processors in 1 minute
- Balanced system ratios for processor:memory:I/O
  - Processor: ≈ N MIPS
  - Large memory: N Mbit/s disk I/O & 2N Mb/s Network
  - Small memory: 2N Mbit/s disk I/O & 2N Mb/s Network
- Serial I/O at 2-4 GHz today (v. 0.1 GHz bus)
- IRAM: ≈ 2-4 GIPS + 2 2-4Gb/s I/O + 2 2-4Gb/s Net
- ISIMM: 16 IRAMs+net swtich+ FC-AL links (+disks)
- 1 IRAM sorts 9 GB, Smart Simm sorts 100 GB

Characterzing IRAM Performance

- Small memory on-chip (25 - 100 MB)
- High vector performance (4 -16 GFLOPS)
- Low latency main memory (20 - 30ns)
- High BW main memory (50 - 200 GB/sec)
- High BW I/O (0.5 - 2 GB/sec via N serial lines)
  - I/O must interact with processor (signal/interrupt), cache (consistency), main memory (bandwidth)
  - Integrated CPU/cache/memory with high memory BW ideal for fast serial I/O

IRAM not a new idea

Stone, '70 "Logic-in memory"
Barron, '78 "Transputer"
Dally, '90 "J-machine"
Patterson, '90 panel session
Kogge, '94 "Execube"



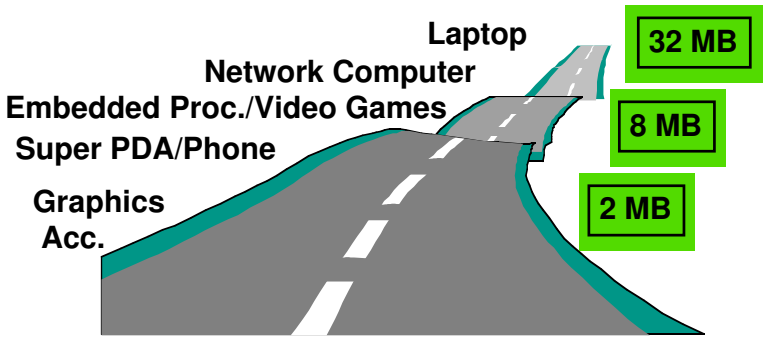
"Architectural Issues for the 1990s" (From Microprocessor Forum 10-10-90):

- **Given:**
  Superscalar, superpipelined RISCs and Amdahl's Law will not be repealed
  => High performance in 1990s is not limited by CPU
- **Predictions for 1990s:**
  "Either/Or" CPU/Memory will disappear *("hit under miss")*
  Multipronged attack on memory bottleneck
  cache conscious compilers
  lockup free caches / prefetching
  All programs will become I/O bound; design accordingly
  Most important CPU of 1990s is in DRAM: "IRAM"
  (Intelligent RAM: 64Mb + 0.3M transistor CPU = 100.5%)
  => CPUs are genuinely free with IRAM

Why IRAM now? Lower risk than before

- Faster Logic + DRAM available now/soon?
- DRAM manufacturers now willing to listen
  - Before not interested, so early IRAM = SRAM
- 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
  ⇒ OK market v. desktop (55M 32b RISC '96)

Commercial IRAM highway is governed by memory per IRAM?



IRAM Challenges

- Chip
  - Speed, area, power, yield, cost in DRAM process?
  - Good performance and reasonable power?
  - BW/Latency oriented DRAM tradeoffs?
  - Testing time of IRAM vs DRAM vs microprocessor?
  - Reconfigurable logic to make IRAM more generic?
- Architecture
  - How to turn high memory bandwidth into performance for real applications?
  - Extensible IRAM: Large program/data solution? (e.g., external DRAM, clusters, CC-NUMA, ...)

IRAM Conclusion

- IRAM potential in bandwidth (memory and I/O), latency, energy, capacity, board area; challenges in power/performance, testing, yield
  - cost-performance v. PC/Server: 2-4X perf. @ ≈0.1 cost
- V-IRAM can show potential (+compilers,+testing)
- 10X-100X improvements based on technology shipping for 20 years (not JJ, photons, MEMS, ...)
- Potential shift in balance of power in DRAM/ microprocessor industry in 5-7 years?

  Who ships the most memory?
  Who ships the most microprocessors?

Interested in Participating?

- Looking for industrial partners to help fab, (design?) test chips and prototype of V-IRAM-1
  - Fast, modern DRAM process
  - Existing RISC CPU core?
- Looking for partners with memory intensive apps
- Contact us if you're interested:
  `http://iram.cs.berkeley.edu/`
  `email: patterson@cs.berkeley.edu`
- Thanks for advice/support: DARPA, Intel, Neomagic, Samsung, SGI/Cray, Sun

Potential DRAM Partners (Chronological order of meetings)

- NEC
- Mitsubishi* (Fab in North Carolina)
- Hyndai Semiconductor
- LG Semiconductor* (Fab in Oregon)
- Samsung
- Alliance Semiconductor* (HQ in Silicon Valley)
- IBM* (Fab in Vermont)
- Micron* (Fab in Idaho)

  * DRAM Fab in US

Backup Slides

(The following slides are used to help answer questions)

Testing in DRAM

- Importance of testing over time
  - Testing time affects time to qualification of new DRAM, time to First Customer Ship
  - Goal is to get 10% of market by being one of the first companies to FCS with good yield
  - Testing 10% to 15% of cost of early DRAM
- Built In Self Test of memory:
  BIST v. External tester?
  Vector Processor 10X v. Scalar Processor?
- System v. component may reduce testing cost

How get Low Power, High Clock rate IRAM?

- Digital Strong ARM 110 (1996): 2.1M Xtors
  - 160 MHz @ 1.5 v = 184 "MIPS" < 0.5 W
  - 215 MHz @ 2.0 v = 245 "MIPS" < 1.0 W
- Start with Alpha 21064 @ 3.5v, 26 W
  - Vdd reduction ⇒ 5.3X ⇒ 4.9 W
  - Reduce functions ⇒ 3.0X ⇒ 1.6 W
  - Scale process ⇒ 2.0X ⇒ 0.8 W
  - Clock load ⇒ 1.3X ⇒ 0.6 W
  - Clock rate ⇒ 1.2X ⇒ 0.5 W
- 6/97: 233 MHz, 268 MIPS, 0.36W typ., $49

V-IRAM-1 Tentative Plan

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

Why a company should try IRAM

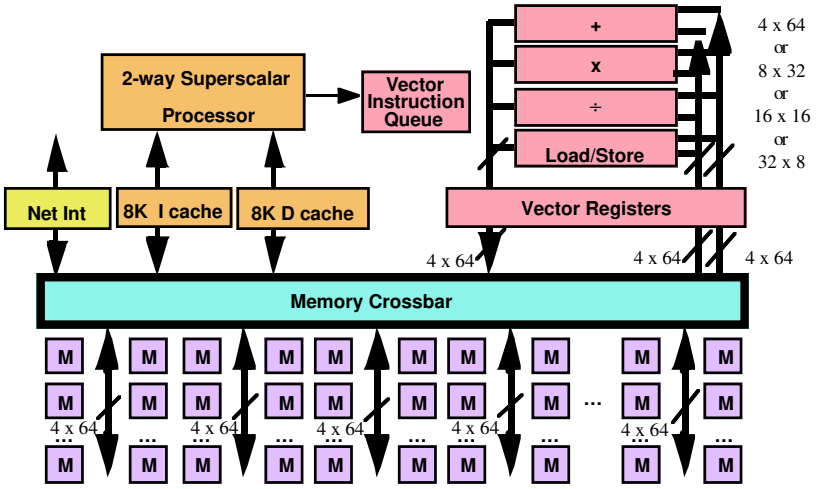
- If IRAM doesn't happen, then someday:
  - $10B fab for 16B Xtor MPU (too many gates per die)??
  - $12B fab for 16 Gbit DRAM (too many bits per die)??
- This is not rocket science. In 1997:
  - 20-50X improvement in memory density;
    ⇒ more memory per die or smaller die
  - 10X -100X improvement in memory performance
  - Regularity simplifies design/CAD/validate: 1B Xtors "easy"
  - Logic same speed
  - < 20% higher cost / wafer (but redundancy improves yield)
- IRAM success requires MPU expertise + DRAM fab

Words to Remember

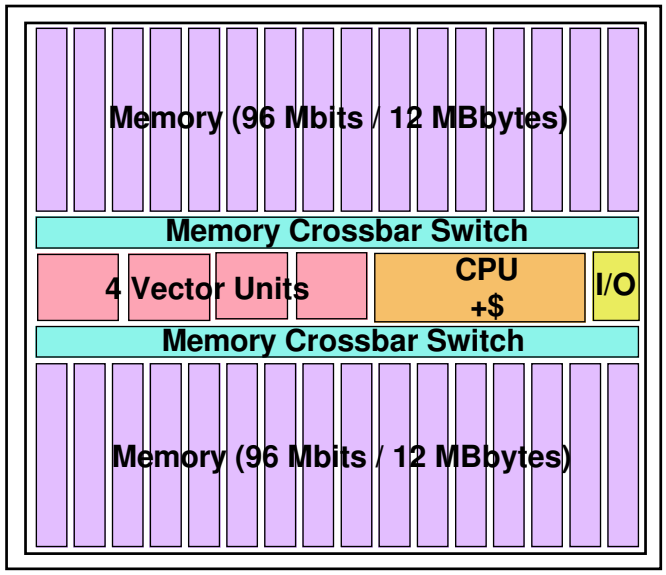
"...a strategic inflection point is a time in the life of a business when its fundamentals are about to change. ... Let's not mince words: A strategic inflection point can be deadly when unattended to. Companies that begin a decline as a result of its changes rarely recover their previous greatness."

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

V-IRAM-1: 0.25 µm, Fast Logic, 500 Mhz 4 GFLOPS(64b)/32 GOPS(8b)/24MB



V-IRAM-1 Floorplan



- 0.25 µm, 256 MbDRAM
- Die size = DRAM die
- 256M Xtors: 80% Memory, 8% Vector, 6% CPU ⇒ regular design

Energy to Access Memory by Level of Memory Hierarchy

- For 1 access, measured in nJoules

	Conventional	IRAM
on-chip L1$(SRAM)	0.5	0.5
on-chip L2$(SRAM v. DRAM)	2.4	1.6
L1 to Memory (off- v. on-chip)	98.5	4.6
L2 to Memory (off-chip)	316.0	*(n.a.)*

  - » Based on Digital StrongARM, 0.35 µm technology
  - » See "The Energy Efficiency of IRAM Architectures," *24th Int'l Symp. on Computer Architecture*, June 1997

21st Century Benchmarks?

- Potential Applications (new model highlighted)
  - **Text**: spelling checker (ispell), Java compilers (Javac, Espresso), content-based searching (Digital Library)
  - **Image**: text interpreter(Ghostscript), mpeg-encode, ray tracer (povray), Synthetic Aperture Radar (2D FFT)
  - **Multimedia**: Speech (Noway), Handwriting (HSFSYS)
  - **Simulations**: Digital circuit (DigSim),Mandelbrot (MAJE)
- Others? suggestions requested!
  - Encryption (pgp), Games?, Object Relational Database?, Word Proc?, Reality Simulation/Holodeck?,

Justification#2: Berkeley has done one "lap"; ready for new architecture?

- RISC: Instruction set /Processor design + Compilers (1980-84)
- SOAR/SPUR: Obj. Oriented SW, Caches, & Shared Memory Multiprocessors + OS kernel (1983-89)
- RAID: Disk I/O + File systems (1988-93)
- NOW: Networks + Clusters + Protocols (1993-98)
- IRAM: Instruction set, Processor design, Memory Hierarchy, I/O, Network, and Compilers/OS (1996-200?)