An Introduction to Intelligent RAM (IRAM)

David Patterson, Krste Asanovic, Aaron Brown, Ben Gribstad, Richard Fromm, Jason Golbus, Kimberly Keeton, Christoforos Kozyrakis, Stelianos Perissakis, Randi Thomas, Noah Treuhaft, Tom Anderson, John Wawrzynek, and Katherine Yelick

patterson@cs.berkeley.edu

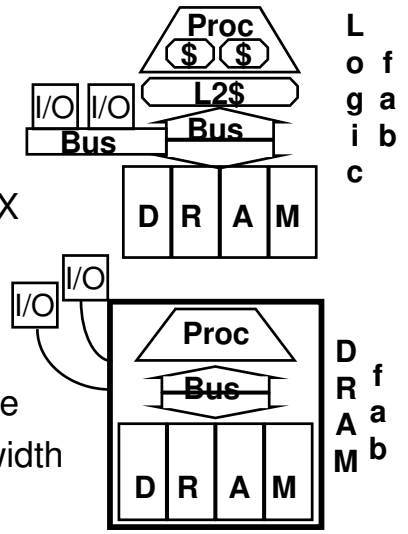
http://iram.cs.berkeley.edu/

EECS, University of California
Berkeley, CA 94720-1776

IRAM Vision Statement

Microprocessor & DRAM on a single chip:

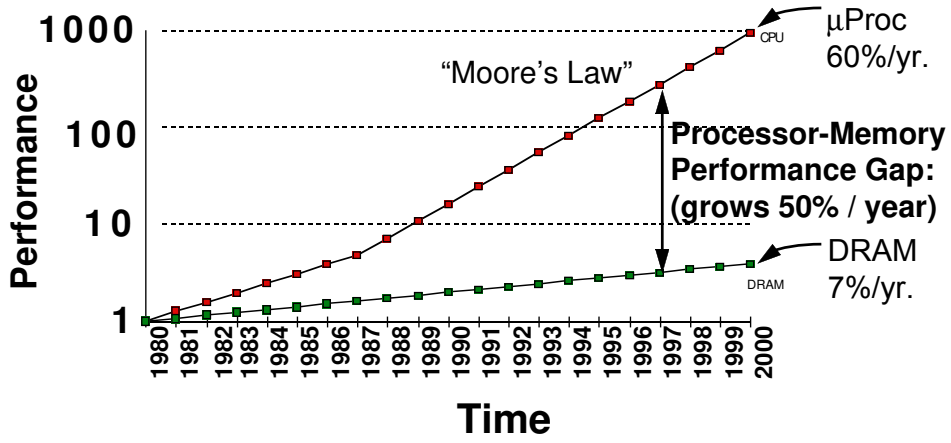
- on-chip memory latency 5-10X, bandwidth 50-100X
- improve energy efficiency 2X-4X (no off-chip bus)
- serial I/O 5-10X v. buses
- smaller board area/volume
- adjustable memory size/width



Outline

- Today's Situation: Microprocessor
- Today's Situation: DRAM
- IRAM Opportunities
- Applications of IRAM
- Directions for New Architectures
- Berkeley IRAM Project Plans
- Related Work and Why Now?
- IRAM Challenges & 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

Today's Situation: Microprocessor

■ Rely on caches to bridge gap

■ 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

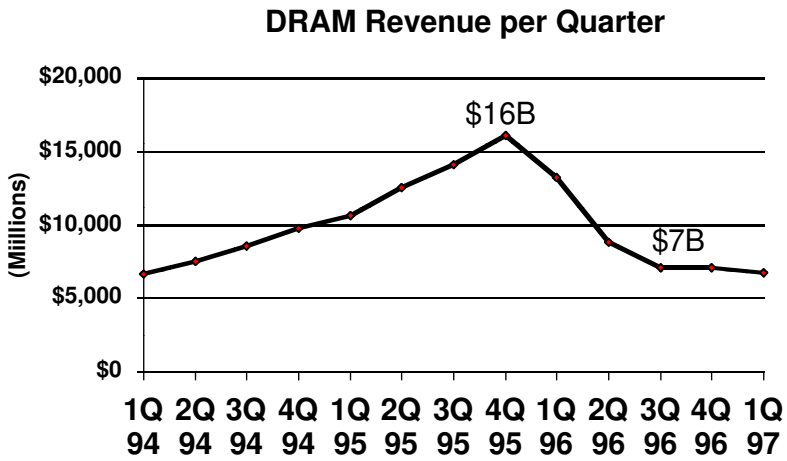
– 1/2X latency x 3X clock rate x 3X Instr/clock ⇒ ≈5X

■ Power limits performance (battery, cooling)

■ Shrinking number of desktop MPUs?

PA-RISC PowerPC MIPS Alpha SPARC IA-64

Today's Situation: DRAM

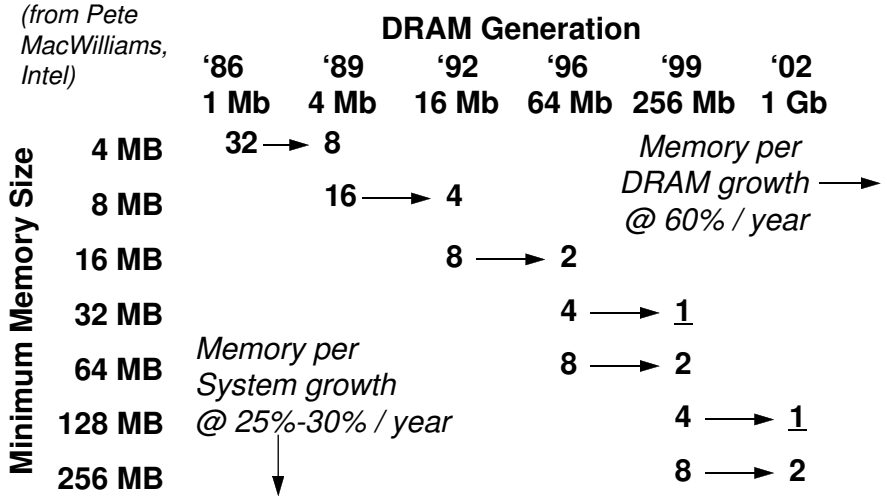


• Intel: 30%/year since 1987; 1/3 income profit

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:
  - 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



Multiple Motivations for IRAM

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

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?
- 10-30 ns for 64b-256b 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
  - 20% @ 20 ns RAS/CAS = 320 GBytes/sec
- If cross bar switch delivers 1/3 to 2/3 of BW of 20% of modules
  ⇒ 100 - 200 GBytes/sec
- FYI: AlphaServer 8400 = 1.2 GBytes/sec
  - 75 MHz, 256-bit memory bus, 4 banks

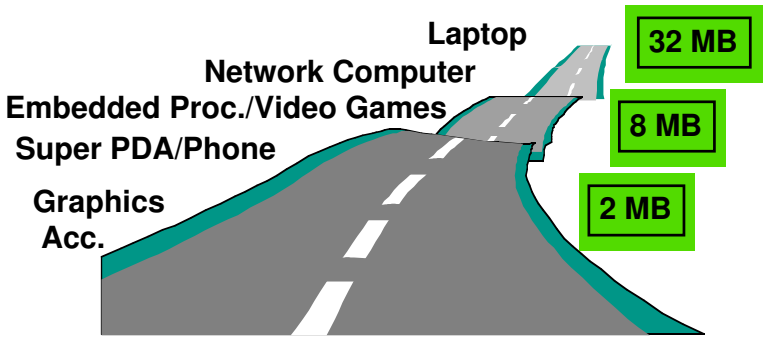
Potential Energy Efficiency: 2X-4X

- 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?

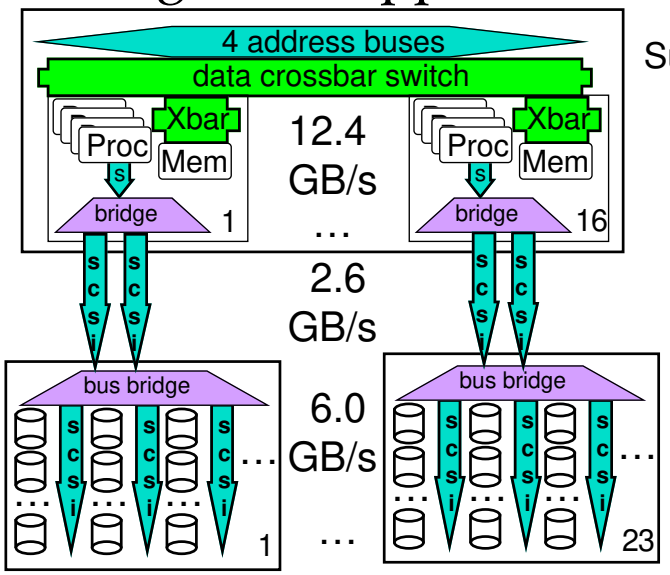
Commercial IRAM highway is governed by memory per IRAM?



Near-term IRAM Applications

- "Intelligent" Set-top
  - 2.6M Nintendo 64 (≈ $150) sold in 1st year
  - 4-chip Nintendo ⇒ 1-chip: 3D graphics, sound, fun!
- "Intelligent" Personal Digital Assistant
  - 1.0M PalmPilots (≈ $300) sold in 1st year:
  - Speech input vs. Learn new Alphabet (α = K, ℸ = T)
  - Camera/Vision for PDA to see surroundings
  - Speech output to converse
  - Play checkers with PDA

Long-term App: Decision Support?

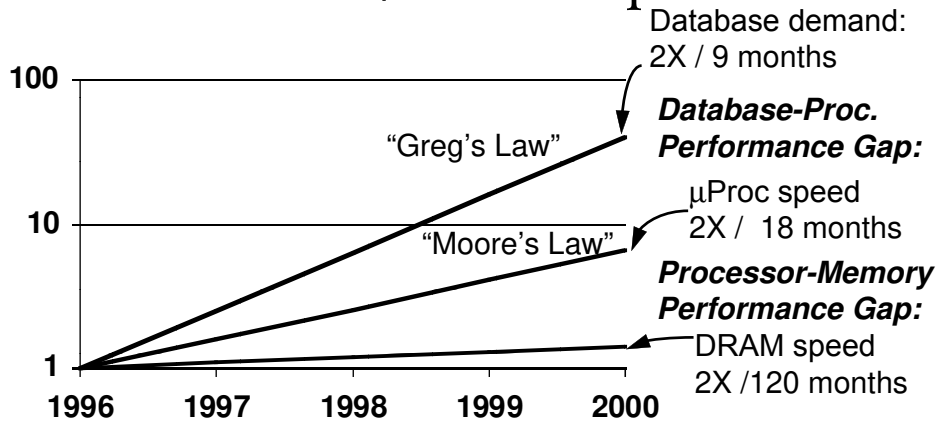


Sun 10000 (Oracle 8):

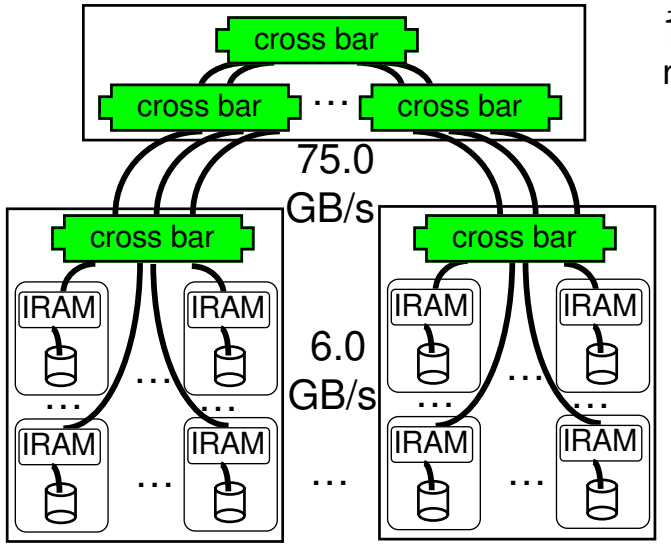
- TPC-D (1TB) leader
- SMP 64 CPUs, 64GB dram, 603 disks

Disks,encl.	$2,348k
DRAM	$2,328k
Boards,encl.	$983k
CPUs	$912k
Cables,I/O	$139k
Misc	$65k
HW total	$6,775k

IRAM Application Inspiration: Database Demand vs. Processor/DRAM speed



"Intelligent Disk": Scalable Decision Support?



1 IRAM/disk + shared nothing database

- 603 CPUs, 14GB dram, 603 disks

Disks (market)	$840k
IRAM (@$150)	$90k
Disk encl., racks	$150k
Switches/cables	$150k
Misc	$60k
Subtotal	$1,300k
Markup 2X?	≈ $2,600k

≈1/3 price, 2X-5X perf

"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: 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

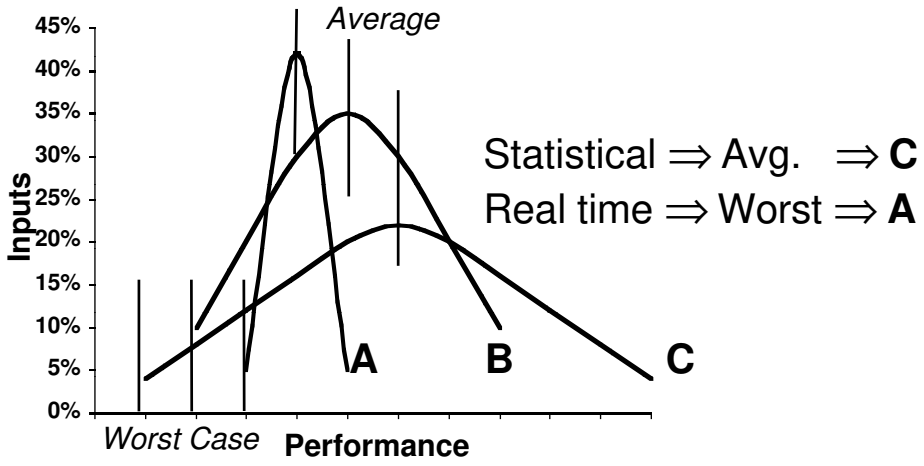
A More Revolutionary Approach: New Architecture Directions

- "...wires are not keeping pace with scaling of other features. … In fact, for CMOS processes below 0.25 micron ... *an unacceptably small percentage of the die will be reachable during a single clock cycle*."
- "Architectures that require long-distance, rapid interaction will not scale well ..."
  - "Will Physical Scalability Sabotage Performance Gains?" Matzke, *IEEE Computer* (9/97)

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)

Which is Faster? Statistical v. Real time Performance



Statistical ⇒ Avg. ⇒ **C**

Real time ⇒ Worst ⇒ **A**

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,8N*8b
  - Easy to get high performance; N operations:
    - » are independent (⇒ short signal distance)
    - » 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!

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,8N*8b

Mediaprocesing Functions (Dubey)

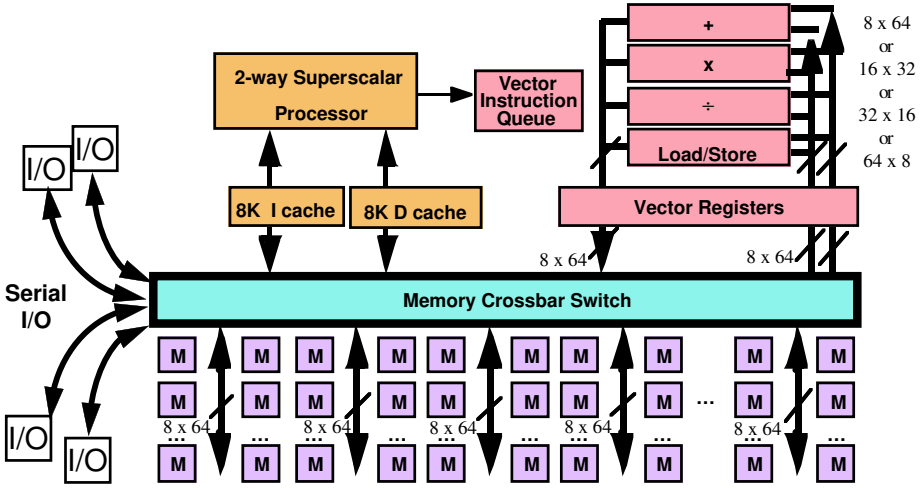
Kernel	Vector length
Matrix transpose/multiply	# vertices at once
DCT (video, comm.)	image width
FFT (audio)	256-1024
Motion estimation (video)	image width, i.w./16
Gamma correction (video)	image width
Haar transform (media mining)	image width
Median filter (image process.)	image width
Separable convolution ("")	image width

(from http://www.research.ibm.com/people/p/pradeep/tutor.html)

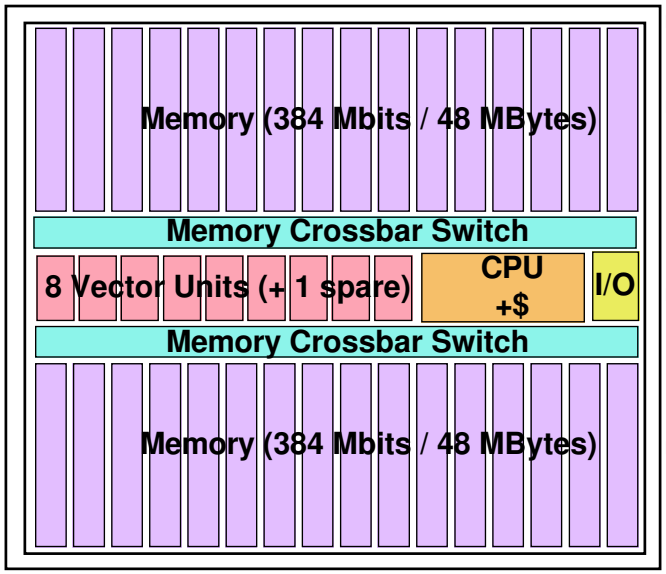
Software Technology Trends Affecting V-IRAM?

- V-IRAM: any CPU + vector coprocessor/memory
  - scalar/vector interactions are limited, simple
  - Example V-IRAM architecture based on ARM 9
- Vectorizing compilers built for 25 years
  - can buy one for new machine from The Portland Group
- Microsoft "Win CE"/ Java OS for non-x86 platforms
- Library solutions (e.g., MMX); retarget packages
- Software distribution model is evolving?
  - New Model: Java byte codes over network?
    + Just-In-Time compiler to tailor program to machine?

V-IRAM-2: 0.13 µm, Fast Logic, 1GHz 16 GFLOPS(64b)/128 GOPS(8b)/96MB

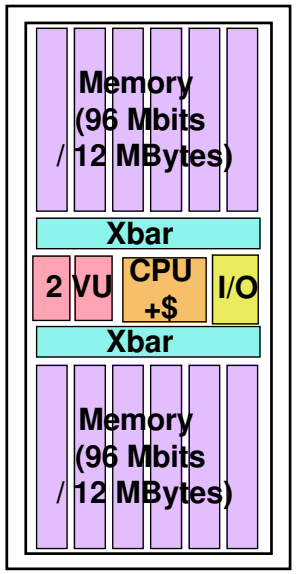


V-IRAM-2 Floorplan



- 0.13 µm, 1 Gbit DRAM
- 1B Xtors: 90% Memory, Xbar, Vector ⇒ regular design
- Spare VU & Memory ⇒ 90% die repairable
- Short signal distance ⇒ speed scales <0.1 µm

Alternative Goal: Low Cost V-IRAM-2



- Scalable design, 0.13 generation
- Reduce die size by 4X by shrinking vector units (25%), caches (25%), memory (25%)
- ≈50 mm², 16-24MB
- High Perf. version: 2.5 w, 1000 MHz, 4 - 32 GOPS
- Low Power version: 0.5 w, 500 MHz, 2 - 16 GOPS

V-IRAM-1 Specs/Goals

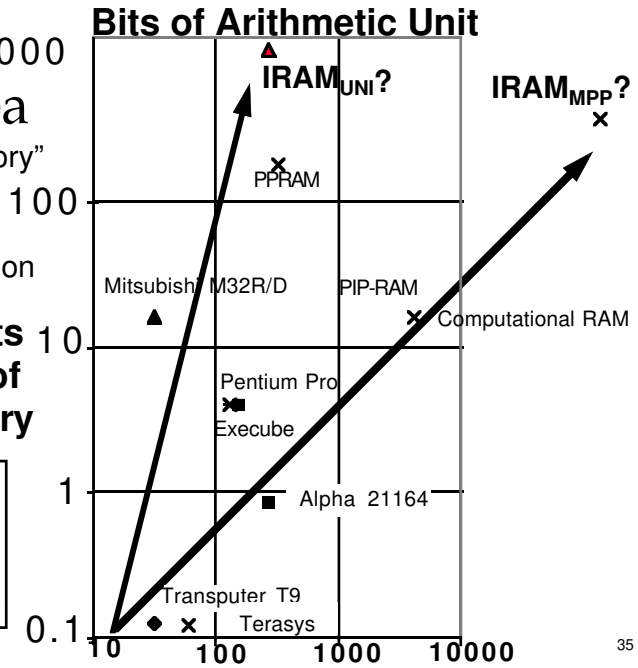
Technology	0.18-0.20 micron, 5-6 metal layers, fast xtor	
Die size	≈200 mm²	
Memory	16-24 MB	
Vector lanes	4 64-bit (or 8 32-bit or 16 16-bit or 32 8-bit)	
Target	Low Power	High Performance
Serial I/O	4 lines @ 1 Gbit/s	8 lines @ 2 Gbit/s
Power	≈2 w @ 1-1.5 volt logic	≈10 w @ 1.5-2 volt logic
$Clock_{univers.}$	200scalar/100vector MHz	250sc/250vector MHz
$Perf_{university}$	0.8 $GFLOPS_{64}$-6 $GFLOPS_8$	2 $GFLOPS_{64}$-16 $GFLOPS_8$
$Clock_{industry}$	400scalar/200vector MHz	500s/500v MHz
$Perf_{industry}$	1.6 $GFLOPS_{64}$-12 $GFLOPS_8$	4 $GFLOPS_{64}$-32 $GFLOPS_8$

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
- First microprocessor ≥ 100M transitors!

IRAM not a new idea

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



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)

IRAM Challenges

- Chip
  - Good performance and reasonable power?
  - Speed, area, power, yield, cost in DRAM process?
  - Testing time of IRAM vs DRAM vs microprocessor?
  - BW/Latency oriented DRAM tradeoffs?
  - 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 mem/IO BW, energy, board area; challenges in power/performance, testing, yield
- 10X-100X improvements based on technology shipping for 20 years (not JJ, photons, MEMS, ...)
- Apps/metrics of future to design computer of future
- V-IRAM can show IRAM's potential
  - multimedia, energy, size, scaling, code size, compilers
- Revolution in computer implementation v. Instr Set
  - Potential Impact #1: turn server industry inside-out?
- Potential #2: shift semiconductor balance of power?

  Who ships the most memory? Most microprocessors?

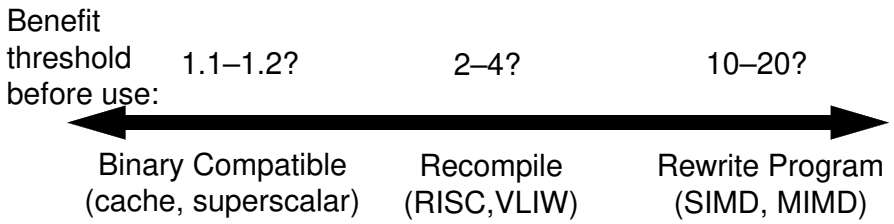
Interested in Participating?

- Looking for ideas of IRAM enabled apps
- Contact us if you're interested:
  `http://iram.cs.berkeley.edu/`
  `email: patterson@cs.berkeley.edu`
- Thanks for advice/support: DARPA, ARM, Intel, LG Semiconductor, Neomagic, Samsung, SGI/Cray, Sun Microsystems

Backup Slides

(The following slides are used to help answer questions)

New Architecture Directions

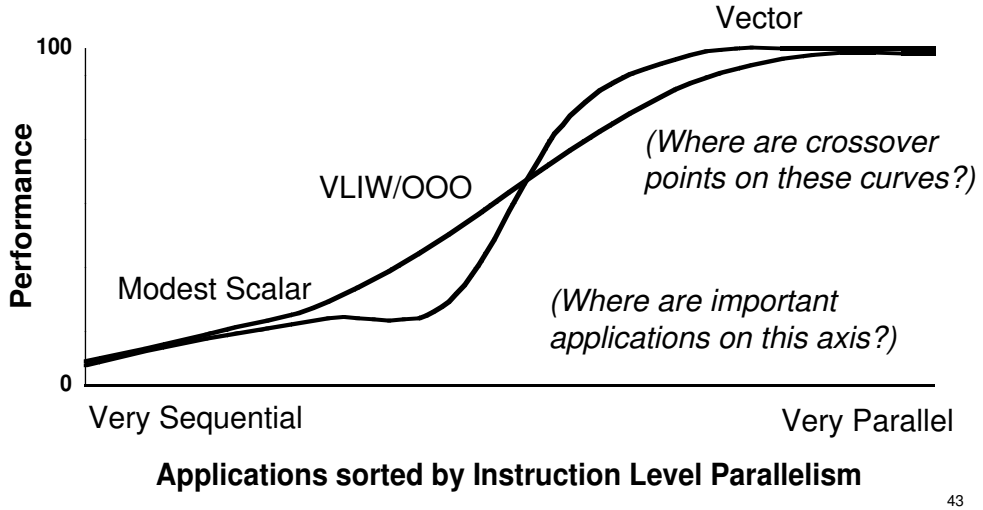


- More innovative than "Let's build a larger cache!"
- IRAM architecture with simple programming to deliver cost/performance for many applications
  - Evolve software while changing underlying hardware
  - Simple ⇒ sequential (not parallel) program; large memory; uniform memory access time

Grading Architecture Options

	Superscalar++	µSMP	VIRAM
Fine grain parallelism	A	A	A
Coarse grain (n chips)	A	B	A
Compiler maturity	B	B	A
MIPS/xtor (cost)	C	B	A
Technology scaling	C	A	A
Real time performance	C	B	A
Energy efficiency	D	A	A
Programmer model	D	B	A
"GPA"	C	B	A

VLIW/Out-of-Order vs. Modest Scalar+Vector



How to 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

Characterizing IRAM Cost/Performance

- Cost ≈ embedded processor + memory
- Small memory on-chip (25 - 100 MB)
- High vector performance (2 -16 GFLOPS)
- High multimedia performance (4 - 64 GOPS)
- Low latency main memory (15 - 30ns)
- High BW main memory (50 - 200 GB/sec)
- High BW I/O (0.5 - 2 GB/sec via N serial lines)
  - Integrated CPU/cache/memory with high memory BW ideal for fast serial I/O

Goal for Vector IRAM Generations

V-IRAM-1 (≈2000)	V-IRAM-2 (≈2003)
256 Mbit generation (0.20)	1 Gbit generation (0.13)
Die size = 256 Mb DRAM die	Die size = 1 Gb DRAM die
1.5 - 2.0 v logic, 2-10 watts	1.0 - 1.5 v logic, 2-10 watts
100 - 500 MHz	200 - 1000 MHz
4 64-bit pipes/lanes	8 64-bit pipes/lanes
1-4 GFLOPS(64b)/6-32G (8b)	2-16 GFLOPS/24-128G
30 - 50 GB/sec Mem. BW	100 - 200 GB/sec Mem. BW
24 MB capacity + DRAM bus	96 MB cap. + DRAM bus
Several fast serial I/O	Many fast serial I/O

"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

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
  - Successes are in servers, which need more memory than found in IRAM
  - How get 10X-20X benefit with 4 processors?
  - Will potential speedup justify rewriting programs?

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 dozen Berkeley grad students?

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, 1-5 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 < 2-4X memory
- Smaller footprint, lower power ⇒ IRAM cluster cost ≈ "DRAM cluster" (SIMM)

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

DRAM v. Desktop Microprocessors

Standards	pinout, package, refresh rate, capacity, ...	binary compatibility, IEEE 754, I/O bus
Sources	Multiple	Single
Figures of Merit	1) capacity, 1a) $/bit 2) BW, 3) latency	1) SPEC speed 2) cost
Improve Rate/year	1) 60%, 1a) 25%, 2) 20%, 3) 7%	1) 60%, 2) little change

DRAM Design Goals

- Reduce cell size 2.5, increase die size 1.5
- Sell 10% of a single DRAM generation
  - 6.25 billion DRAMs sold in 1996
- 3 phases: engineering samples, first customer ship(FCS), mass production
  - Fastest to FCS, mass production wins share
- Die size, testing time, yield => profit
  - Yield >> 60%
    (redundant rows/columns to repair flaws)

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 switch+ FC-AL links (+disks)
- 1 IRAM sorts 9 GB, Smart SIMM sorts 100 GB

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?)

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