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

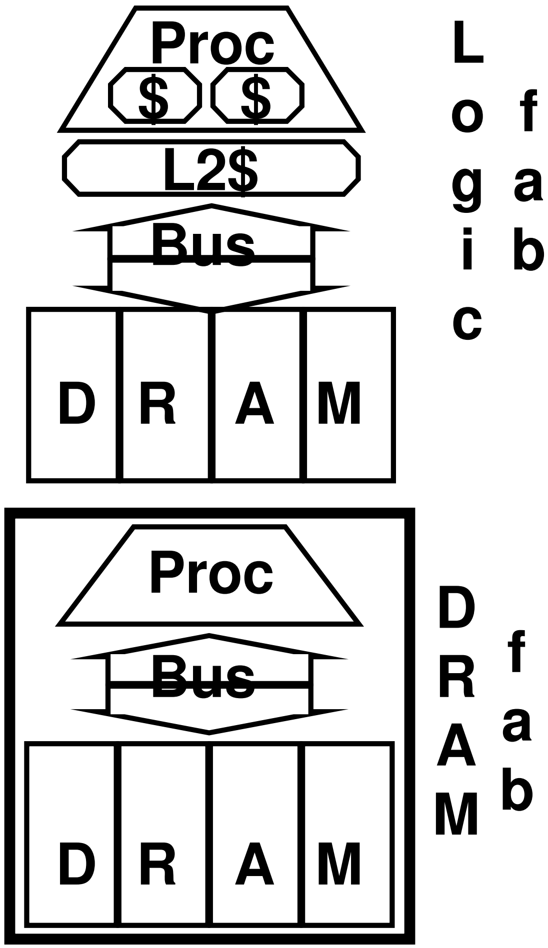
http://iram.cs.berkeley.edu/

EECS, University of California
Berkeley, CA 94720-1776

IRAM Vision Statement

Microprocessor & DRAM on a single chip:

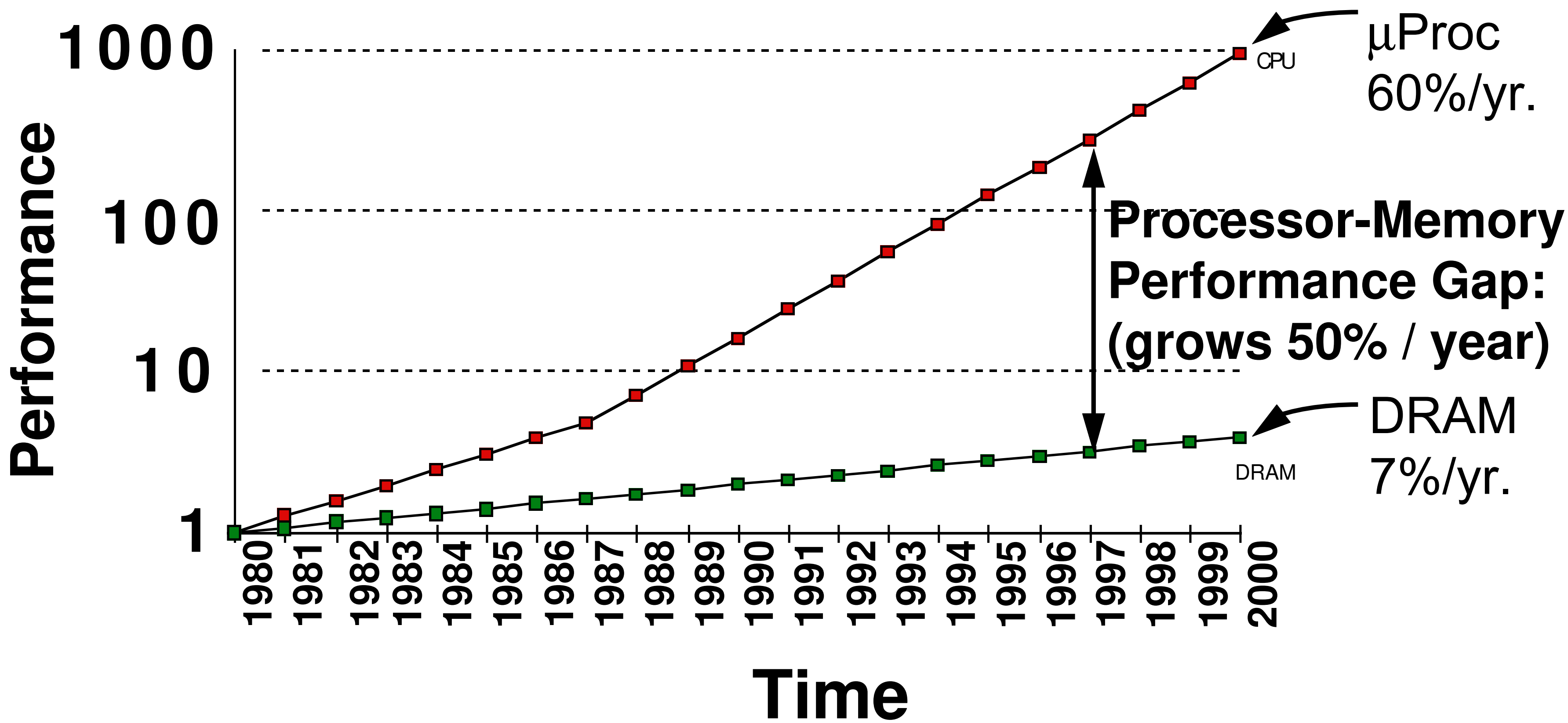
- bridge processor-memory performance gap via on-chip latency 5-10X,bandwidth 100X
- improve energy efficiency 2X-4X (no DRAM bus)
- adjustable memory size/width (designer picks any amount)
- smaller board area/volume



Outline

- Today's Situation: Microprocessor
- Today's Situation: DRAM
- IRAM Opportunities
- IRAM Architecture Options
- 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 $\Rightarrow \approx$5X
- Power limits performance (battery, cooling)
- Rely on caches to bridge gap
  - Doesn't work well for a few apps: data bases, …

Today's Situation: DRAM

- Commodity, second source industry
  $\Rightarrow$ 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
  - Starting to question buying larger DRAMs?

Fewer DRAMs/System over Time

(from Pete MacWilliams, Intel)

DRAM Generation

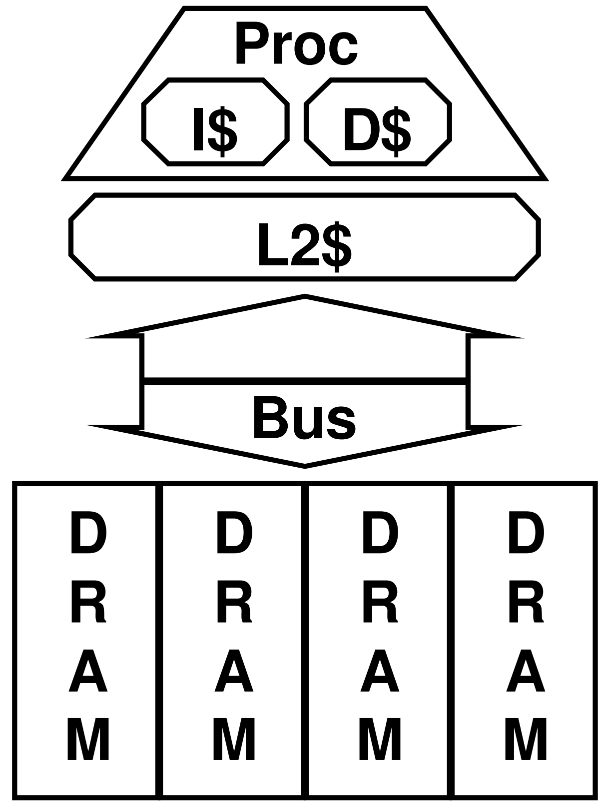
Minimum Memory Size	**'86 1 Mb**	**'89 4 Mb**	**'92 16 Mb**	**'96 64 Mb**	**'99 256 Mb**	**'02 1 Gb**
4 MB	32 →	8				
8 MB		16 →	4			
16 MB			8 →	2		
32 MB				4 →	1	
64 MB				8 →	2	
128 MB					4 →	1
256 MB					8 →	2

Memory per DRAM growth @ 60% / year →

Memory per System growth @ 25% / year ↓

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 increase BW but higher latency
- EDO DRAM, Synch DRAM < 5% performance in PCs



Multiple Motivations for IRAM

- Some apps: energy, board area, memory size
- Gap means performance limit is memory
- Dwindling interest in future DRAM: 256Mb/1Gb?
  - Too much memory per chip?
  - Industry supplies higher bandwidth at <u>higher</u> latency, but computers need lower latency
- Alternatives: 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, each 1Kb wide(1Gb)
  - 10% @ 40 ns RAS/CAS = 320 GBytes/sec
- If cross bar switch or multiple busses deliver 1/3 to 2/3 of total 10% of modules
  $\Rightarrow$ 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:21164,SA 110

cache/logic	SRAM/SRAM	DRAM/DRAM
25-50	10	1

Potential Innovation in Standard DRAM Interfaces

- Optimizations when chip is a system vs. chip is a memory component
  - Lower power with more selective module activation?
  - Lower voltage if all signals on chip?
  - Improved yield with variable refresh rate?
- IRAM advantages even greater if innovate inside DRAM memory modules?

"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 $\Rightarrow$ 1.2 to 1.8 times slower
  – Database $\Rightarrow$ 1.1 times slower to 1.1 times faster
  – Sparse matrix $\Rightarrow$ 1.2 to 1.8 times faster
- Conventional architecture/benchmarks/DRAM <u>not</u> 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
- 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

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
- Vector architecture in IRAM: More promising?
  - Simple model: seq. program, uniform mem. access
  - Multimedia apps (MMX) broaden vector relevance
  - Can tradeoff more hardware for slower clock rate
  - Cray on a chip:vector processor+interleaved memory

Why Vector? Isn't it dead?

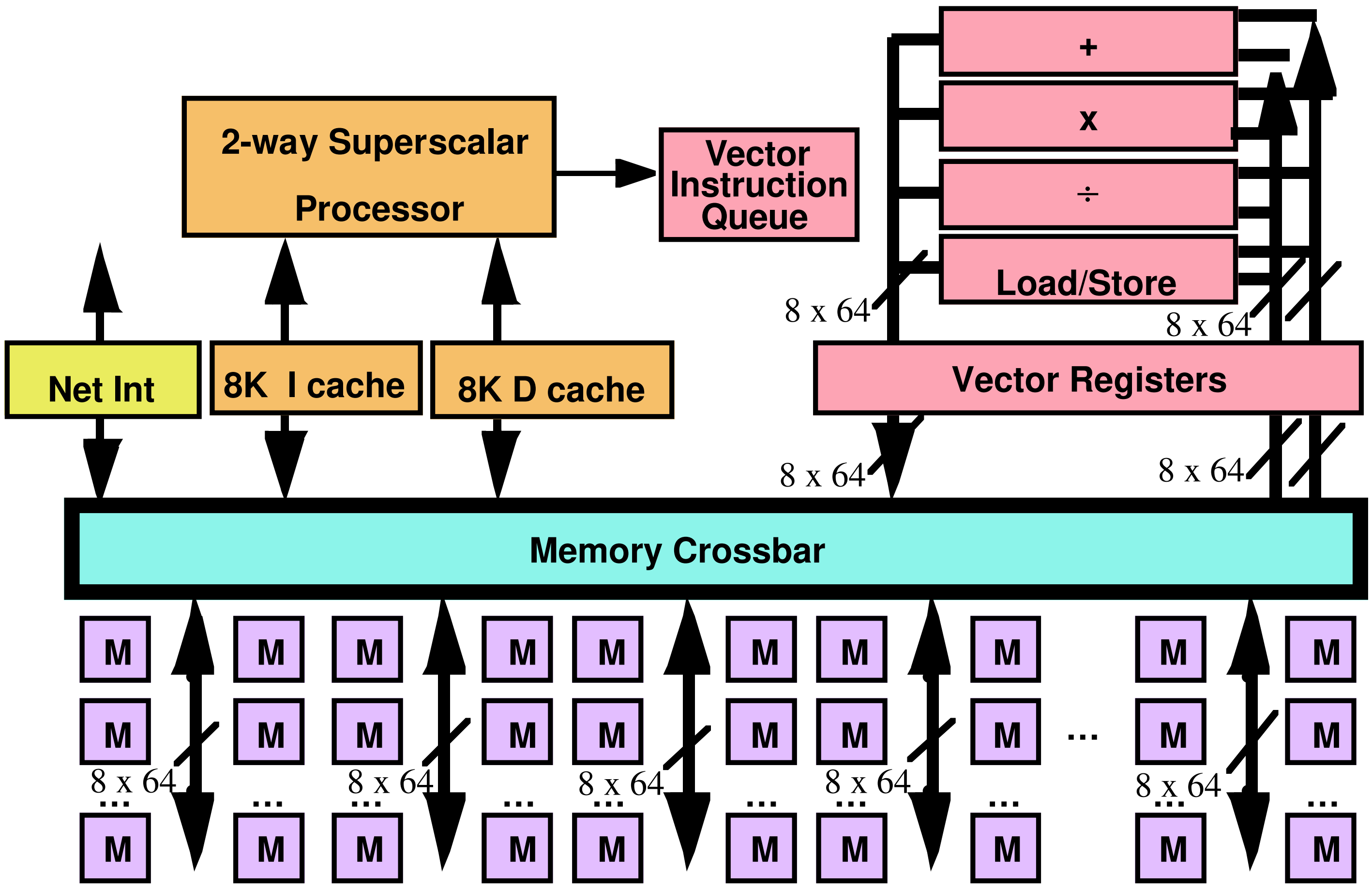
- High cost: ≈ $1M / processor?
- ≈5-10M transistors for vector processor?
- Low latency, high BW memory system?
- Energy?
- Poor scalar performance?
- Limited to scientific applications?

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

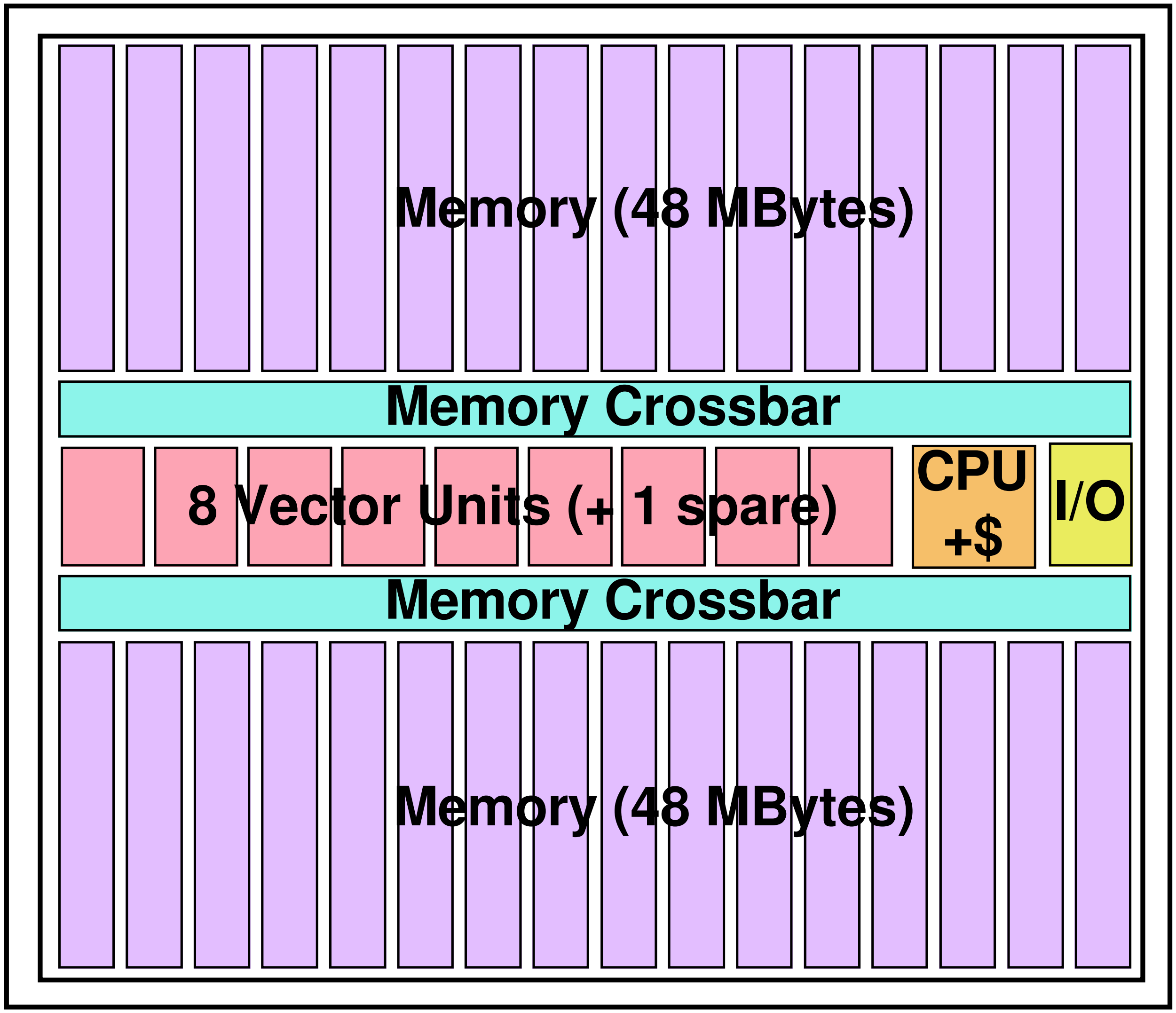
Software Technology Trends Affecting V-IRAM?

- V-IRAM: <u>any</u> CPU + V-IRAM co-processor on-chip
  - scalar/vector interactions are limited, simple
- Vectorizing compilers built for 25 years
  - can buy one for new machine from The Portland Group
- Library solutions; retarget packages (e.g., MMX)
- SW distribution model is evolving?
  - Old Model SW distribution: binary for processor on CD
  - New Model #1: Binary translation to new machine?
  - New Model #2: 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 ⇒ ≈80% die repairable

Vector IRAM Generations

- V-IRAM-1 (≈1999)
- 256 Mbit generation (0.25)
- Die size = DRAM (290 mm^2)
- 1.5 - 2.0 volts (logic)
- 0.5 - 2.0 watts
- 300 - 500 MHz
- 4 64-bit pipes/lanes
- 4 GFLOPS(64b)/32GOPS(8b)
- 24 MB capacity + DRAM bus
- PCI bus/Fast serial lines

- V-IRAM-2 (≈2002)
- 1 Gbit generation (0.18)
- Die size = DRAM (420 mm^2)
- 1.0 - 1.5 volts (logic)
- 0.5 - 2.0 watts
- 500 - 1000 MHz
- 8 64-bit pipes/lanes
- 16 GFLOPS/128GOPS
- 96 MB cap. + DRAM bus
- Firewire/FC-AL serial lines

IRAM Applications

- "Supercomputer on a AA battery"
  - Super PDA/Smart Phone:
    speech I/O + "voice" email + pager + GPS +...
  - Super Gameboy/Portable Network Computer:
    3D graphics + 3D sound + speech I/O+ Gbit link + ...
- Intelligent SIMM ("ISIMM")
  - Put 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
  - Apps? Full text search? Fast sort? No index database?
- Intelligent Disk ("IDISK") 2.5" disk + IRAM + net.

ISIMM/IDISK Example: Sort

- Berkeley NOW cluster has world record sort: 6GB disk-to-disk using 64 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
- IRAM: ≈ 2-4 GIPS + 2 2-4Gb/s I/O + 2 2-4Gb/s Net
- ISIMM: 8 IRAMs + net swtich + FC-AL links + disks
- IDISK: Intelligent Disks + switch = cluster

Why IRAM now?
Lower risk than before

- DRAM manufacturers now facing challenges
  – Before not interested, so early IRAM = SRAM
- Past efforts memory limited ⇒ multiple chips
  ⇒ <u>1st</u> solve the unsolved (parallel processing)
  – Gigabit DRAM ⇒ ≈100 MB; OK for many apps?
- Fast Logic + DRAM available now/soon?
- Embedded apps leverage energy efficiency, adjustable mem. capacity, smaller board area
  ⇒ OK market v. desktop (55M 32b RISC '96)

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 yield, power, testing, memory size
- 10X-100X improvements based on technology shipping for 20 years (not photons, MEMS, ...)
- Potential shift in balance of power in DRAM/ microprocessor (μP) industry in 5-7 years?

  μP-oriented vs. DRAM-oriented manufacturers:

  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

Backup Slides

(The following slides are used to help answer questions)

Why a company should try IRAM

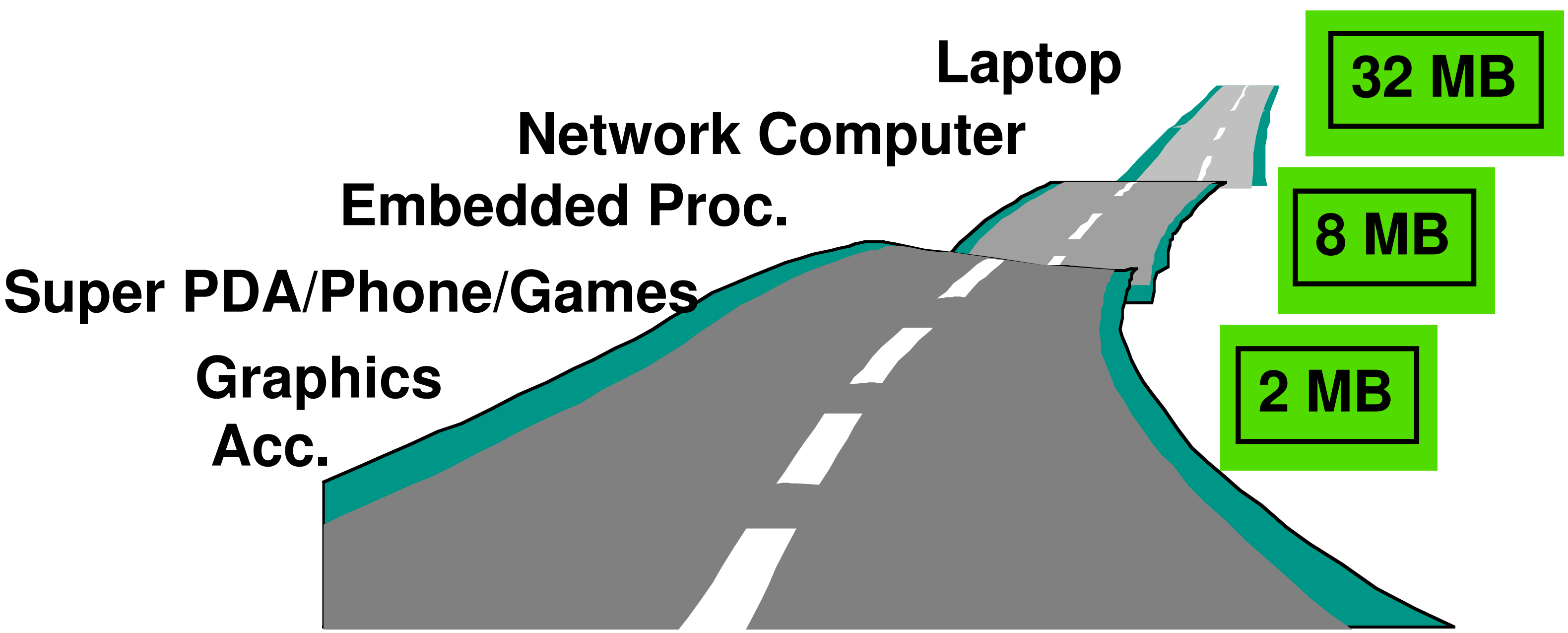
- If IRAM doesn't happen, then someday:
  - $10B fab for 16B Xtor MPU (too many Xtors per die)??
  - $10B fab for 16 Gbit DRAM (too many bits per die)??
- This is not rocket science. In 1997:
  - 25-50X improvement in memory density;
    ⇒ more memory per die or smaller die
  - 10X -100X improvement 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

Commercial IRAM highway is governed by memory per IRAM?



Energy to Access Memory by Level of Memory Hierarchy

- For 1 access, measured in nJoules

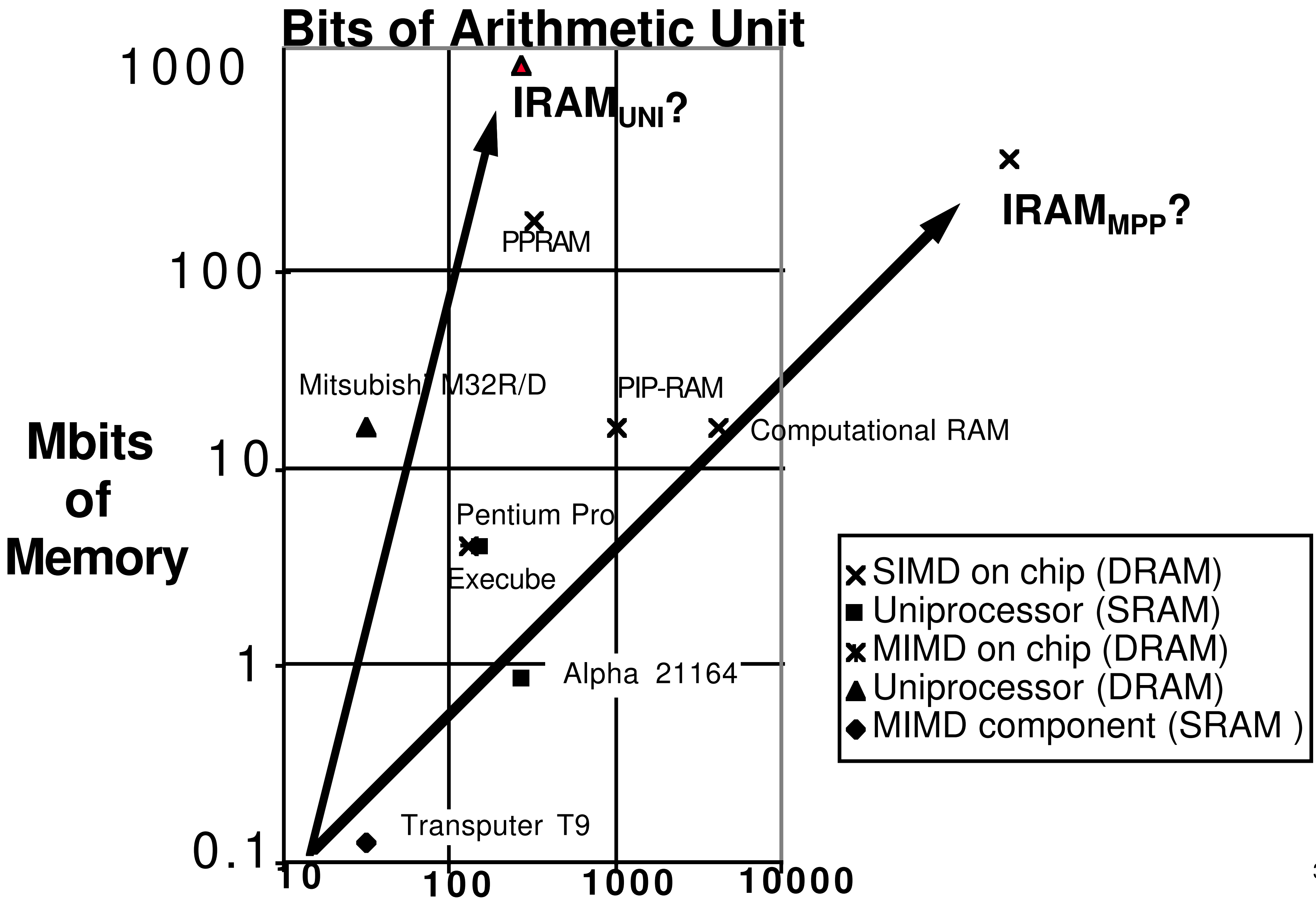
	Conventional	IRAM
on-chip L1$(SRAM)	0.5	0.5
on-chip L2$(SRAM v. DRAM)	2.4	1.6
L1 to Memory (off- v. on-chip)	98.5	4.6
L2 to Memory (off-chip)	316.0	*(n.a.)*

  » Based on Digital StrongARM, 0.35 µm technology

  » See "The Energy Efficiency of IRAM Architectures," *24th Int'l Symp. on Computer Architecture*, June 1997

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?
  - Wider DRAMs more expensive: bigger die, test time



Bits of Arithmetic Unit
1000
100
10
1
0.1
10
100
1000
10000
Mbits
of
Memory
IRAM$_{UNI}$?
IRAM$_{MPP}$?
PPRAM
Mitsubishi M32R/D
PIP-RAM
Computational RAM
Pentium Pro
Execube
Alpha 21164
Transputer T9
SIMD on chip (DRAM)
Uniprocessor (SRAM)
MIMD on chip (DRAM)
Uniprocessor (DRAM)
MIMD component (SRAM)

IRAM Conclusion

- Research challenge is quantifying the evolutionary-revolutionary spectrum

Evolutionary

Packaging

Standard CPU in DRAM process

Standard ISA+ Vector CPU in DRAM process

New ISA CPU in DRAM process

New ISA + FPGA in DRAM process

Revolutionary

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 + Protocols (1993-98)
- IRAM: Instruction set /Processor design/Memory Hierarchy and Compilers/OS (1996-200?)

21st Century Benchmarks?

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