A Case for Intelligent DRAM: IRAM

David Patterson,

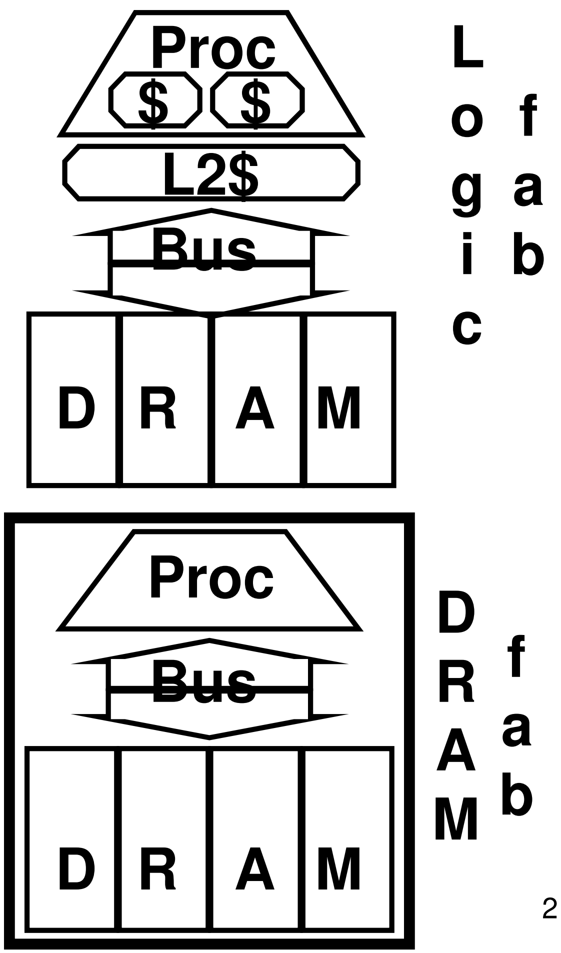
Tom Anderson, Kathy Yelick

Computer Science Division
University of California at Berkeley
http://iram.cs.berkeley.edu/

Hot Chips VIII   August, 1996

IRAM Vision Statement

- Microprocessor & DRAM on single chip:
  - » bridge the processor-memory performance gap via on-chip latency & bandwidth
  - » improve power-performance (no DRAM bus)
  - » lower minimum memory size (designer picks any amount)



Outline

- Today’s Situation: Microprocessor
- Today’s Situation: DRAM
- Alternatives to Today’s Situation
- IRAM Opportunities
- Related Work
- Research Agenda
- IRAM Potential Impact

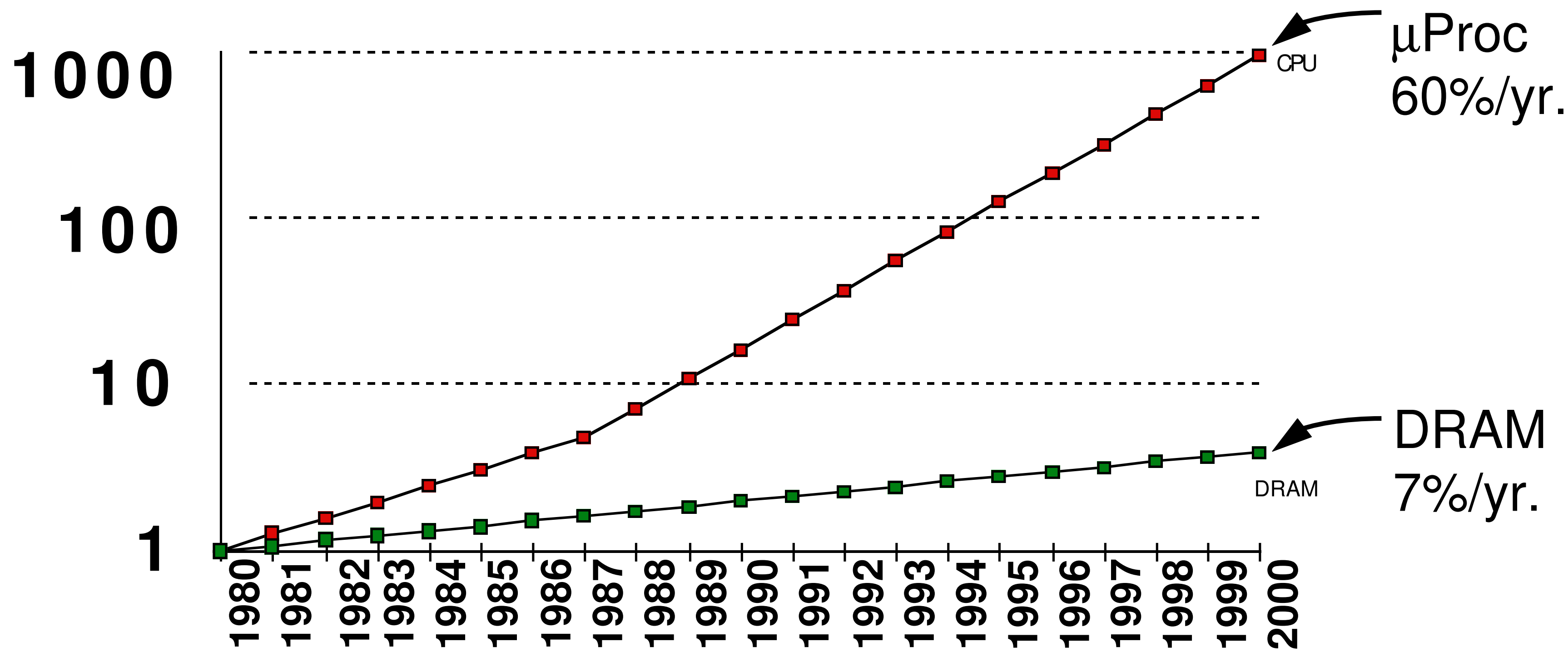
Today's Situation: Microprocessor

- Microprocessor-DRAM performance gap
  - » full cache miss time = 100s instructions

    (Alpha 7000: 340 ns/5.0 ns = 68 clks x 2 or 136)

    (Alpha 8400: 266 ns/3.3 ns = 80 clks x 4 or 320)
- Rely on locality + caches to bridge gap
- Still doesn't work well for some applications: data bases, CAD tools, sparse matrix, ...
- Power limits performance (battery, cooling)

Processor-DRAM Gap (latency)



I cache '84; I+D '86; I+D+L2 '95

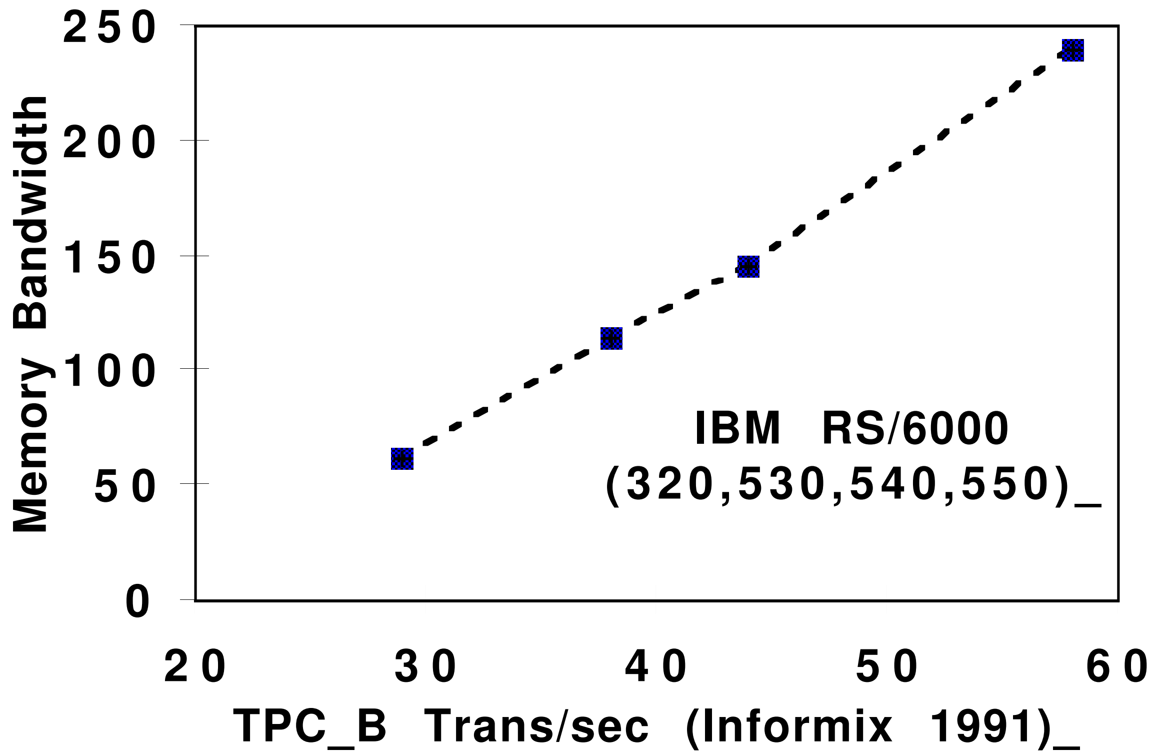
0.25K; 2x0.25K; 2x8K+96K

Works poorly for some applications

- Sites and Perl [1996]
  - » Alpha 21164, 300 MHz, 4-way superscalar
  - » Running Microsoft SQLserver database on Windows NT operating system, it operates at 12% of peak bandwidth
    (Clock cycles per instruction or CPI = 2.0)
  - » “The implication of this is profound -- caches don’t work.”

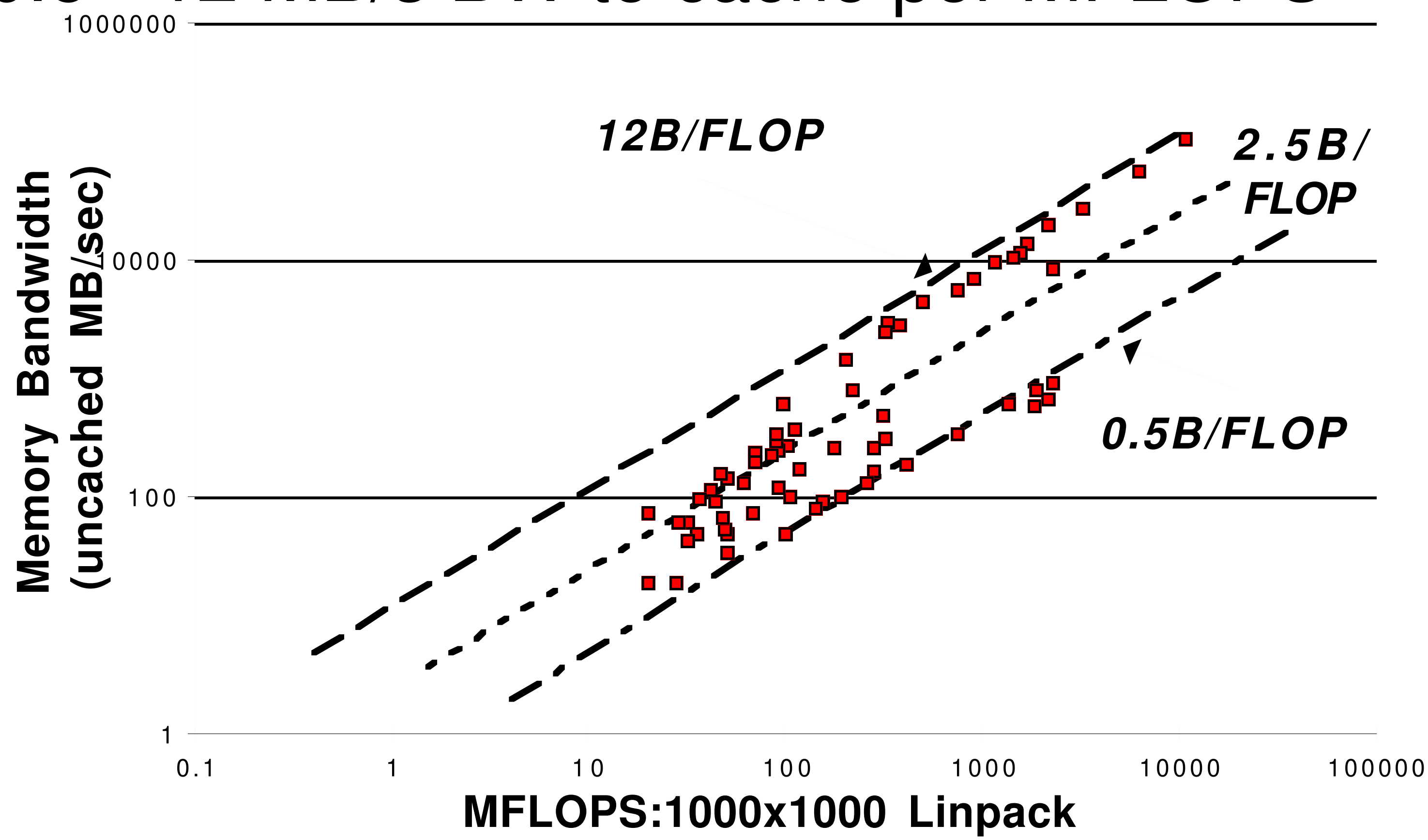
Speed tied to Memory BW: Database

- ≈3 MB/s BW to cache per Trans/s



Speed tied to Memory BW: Linpack

- 0.5 - 12 MB/s BW to cache per MFLOPS



Available Options: Microprocessor

- Memory controller on chip
- Packaging breakthrough: fast DRAMs with 100s of pins, MPers with 1000s?
  - » Cost? Bare die? Standard? Latency?
- More levels of caches (L4?), prefetching?
- Larger instruction window, more outstanding memory references?
- IRAM: processor + DRAM on same chip?

Today's Situation: DRAM

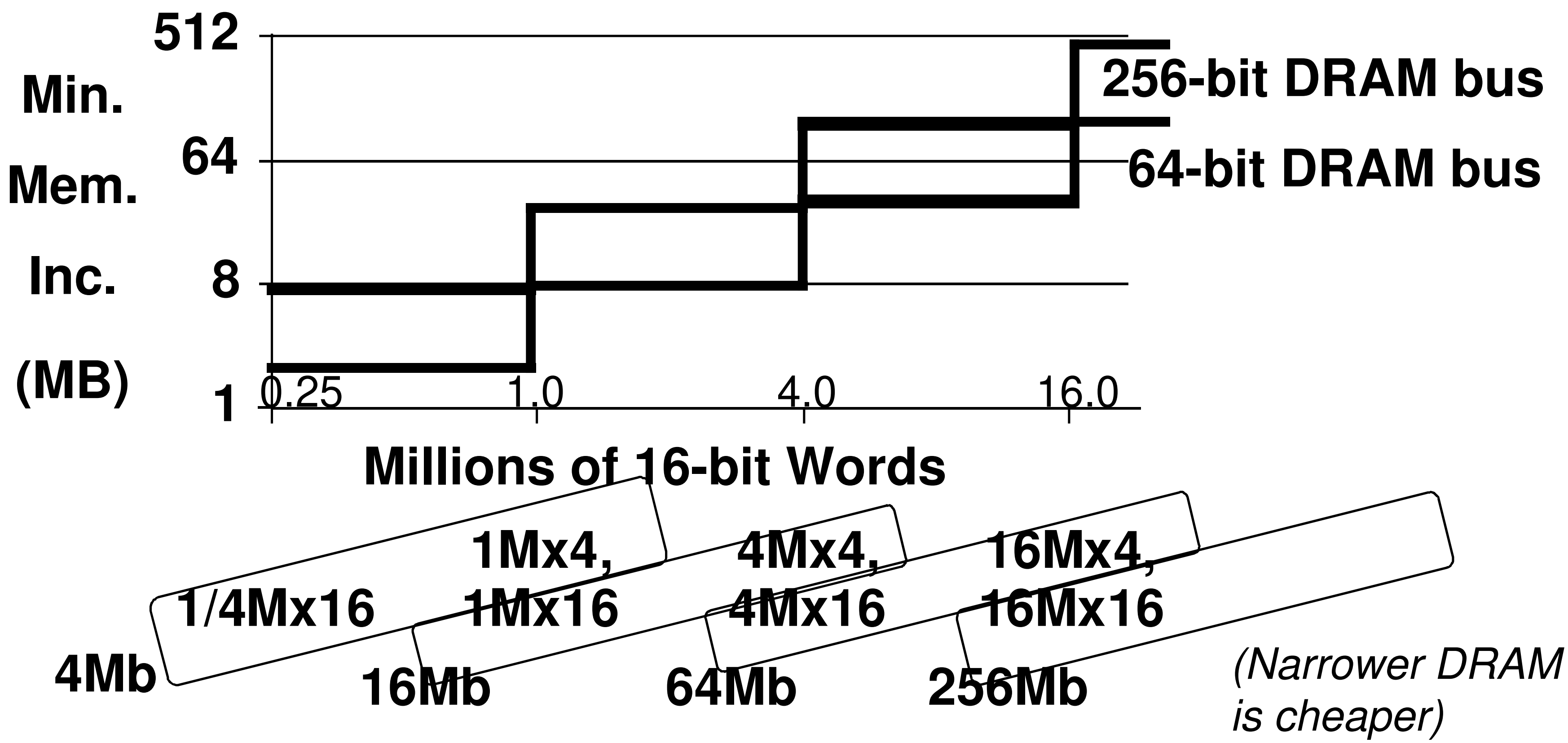
- Commodity, second source industry
  => high volume, low profit, conservative
  - » Little organization innovation in 20 years:
    page mode, EDO, probably Synch DRAM?
- Order of importance: 1) Cost/bit 2) Capacity
  - » RAMBUS: 10X BW, +30% cost => little impact
- Fewer DRAMs/computer over time
- Starting to question buying larger DRAMs?
  - » Limited BW from larger DRAMs & DRAM latency ≠ BW

DRAMs per System over Time

DRAM Generation

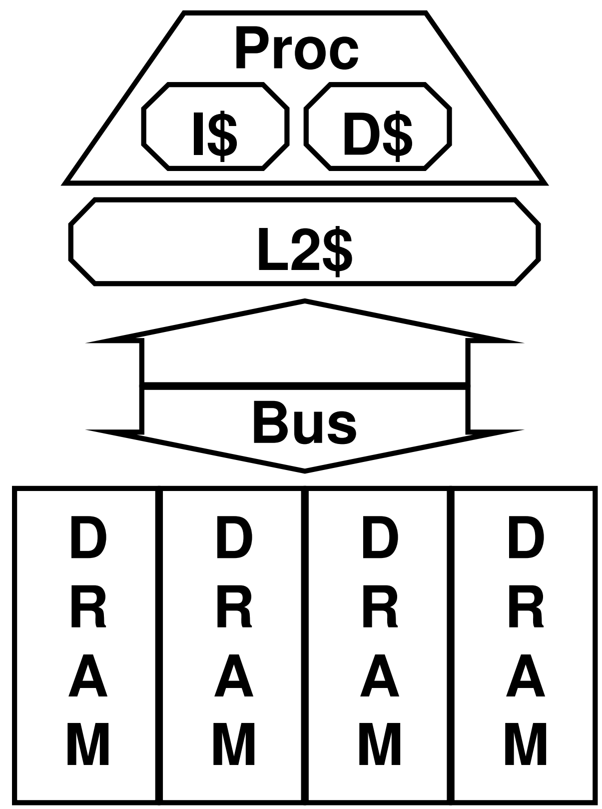
Minimum Memory Size	'86 1 Mb	'89 4 Mb	'92 16 Mb	'96 64 Mb	'99 256 Mb	'02 1 Gb
4 MB	32 →	8				
8 MB		16 →	4			
16 MB			8 →	2		
32 MB				4 →	1	
64 MB				8 →	2	
128 MB					4 →	1
256 MB					8 →	2

Bus Width, DRAM gen., Memory Size



DRAM Latency >> BW

- More App Bandwidth =>
  Cache misses
  => DRAM RAS/CAS
- Application BW =>
  Lower DRAM <u>Latency</u>
- RAMBUS, Synch DRAM
  increase BW but <u>higher</u>
  latency
- EDO DRAM < 5% in PC



Available Options: DRAM

- Packaging breakthrough allowing low cost, high speed DRAMs with 100s of pins, microprocessors with 1000s of pins
  - » Cost? Bare Die? Standard? Latency?
- 2.5X cell/area & <u>smaller</u> die DRAM => lower cost, fixed capacity per chip
  - » DRAM industry invest?
- IRAM: processor + DRAM on same chip

Multiple Motivations for IRAM

- Performance gap increasingly means performance limit is memory
- Dwindling interest in future DRAM generations: 64 Mb? 256 Mb? 1 Gb?
  - » Higher capacity/DRAM => system memory BW <u>worse</u>
  - » Higher BW/DRAM => <u>higher</u> cost/bit & memory latency/ app BW <u>worse</u>
- Caches don’t work for all apps

IRAM Pros and Cons

- Potential IRAM Advantages:
  - » Minimum memory increment adjustable
  - » Width, Bandwidth => greater performance
  - » Lower latency => greater performance
  - » Fewer pins => less power (cost?)
- IRAM Challenges
  - » Speed, area, yield vs. conventional designs?
  - » Expandable memory solution?
  - » Business model: volume? 2nd source? cost/bit?

1 Gbit DRAM Parameters

	Mitsubishi	Samsung
Blocks	512 x 2 Mbit	1024 x 1 Mbit
Clock	200 MHz	250 MHz
Pins	64	16
Die Size	24 x 24 mm	31 x 21 mm
Metal Layers	3	4
Technology	0.15 micron	0.16 micron

(ISSCC ‘96; production ‘02?)

<u>Potential</u> 1 Gbit IRAM BW: 100X

- 1024 1Mbit modules, each 1Kb wide
  - » 10% @ 40 ns RAS/CAS = 320 GBytes/sec
- If 1Kb bus = 1mm @ 0.15 micron
  => 24 x 24 mm die could have 16 busses
- If bus runs at 50 to 100 MHz on chip
  => 100-200 GBytes/sec
- FYI: AlphaServer 8400 = 1.2 GBytes/sec
  - » 75 MHz, 256-bit memory bus, 4 banks

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"?
- FYI:
  AlphaSta. 600: 180 ns=128b, 270 ns= 512b
  AlphaSer. 8400: 266 ns=256b, 280 ns= 512b

Potential Power Advantage: 2 - 3X

- CPU + memory ≈ 40% power in portable
- Memory power = f(cache, bus, memory)
  - Smaller cache => less power for cache but use bus & memory more
  - As vary cache size/hit rate, bus ≈ 24% power
- Larger DRAM on-chip cache, on-chip bus => IRAM improve power 2X to 3X?

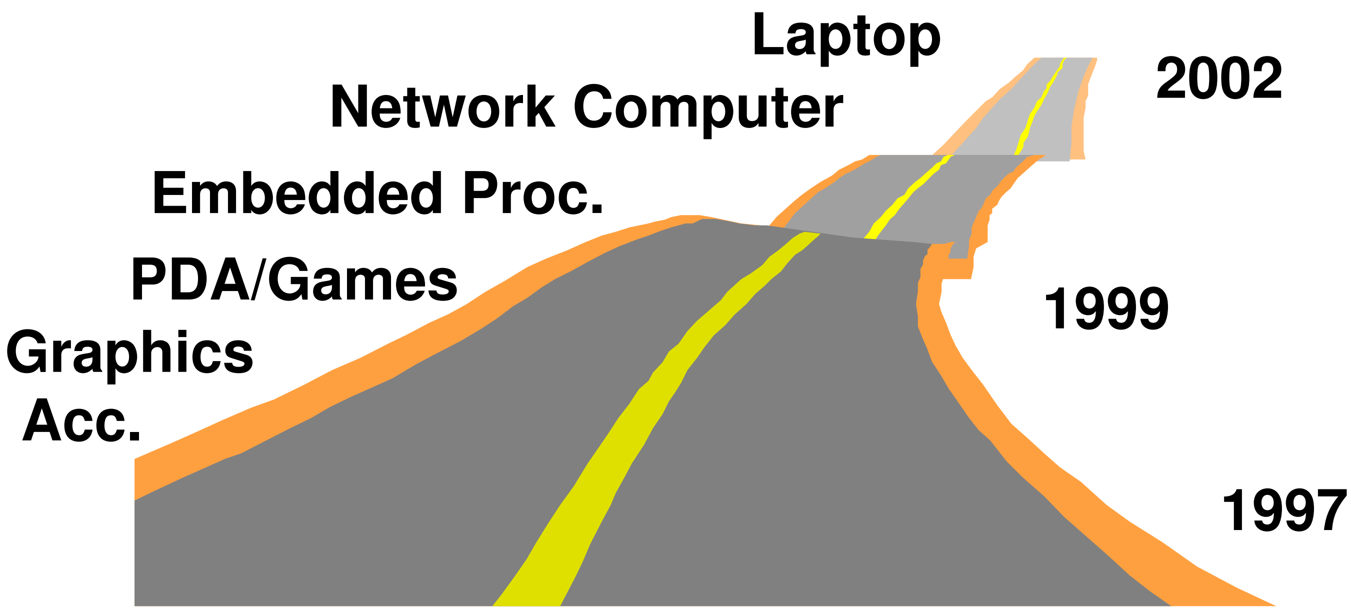
IRAM Challenges

- Chip
  - » Speed, area, power, yield in DRAM process?
  - » Good performance and reasonable power?
  - » BW/Latency oriented DRAM tradeoffs?
- Architecture
  - » How to turn high memory bandwidth into performance?
    - – Vector: (n elements/clock) vector units?
    - – Extensive Prefetching?
  - » Extensible IRAM: Large pgm/data solution?

Why might IRAM succeed this time?

- DRAM manufacturers facing challenges
  - » Before not interested, so early IRAM = SRAM
- Past efforts memory limited => multiple chips => <u>1st</u> solve parallel processing
  - » Gigabit DRAM => 128 MB; OK for many?
- Embedded applications offer large 2nd target to conventional computing (business)
- 1st Customer Ship of IRAM closer to 1st Customer Ship of system

IRAM Highway?



IRAM Conclusion

- Research challenge is quantifying the evolutionary-revolutionary spectrum
- IRAM rewards creativity as well as manufacturing; shift balance of power in DRAM/microprocessor industry?

Evolutionary

Packaging

Standard CPU in DRAM process

Prefetching CPU in DRAM process

Vector CPU in DRAM process

CPU+ FPGA in DRAM process

Revolutionary

Backup Slides

(The following slides are used to help answer questions)

Is IRAM a New Idea?

	≤ 1 Mb	4 - 16 Mb	≥ 64 Mb
Memory	Video DRAM	3D DRAM	
UniPro-cessor		SHARC NEC	Mitsubishi
MIMD node	Transput., J-mach.		[Sau 96]
S/MIMD on chip	MM32k	Execube, PIP-RAM	

- More as get more memory? DRAM ASIC?

IRAM Related References

- `http://www.cs.berkeley.edu/~patterson/lecture294.html`

[Bur96] Burger, Doug; Kagi, Alain; Goodman, James R. “Memory Bandwidth Limitations of Future Microprocessors,” ISCA, Philadelphia, PA USA, May 1996.

[Dal92] Dally, W.J.; et al. The message-driven processor: a multicomputer processing node with efficient mechanisms. IEEE Micro, April 1992, vol.12, (no.2):23-39.

[Dee94] Deering, M.F.; Schlapp, S.A.; Lavelle, M.G. “FBRAM: a new form of memory optimized for 3D graphics,” SIGGRAPH 94 Conference Proceedings. Orlando, FL, USA, 24-29 July 1994). p. 167-74.

[Kog95] Kogge, P.M.; Sunaga, T.; Miyataka, H.; Kitamura, K.; and others. “Combined DRAM and logic chip for massively parallel systems.” Proceedings. Sixteenth Conference on Advanced Research in VLSI, Chapel Hill, NC, USA, 27-29 March 1995, p. 4-16.

[Lip92] Lipovski, G.J. “A Four Megabit Dynamic SYstolica Associative Memory Chip.” Journal of VLSI Signal Processing, vol. 4, 37-51.

[Sau96] Ashley Saulsbury, Fong Pong, Andreas Nowatzk, “Missing the Memory Wall: The Case for Processor/Memory Integration,” ISCA , Philadelphia, PA USA, May 1996.

[Wul95] Wulf, W.A.; McKee, S.A. Hitting the memory wall: implications of the obvious. Computer Architecture News , March 1995, vol.23, (no.1):20-24.

Other References

[Cve94] Cvetanovic, Z.; Bhandarkar, D. Characterization of Alpha AXP performance using TP and SPEC workloads. ISCA, Chicago, IL, USA, 18-21 April 1994, p. 60-70.

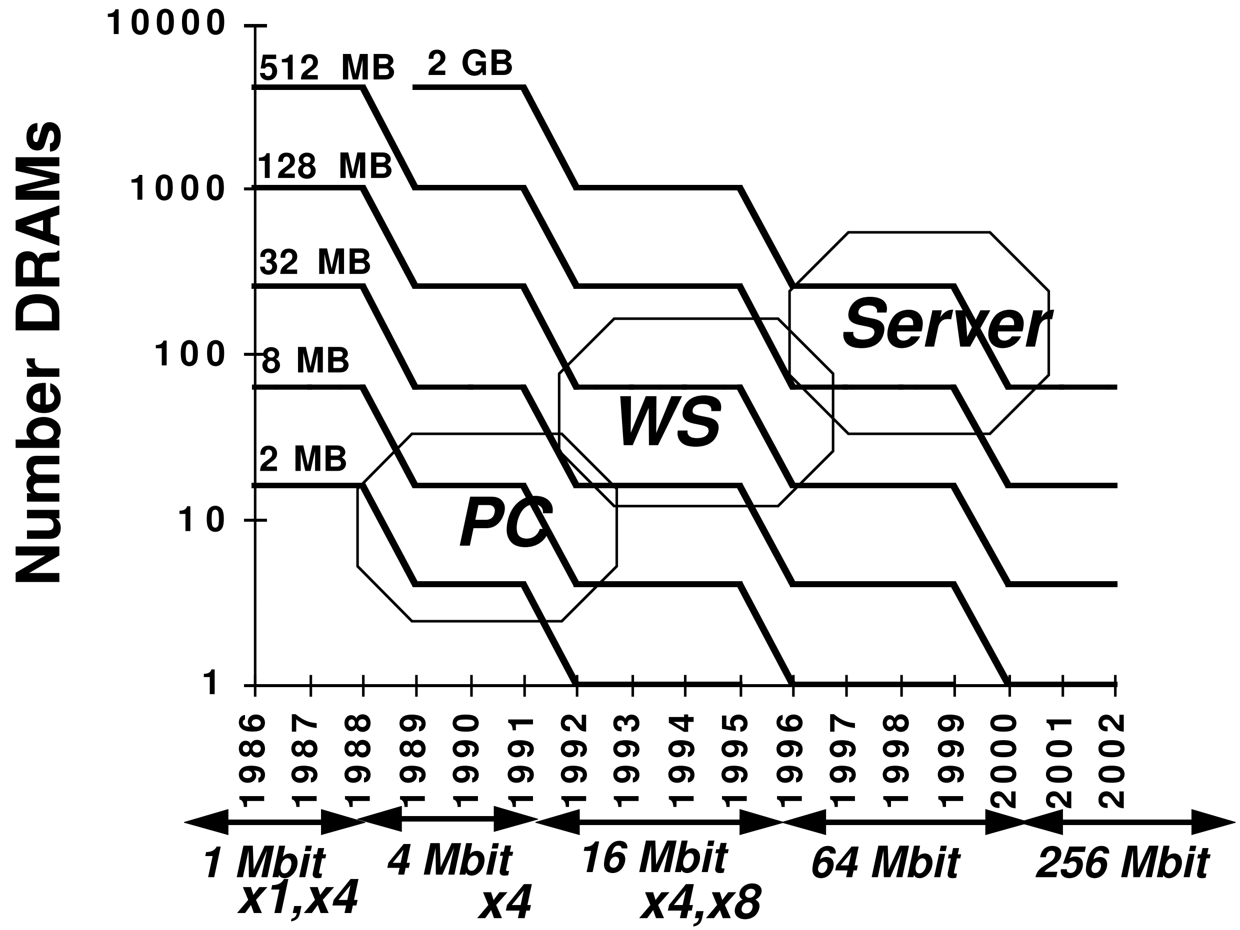
[Prz94] Przybylski, Steven A. New DRAM Technologies: A Comprehensive Analysis of the New Architectures, MicroDesign Resources, Sebastopol, California, 1994.

[Sei92] Seitz, C.L. Mosaic C: an experimental fine-grain multicomputer. Future Tendencies in Computer Science, Control and Applied Mathematics. Paris, France, 8-11 Dec. 1992). Edited by: Bensoussan, A.; Verjus, J.-P. Berlin, Germany: Springer-Verlag, 1992. p. 69-85.

[Tar91] Tarui, Y.; Tarui, T. New DRAM pricing trends: the bi rule. IEEE Circuits and Devices Magazine, March 1991, vol.7, (no.2):44-5.

[Sit96] Sites, Richard L.; Perl, Sharon A. “PatchWrx--A Dynamic Execution Tracing Tool,” 1996. http://www.research.digital.com/SRC/staff/sites/bio.html.

DRAMs per System (Based on Figure 2-7 and 2-8, Przybylski 1994)



Works poorly for some benchmarks

- Cache (1995) vs. Vector (1991)

	Alpha 8400 5/300	Cray C90
Clock	300 MHz	240 MHz
Cache	8K+8K+96K+4MB	1 KB
su2cor	7 SPECbase95	25 (3.5x)
swim	19 SPECbase95	141 (7.4x)

- Nonconventional memory system, instruction set offers 3.5X to 7.4X speedup

Applications

- Apps limited by amount of DRAM
  - » Fast DRAM limited by cost/bit vs. generic
  - » Memory in Processor vs. Multiprocessor in Memory
- Graphics: VDRAM, 3D DRAM
- Cache: IBM 1MB L2 cache (1996)
- Low power, vector => DSP, Embedded
- Vector: what % new apps vectorize? (multimedia, encryption, compression)

Related Work: Accelerators

- Apps limited by amount of DRAM
- Graphics:
  - » Video RAM (TI, 1983): 10% DRAM market
  - » 3D DRAM from Sun/Mitsubishi (1995)
  - » Many startups: Silicon Magic, Neomagic
- Cache: IBM 1MB L2 cache (1996)
- Toshiba (and others): 1 - 8 MB DRAM macrocell in ASIC logic technology

Related Work: SIMD Multiprocessors

- SIMD on chip : Logic in Memory
  - » Array processing: ALUs in memory (1960s)
  - » Half dozen examples in DRAM process
  - » MM32k: 2048 1-bit PE + 1Mb DRAM/chip, 32K PE in 16 chips for Neural Net apps
  - » Comp RAM: 4096 1-bit PE + 16Mb DRAM/chip for DSP (MOSAID, ISSCC 96)
  - » PIP-RAM: 128 8-bit PE + 16 Mb DRAM/chip for image processing (NEC, ISSCC 96)

Related Work: MIMD Multiprocessors

- MIMD: Processor + SRAM + net interface
  - » Inmos Transputer (1978-96): 1- 4 KB + 16-bit CPU @ 5-20 MHz (T9@20MHz,32-bit,16KB)
  - » MIT J-machine ‘91:18 KB+16 MHz,36-bit CPU
  - » Caltech Mosaic-C ‘91: 64KB DRAM + 30 MHz, 16-bit CPU
- MIMD on chip: IBM Execube (1994)
  - » <u>Eight</u> 25 MHz,16-bit CPUs + 8 x 64 KB (4Mbit)

Related Work: Faster DRAM

- Faster DRAM/ “processors are free”
- Cost/bit dominates: See RAMBUS
- Paced by success of DRAM generation
- Salisbury, Notwacyk study (ISCA 1996):
  - » 1 scalar, 5-stage SPARC + 256 Mbit DRAM
  - » Limit to 10% die area; 2 64-bit buses
  - » Small, wide block caches,
  - » SPEC95: @ 200 MHz ≈ same integer perf. , ≈ 1/2 floating point perf. as 300 MHz Alpha

Related Work: Uniprocessors

- Apps limited by amount of DRAM
- A/D SHARC: 1MB SRAM + 100 MFLOPS (32b) DSP
- Mitsubishi: 8MB + “multimedia” or “PDA” RISC (ISSCC 1996)
- NEC: 4?MB + MIPS Core video game?

Simpler processor might win?

- Historical precedent: IBM 360
  - » 360/91 with register renaming, out-of-order execution (forerunner of IBM PowerPC 620, Intel Pentium Pro, MIPS R10000)
  - » 360/85 with cache (a better memory system)
  - » Clock rate of 91 vs. 85: 60 ns vs. 80 ns
  - » Memory speed: 750 ns vs. 1040 ns
  - » Memory interleaving: 8-way vs. 4-way
  - » 360/85 faster on 8 of 11 programs