Lecture 10:
Intelligent IRAM and
Doing Research in the Information Age

Professor David A. Patterson

Computer Science 252

Fall 1996

Review: Cache Optimization

$$CPUtime = IC \times \left(CPI_{Execution} + \frac{Memory\ \ accesses}{Instruction} \times \textbf{Miss rate} \times Miss\ \ penalty \right) \times Clock\ \ cycle\ \ time$$

Technique	*MR*	*MP*	*HT*	*Complexity*
Larger Block Size	+	–		0
Higher Associativity	+		–	1
Victim Caches	+			2
Pseudo-Associative Caches	+			2
HW Prefetching of Instr/Data	+			2
Compiler Controlled Prefetching	+			3
Compiler Reduce Misses	+			0
Priority to Read Misses		+		1
Subblock Placement		+	+	1
Early Restart & Critical Word 1st		+		2
Non-Blocking Caches		+		3
Second Level Caches		+		2
Small & Simple Caches	–		+	0
Avoiding Address Translation			+	2
Pipelining Writes			+	1

Review: Main Memory

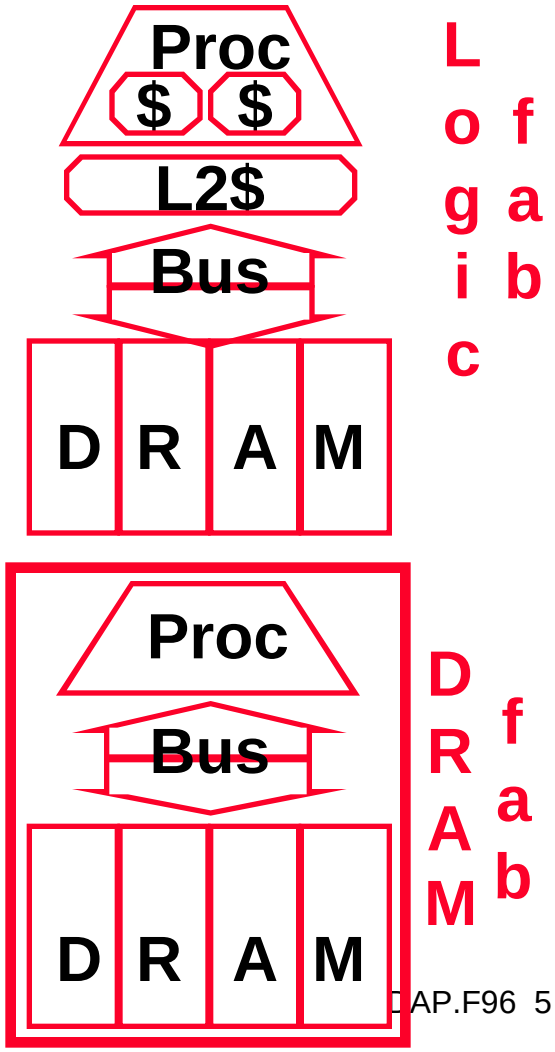
- **Wider Memory**
- **Interleaved Memory: for sequential or independent accesses**
- **Avoiding bank conflicts: SW & HW**
- **DRAM specific optimizations: page mode & Specialty DRAM (EDO, RAMBUS, Synchronous)**
- **DRAM future less rosy?**

Review: Reducing Miss Penalty

- Five techniques
  - Read priority over write on miss
  - Subblock placement
  - Early Restart and Critical Word First on miss
  - Non-blocking Caches (Hit Under Miss)
  - Second Level Cache
- Can be applied recursively to Multilevel Caches
  - Danger is that time to DRAM will grow with multiple levels in between

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)**



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)**

Memory Latency: System vs. Chip

• Processor	Alpha 21164
• Machine	AlphaServer 8200
• Clock Rate	300 MHz
• Caches	8K I, 8K D, 96K L2, 4M L3
• I Cache Latency	6.7 ns (2 clocks)
• D Cache	6.7 ns (2 clocks)
• L2 Cache	20 ns (6 clocks)
• L3 Cache	20 ns (6 clocks)
• Main Memory	253 ns (76 clocks)
• Single DRAM component	external: ≈60ns (18 clocks)
	internal: ≈40ns (12 clocks)

Processor-Memory Gap Penalty

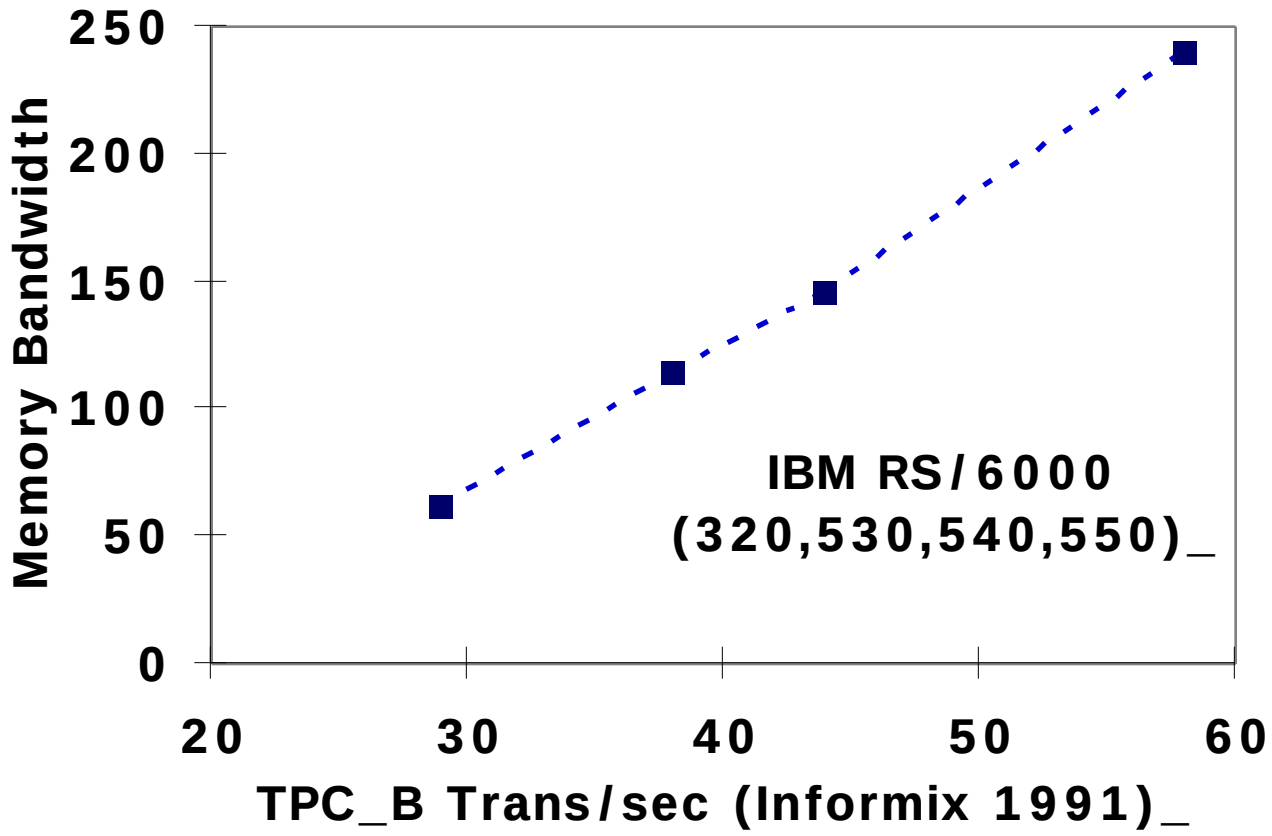
Microprocessor	Caches	% Area	% Transistors
• Alpha 21064	I: 8 KB, D: 8 KB	21%	60%
• Alpha 21164	I:8KB,D:8KB,L2:96KB	37%	77%
• Strong-Arm	I: 16 KB, D: 16 KB	61%	94%
• 80386 (‘85)	0 (12 B Prefetch)	6%	≈5%
• 80486 (‘89)	8 KB	20%	≈50%
• Pentium (‘93)	I: 8 KB, D: 8 KB	32%	≈32%
• Pentium Pro	I: 8 KB, D: 8 KB, + L2: 512 KB	P: 22% +L2: 100% (Total: 64%)	P: 18% +L2: 100% (Total: 88%)

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.”**

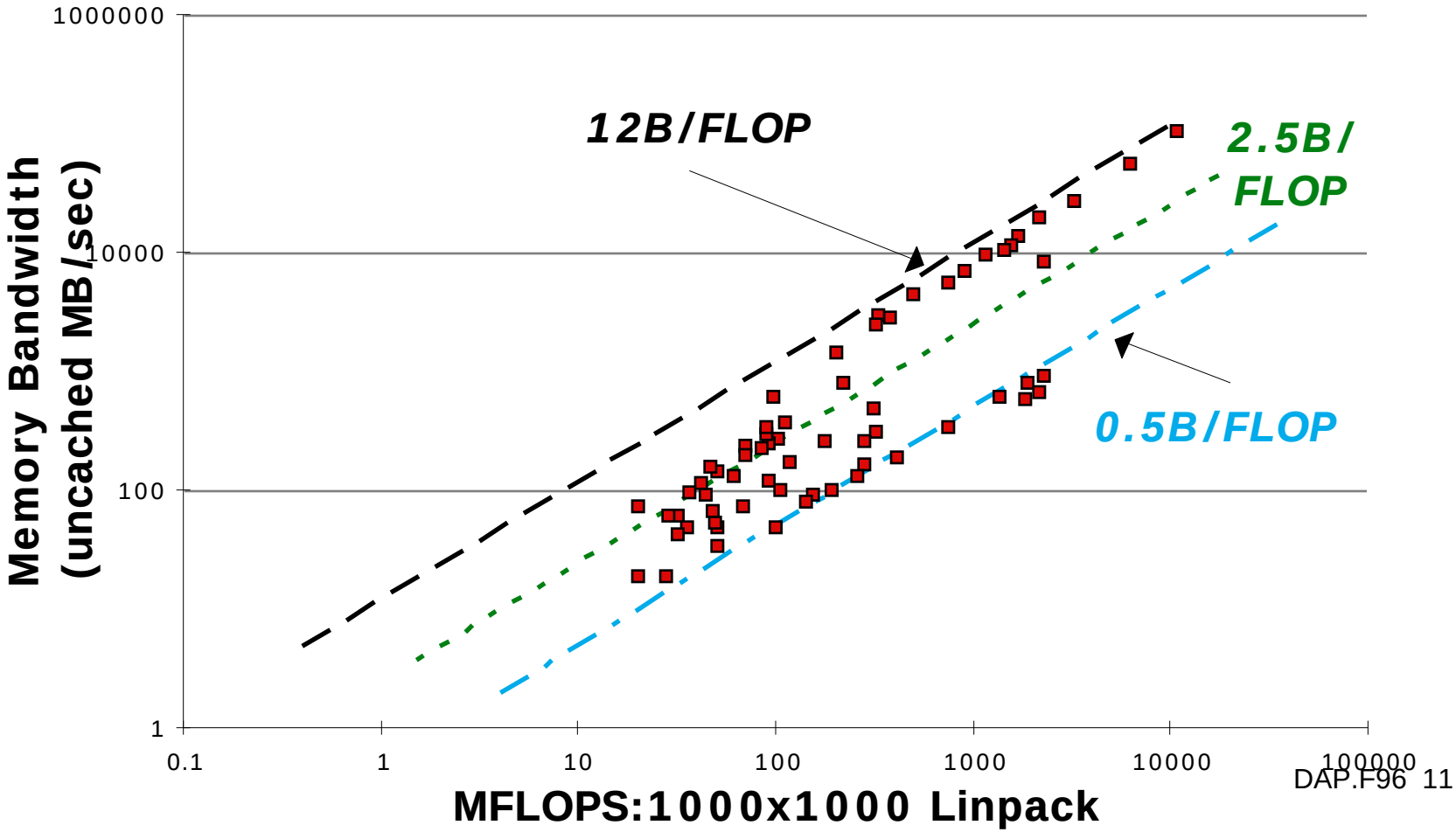
Speed tied to Memory BW: Database

- ≈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

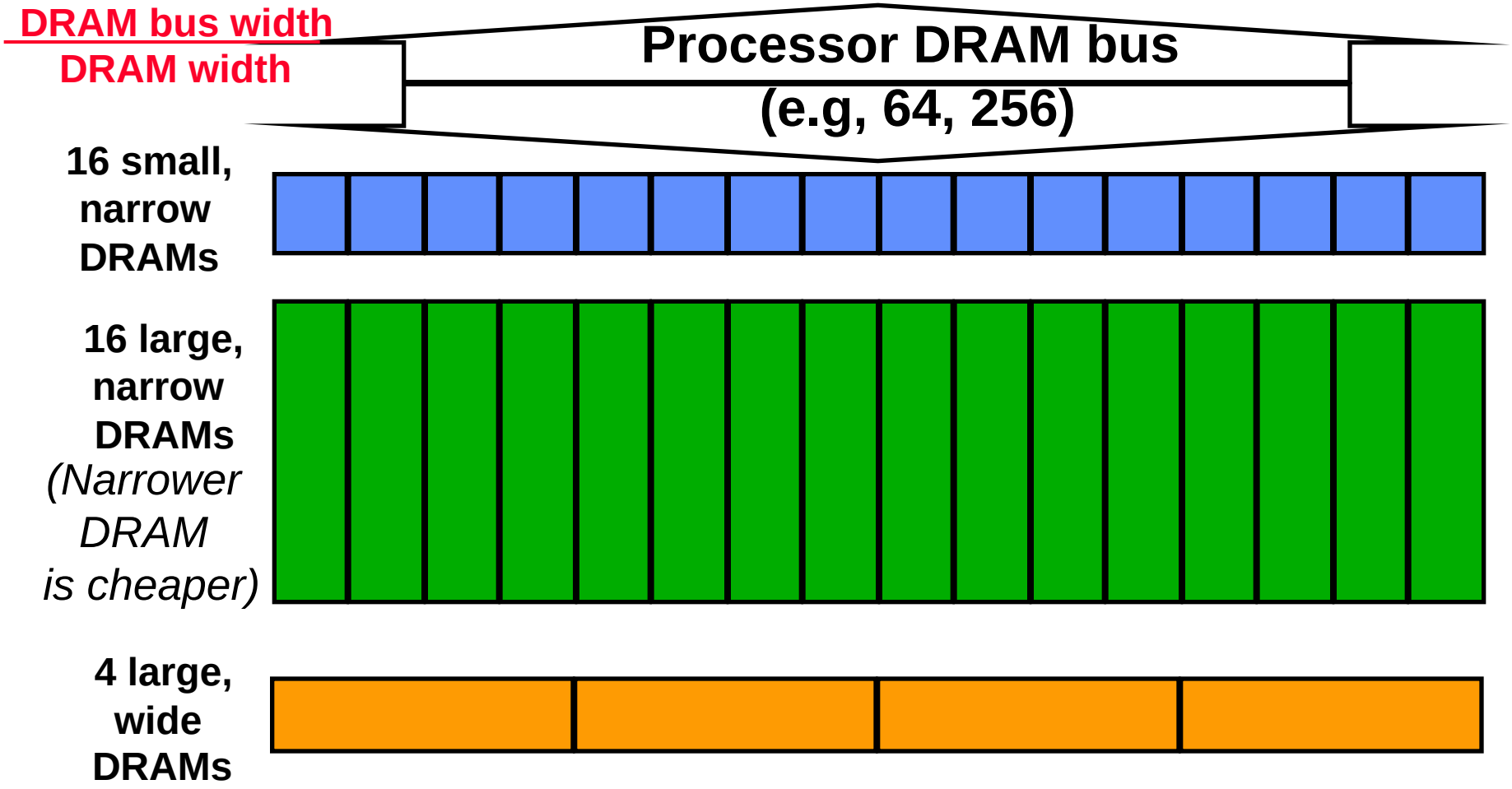
- Memory controller on chip
- Packaging breakthrough: fast DRAMs with 100s of pins, MPers with 1000s?
  - Cost? Bare die? Standard? Latency?
- More levels of caches (L4?), prefetching?
- Larger instruction window, more outstanding memory references?
- IRAM: processor + DRAM on same chip?

Potential
DRAM Crossroads?

- Fewer DRAMs per computer over time
- Starting to question buying larger DRAMs?
  - Limited BW from larger DRAMs
- After 20 years of 4X every 3 years, running into wall? (64Mb - 1 Gb)
- How can keep $1B fab lines full if buy fewer DRAMs per computer?
- Cost/bit –30%/yr if stop 4X/3 yr?
- What will happen to $40B/yr DRAM industry?

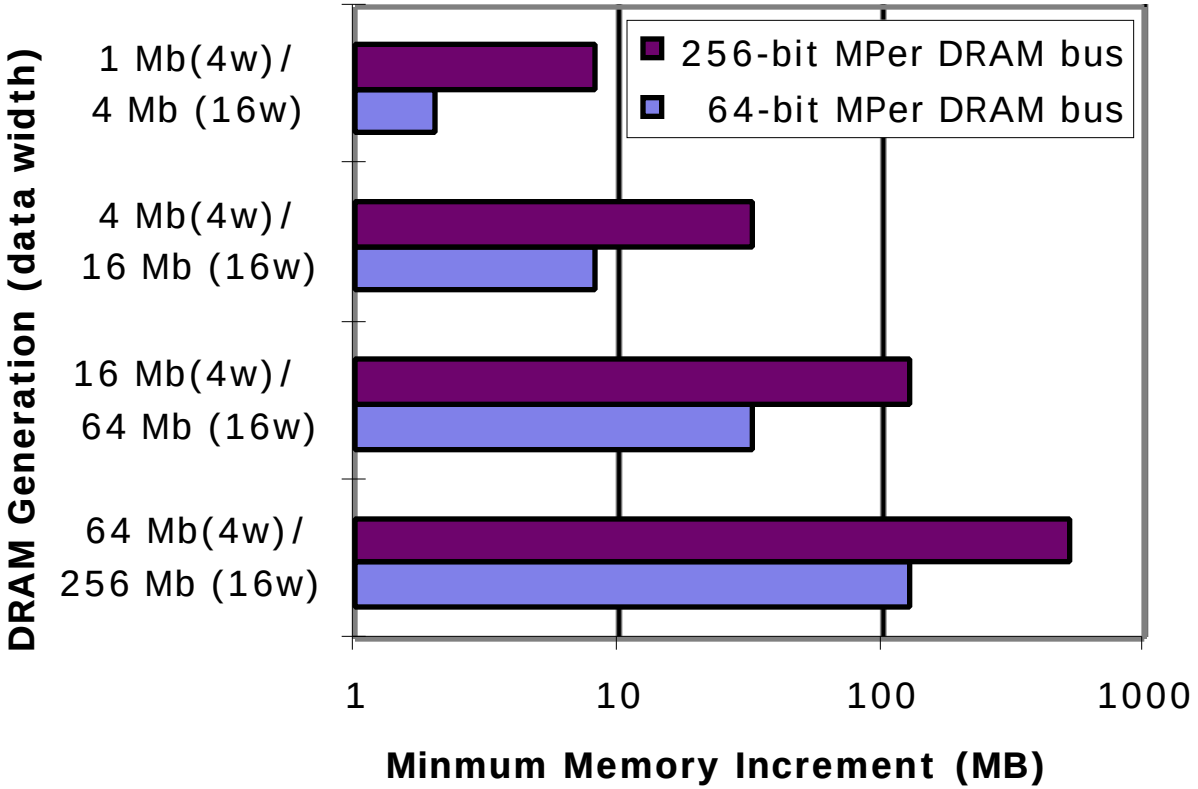
Bus Width, DRAM Width, Minimum Memory Size

 # DRAMs = DRAM bus width / DRAM width



- **4x capacity/DRAM => 4X minimum memory or 4x wider DRAM (and higher cost per bit)**

Minmimum Memory Size vs. DRAM Generation and Width



- **Too large unless wider but wider more expensive per bit (10% more for 16b vs. 4b)**

Memory Width vs. Error Checking

- IBM PC's 75% of DRAM market
- 1 bit Parity/8 bits data is standard, but its optional in SIMMs with change to BIOS
- 64b data + 8 bit parity = 72 bits
- 64b data + 0 bit parity = 64 bits
- 4-bit wide DRAM => 18 or 16 chips
- 16-bit wide DRAM? 32-bit? 64-bit?
- Other systems need Single Error Correction, Double Error Detection:
  - 256b data + 10 bits SEC/DED

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 & smaller 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 worse**
  - **Higher BW/DRAM => higher cost/bit & memory latency/ app BW worse**
- **Caches don't work for all apps**

Potential 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?

Case Study #1: Alpha IRAM

- **Use measurement of existing Alpha to estimate hypothetical Alpha in IRAM**
- **Not optimal for IRAM, but gives one estimate of performance**
- **Use both optimistic and pessimistic slowdown factors for logic and SRAM to bound performance estimate**

Alpha 21164 Performance

Program	SPECint92	SPECfp92	Database	Sparse
CPI	1.2	1.2	3.6	3.0
I$ misses/k instr	7	2	97	0
D$ misses/k instr	25	47	82	38
L2 misses/k instr	11	12	119	36
L3 misses/k instr	0	0	13	23
processor time	0.78	0.68	0.23	0.27
I$ miss time	0.03	0.01	0.16	0.00
D$ miss time	0.13	0.23	0.14	0.08
L2 miss time	0.05	0.06	0.20	0.07
L3 miss time	0.00	0.02	0.27	0.58
Total time	1.00	1.00	1.00	1.00

IRAM Performance Factors

	Optimistic	*Pessimistic*
• Logic	1.3	2.0
• SRAM	1.1	1.3
• DRAM	0.1	0.2

- Apply logic time factor to processor time
  - Although might argue critical path is I or D cache hit time
- Apply SRAM time factor to I & D cache miss time
- Assume L2 cache miss time same speed in IRAM
- Apply DRAM time factor L3 cache miss time

Alpha 21164 Performance

Program	SPECint92		SPECfp92		Database		Sparse	
	Opt.	*Pes.*	*Opt.*	*Pes.*	*Opt.*	*Pes.*	*Opt.*	*Pes.*
• processor time	1.02	1.57	0.89	1.36	0.30	0.46	0.35	0.54
• I$ miss time	0.04	0.05	0.01	0.01	0.18	0.21	0.00	0.00
• D$ miss time	0.14	0.17	0.26	0.30	0.15	0.18	0.08	0.10
• L2 miss time	0.05	0.05	0.06	0.06	0.20	0.20	0.07	0.07
• L3 miss time	0.00	0.00	0.00	0.00	0.03	0.05	0.06	0.12
• Total time	1.25	1.83	1.21	1.74	0.85	1.10	0.56	0.82

(ratio of time vs. alpha; >1 means IRAM slower)

Would IRAM be Fast Enough?

	HP PA-8000	Alpha 21164	Pentium Pro	PPC 604
• SPECint95base	10.8	9.8	8.7	4.6
• SPECfp95base	18.3	12.8	6.0	3.6
• Year	1996	1994	1995	1994
• Clock Rate	180 MHz	400 MHz	200 MHz	133 MHz

- 2 most popular chips are also 2 slowest chips
- Ratio fastest (HP) to slowest (PowerPC):
  - 2.3 for SPECint95base
  - 5.1 for SPECfp95base
- IRAM fast enough even using CPU intensive apps?
- IRAM even better as memory gap increases?

Case Study #2: Vector IRAM

- **Advantages of vector processing**
  - **Fewer instruction fetches**
  - **Indepdent results => deep pipelines, high clock rate**
  - **Multiple pipes/clock => tradeoff HW vs. clock rate**
  - **Amortize memory latency by block access (e.g., 64 words)**
  - **Known access patterns => memory BW via multiple banks**
- **Requirements of vector processing**
  - **Many pipelined functional units: 6 to 16**
  - **Lots of registers: 8x64x64b (32Kb) to 16x128x64b (128Kb)**
  - **Low latency main memory: often SRAM**
  - **Many memory banks: 32 to 1024**
  - **Several memory ports: 3 to 8**

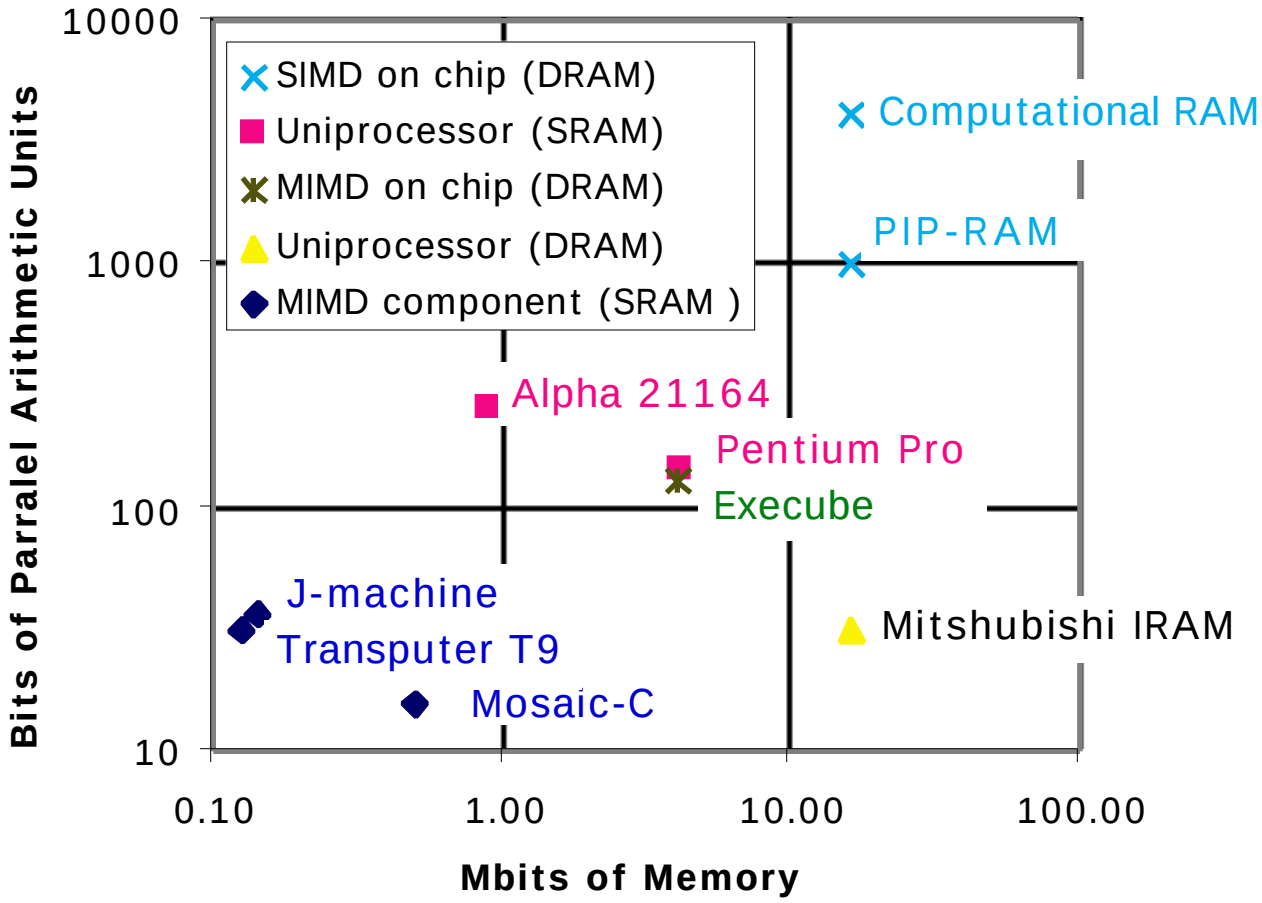
Hypothetical Vector IRAM

- **Gigabit era Vector IRAM (0.15 micron, 600 mm^2 die)**
  - **8 pipelined functional units, 8 pipes/FU @ 500 MHz**
  - **Vector registers: 16x128x64b (128Kb)**
  - **<30 ns main memory latency**
  - **512 memory banks, 1M bits/bank**
  - **16 1024-bit memory ports @ 50 MHz = 100 GB/sec**
- **1000x1000 Linpack**
  - **1.5 GFLOPS on uniprocessor Cray T-90 in 1996**
    - » **6 GFLOPS in 2002?**
  - **Vector IRAM = 8 GFLOPS in 2002?**
    - » **500 MHz * 8 * 2 = 8 GFLOPS in processor**
    - » **(100 GB/sec) / (12B/FLOP) = 8.5 GFLOPS of memory BW**

IRAM Challenges

- **Chip**
  - **Speed, area, power, yield of logic in DRAM process?**
  - **Speed, area, power, yield of SRAM in DRAM process?**
  - **Good performance and reasonable power?**
  - **BW/Latency oriented DRAM tradeoffs?**
- **Architecture**
  - **How to turn high memory bandwidth into performance?**
    - » **Vector?**
    - » **Extensive Prefetching?**
  - **Extensible IRAM: Large pgm/data solution?**
  - **Redudancy in processor to match redundancy in DRAM?**

Ordering of IRAM Projects



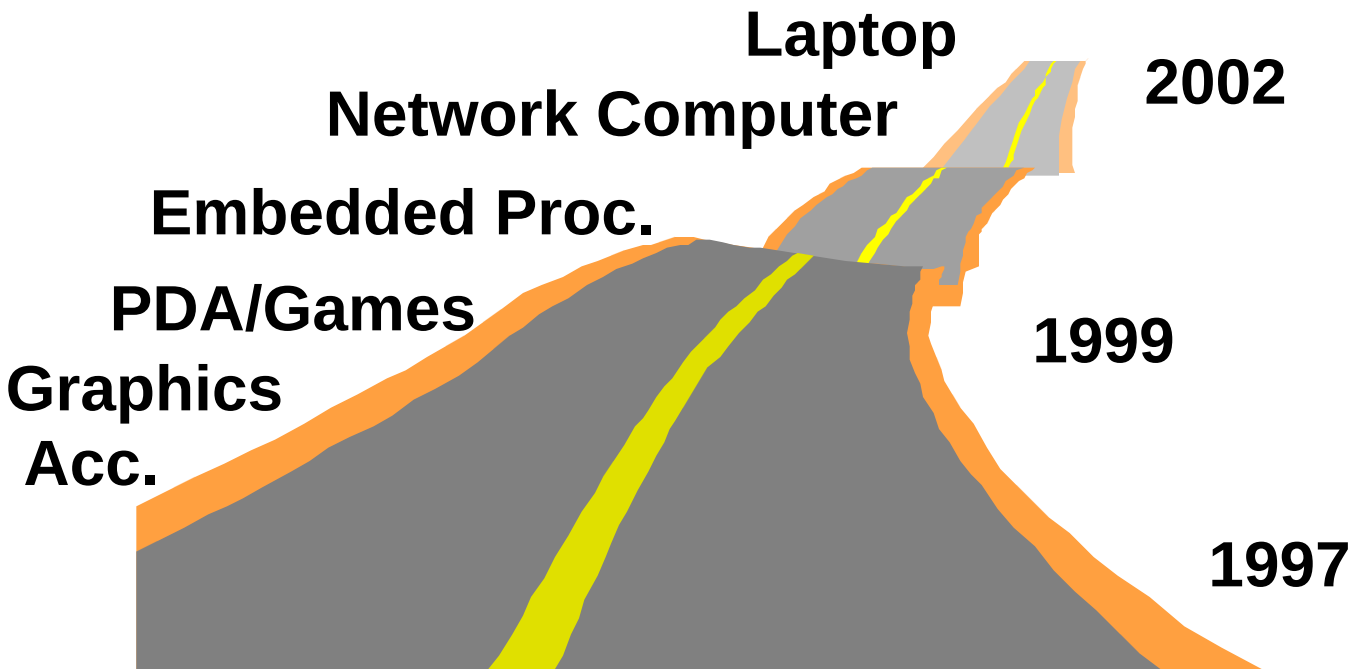
2 IRAM Paths

- **up and right: scale processors and memory**
- **low and right: scale processor speed and memory**

Why might IRAM succeed this time?

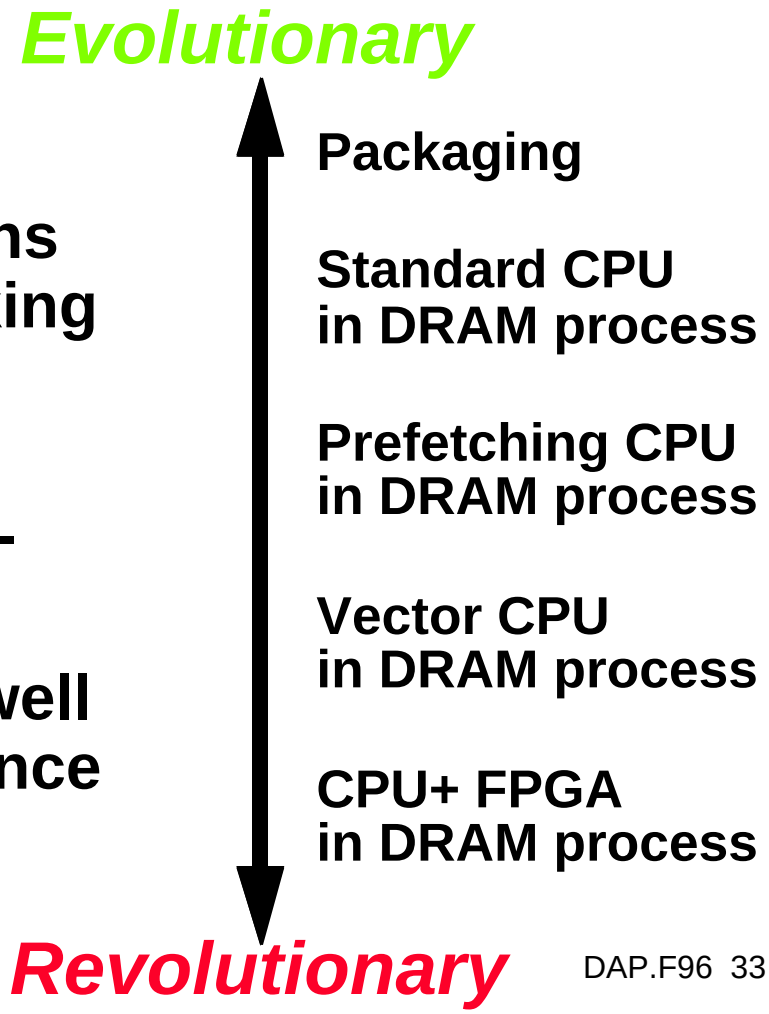
- **DRAM manufacturers facing challenges**
  - **Before not interested, so early IRAM = SRAM**
- **Past efforts memory limited => multiple chips => 1st solve parallel processing**
  - **Gigabit DRAM => 128 MB; OK for many?**
- **Embedded applications offer large 2nd target to conventional computing (business)**

IRAM Roadmap?



IRAM Conclusion

- **Good IRAM applications: predictable access patterns to use 100X bandwidth *or* unpredictable access patterns to use 1/10X latency (& working set too large for caches)**
- **Research challenge is quantifying the evolutionary-revolutionary spectrum**
- **IRAM rewards creativity as well as manufacturing; shift balance of power in DRAM/ microprocessor industry?**



5 minute Class Break

- **Lecture Format:**
  - **≈ 1 minute: review last time & motivate this lecture**
  - **≈ 20 minute lecture**
  - **≈ 3 minutes: discuss class manangement**
  - **≈ 25 minutes: lecture**
  - **5 minutes: break**
  - **≈25 minutes: lecture**
  - **≈1 minute: summary of today's important topics**

Doing Research in the Information Age

- **Online at UCB**
  - **Finding articles**
    - » **INSPECT database**
    - » **COMP database**
  - **Printing IEEE articles**
  - **Finding Books: MELVYL and GLADIS**
- **WWW Search Engines**
  - **Alta Vista, HotBot, Yahoo!**
- **Computer Architecture Resources**
  - **Architecture Homepage, Benchmark Database...**

INSPECT Database

- **Finding articles**
  - **Dates**
  - **Authors**
  - **Author's Institution**
  - **Subject**
  - **Type of publication**
- **Viewing abstracts**
- **Mailing results**
  - **saving to list**
- **Printing IEEE papers**

COMP Database

- **Has full text of many articles!**
- **Types of publications: trade magazines**
  - **Includes Microprocessor Report!**

MELVYL and GLADIS

- **MELVYL: Finding books in UC system**
- **GLADIS: Is book at Berkeley available?**
- **MELVYL**
  - **Finding by author**
  - **FInding by subject**
  - **Adding to search**
  - **Mailing results**
- **GLADIS**
  - **checking status of books**

WWW Search Engines

- **What to look for on WWW:**
  - **Latest version things that vary over time: products, goverment indexes, …**
  - **What research is on-going: technical reports, project home pages, pre-released papers**
  - **Serendipity: trying to see what's out there on topic, trying to find a person**
  - **Other suggestions?**
- **What *not* to look for on WWW:**
  - **Historical record of things that vary over time**
  - **Research projects long completed**
  - **(thus far) refereed, authoratative publications**
  - **Other suggestions?**

WWW Search Engines

- **Alta Vista, Hot Bot**
  `http://altavista.digital.com`
  `http://www.hotbot.com`
  - **Full text index of HTML on WWW**
  - **Also of USENIX interest groups**
- **Yahoo!**
  `http://www.yahoo.com`
  - **Table of Contents of the WWW**

Computer Architecture and WWW

- **Computer Architecture Home Page**
  `http://www.cs.wisc.edu/~arch/www`
- **Benchmark Database**
  `http://performance.netlib.org/`
- **comp.arch interest group**

Academic Scholarship and the WWW?

- **Berkeley, MIT, Stanford values impact**
- **Traditional academic "Coin of the Realm" is journal publications (benchmark problem?)**
  - **What is the ratio of readers to writers for journal publications?**
  - **What would happen to professional societies if went to "pay per view" for jounral publications?**
  - **Membership in all professional societies is dropping; lowered interest in journals?**
- **What will research evaluation metric become?**