Intelligent RAM (IRAM): Chips that remember and compute

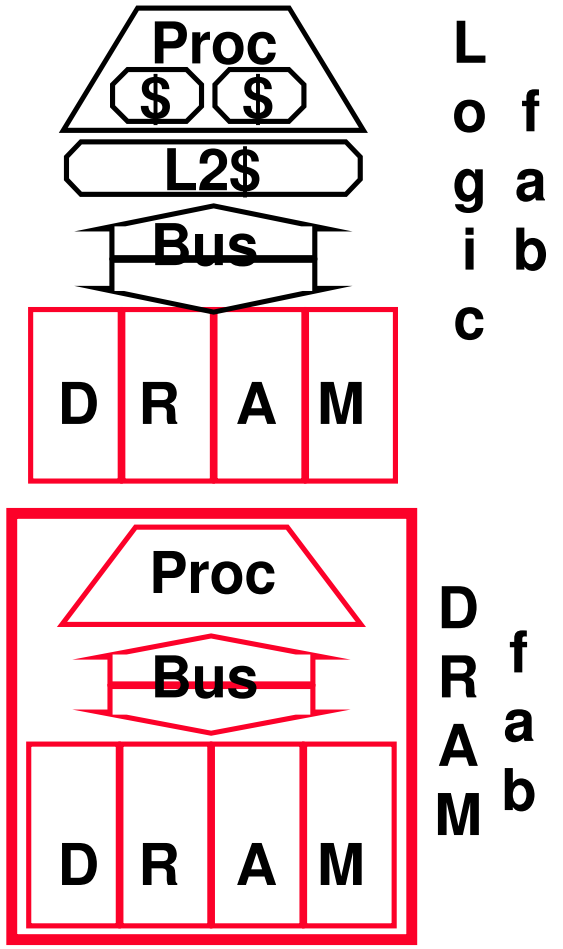
David Patterson, Thomas Anderson,
Neal Cardwell, Richard Fromm,Kimberly Keeton,
Christoforos Kozyrakis, Randi Thomas, and
Katherine Yelick

Computer Science Division
University of California
Berkeley, CA 94720-1776

IRAM Vision Statement

Microprocessor & DRAM on single chip:

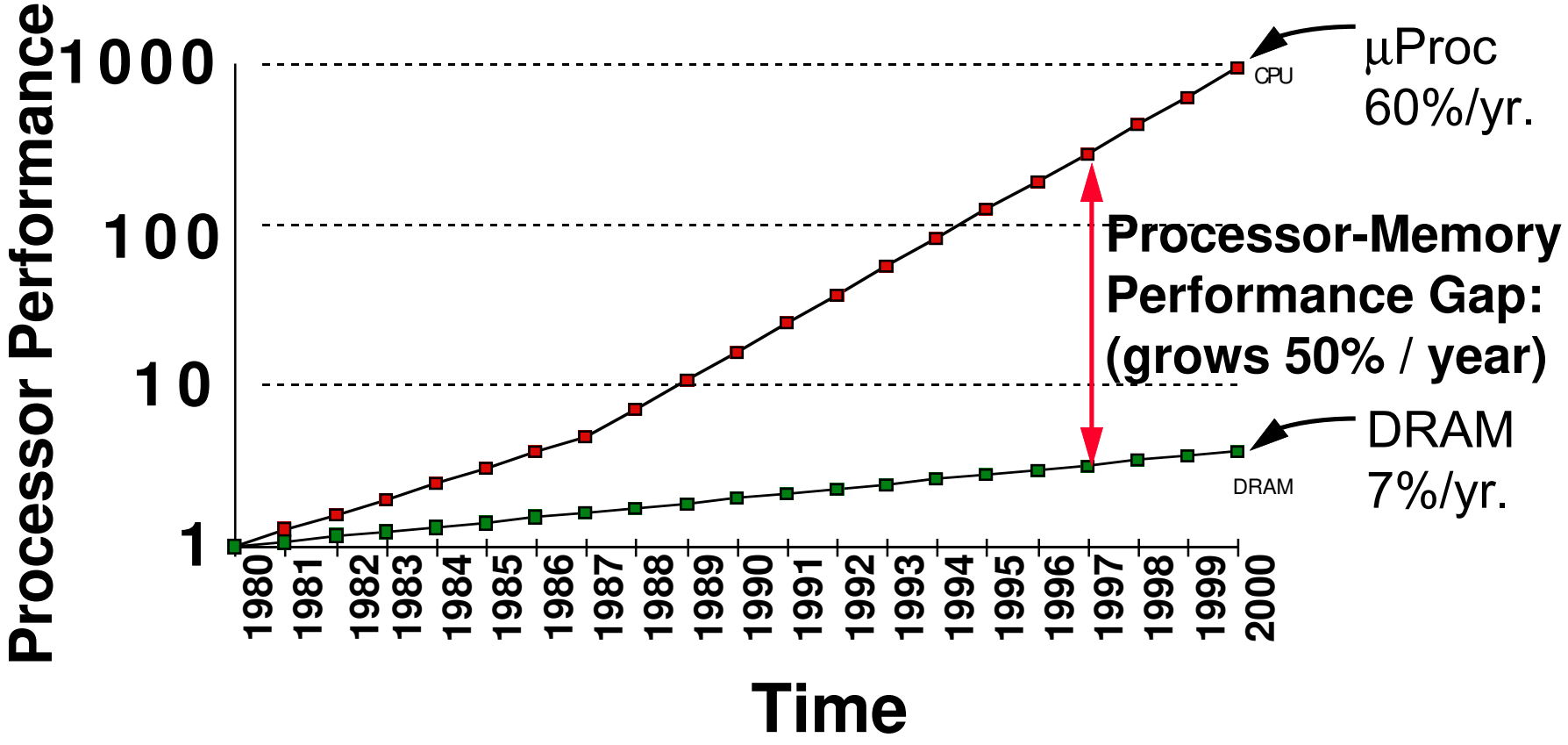
- » bridge processor-memory performance gap via on-chip latency 5-10X,bandwidth 100X
- » improve energy efficiency 2X-4X (no DRAM bus)
- » adjustable memory size/width (designer picks any amount)
- » smaller board area



Outline

- Today's Situation: Microprocessor
- Today's Situation: DRAM
- IRAM Opportunities and Challenges
- IRAM Options
- Related Work
- IRAM Potential 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
  - » full cache miss time = 100s instructions

  1st Alpha (7000): 340 ns/5.0 ns = 68 clks x 2 or136

  2nd Alpha (8400): 266 ns/3.3 ns = 80 clks x 4 or 320

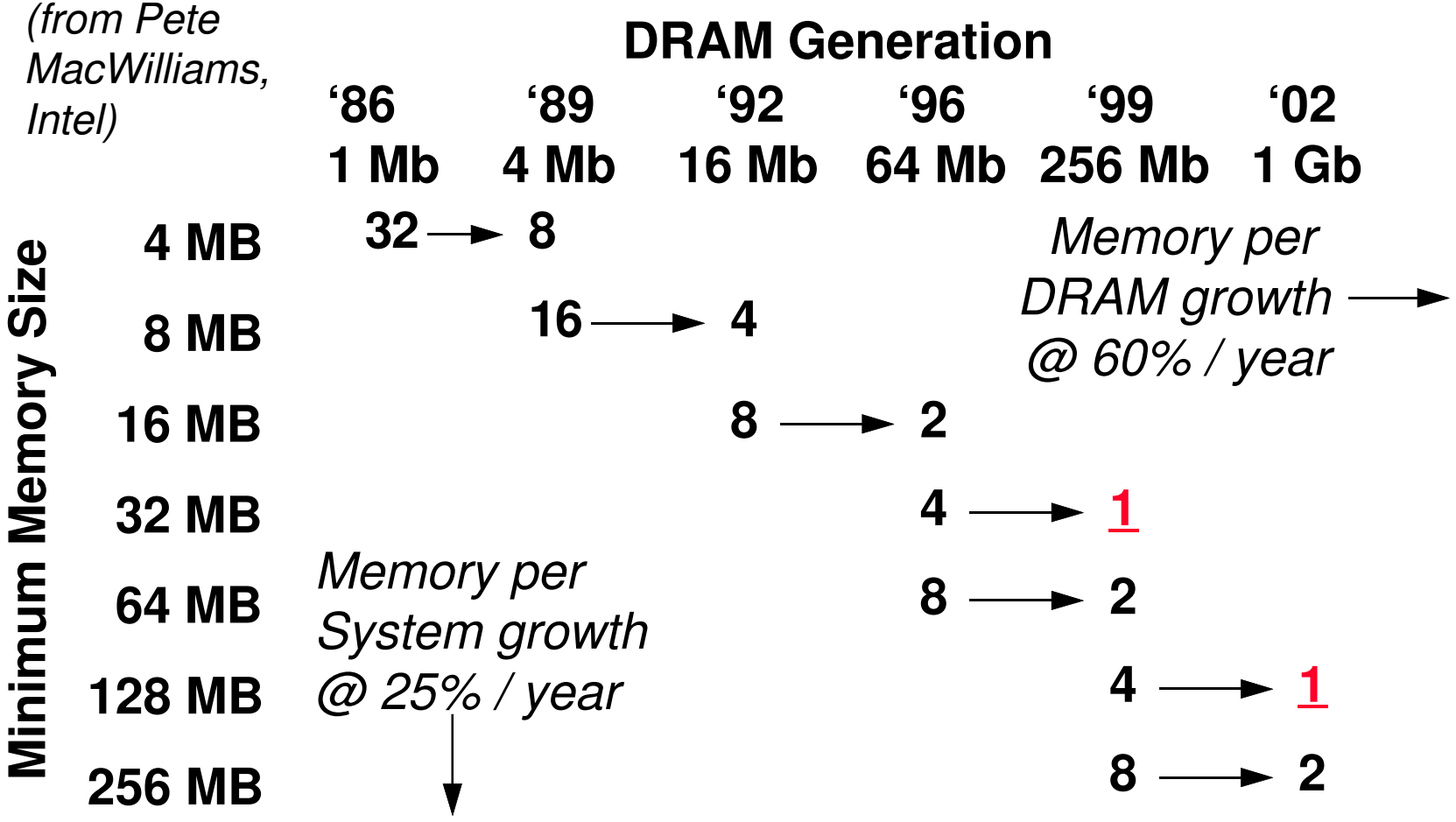
  3rd Alpha (): 180 ns/1.7 ns =108 clks x 6 or 648
- Rely on caches to bridge gap
  - » Doesn't work well for some apps: data bases, …
- Power limits performance (battery, cooling)

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
- Fewer DRAMs per computer over time
- Starting to question buying larger DRAMs?

Fewer DRAMs/System over Time

(from Pete MacWilliams, Intel)

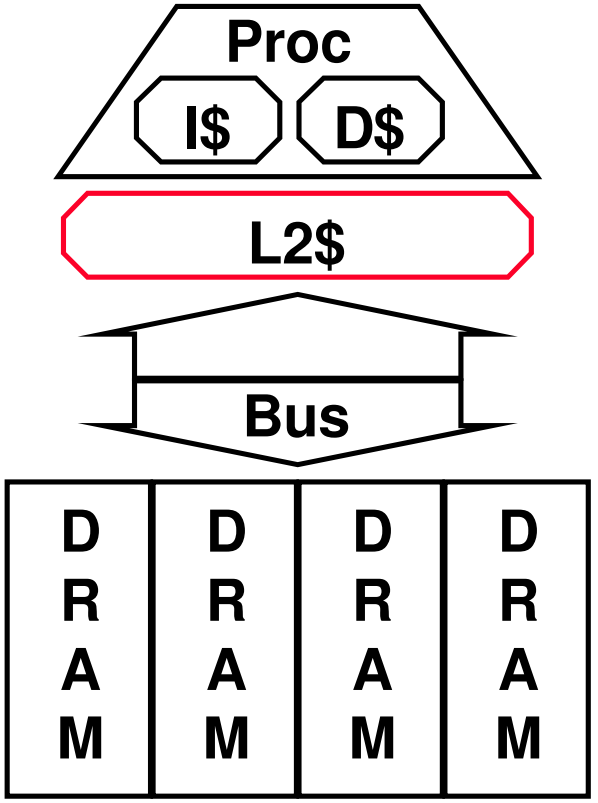


Reluctance for New DRAMs: Proc. v. DRAM BW, Min. Mem. size

- 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

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

- Gap means performance limit is memory
- Some apps: energy, board area, memory size
- Dwindling interest in future DRAM:256Mb/1Gb?
  - » Higher capacity/DRAM
    => system memory bandwidth worse
  - » Industry supplies higher bandwidth / DRAM
    => higher latency (& cost/bit)=> app bandwidth worse
- Alternatives: packaging breakthru, 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
  AlphaSer. 8400: 266 ns=256b, 280 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 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 Energy Efficiency: 2X-4X

- Case study of StrongARM memory hiearchy 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 optimisitic and pessimistic factors for logic (1.3-2.0 slower), SRAM (1.1-1.3 slower), DRAM speed (5X-10X faster)
  - » 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

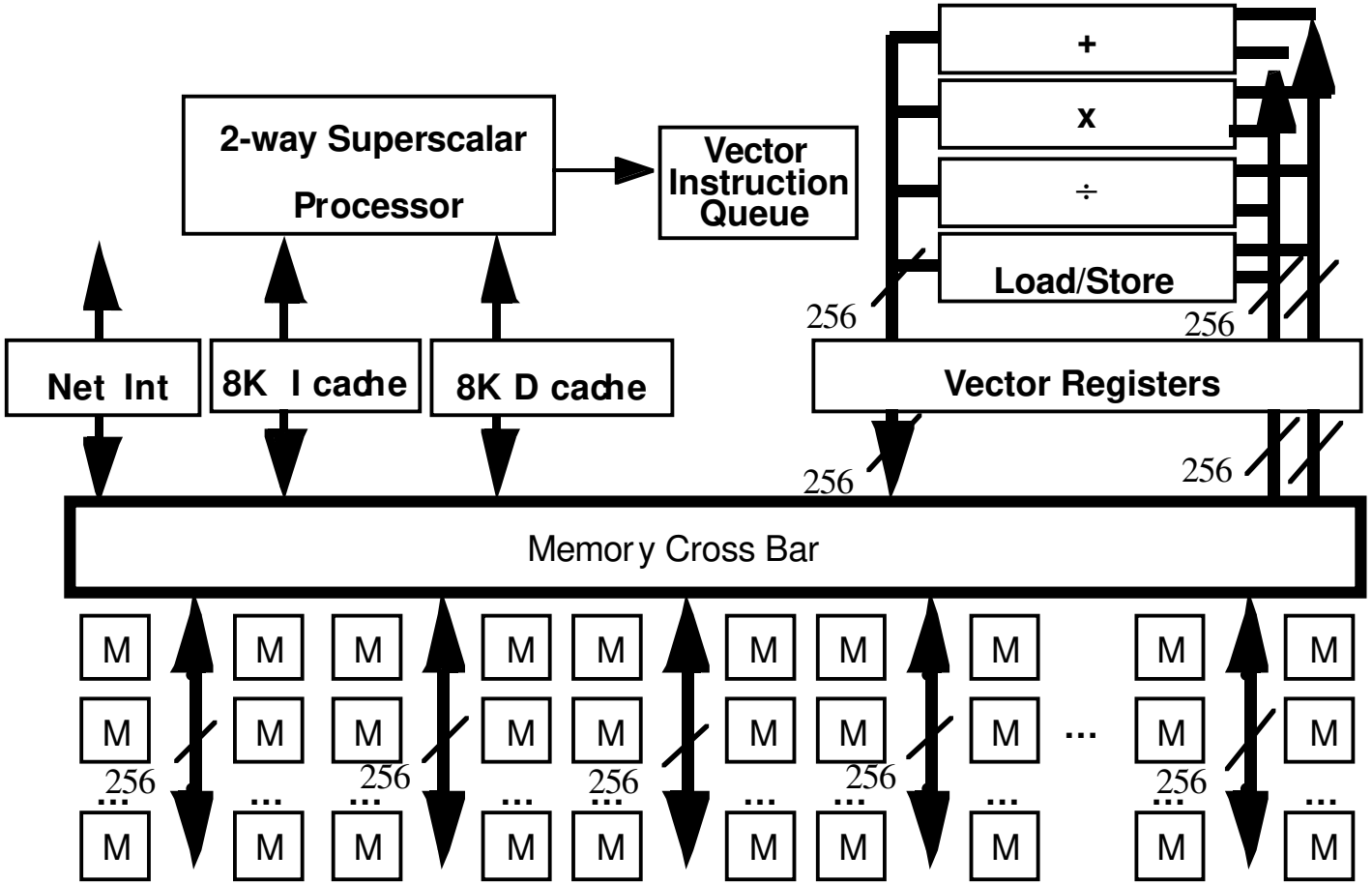
A More Revolutionary Approach

- Faster logic in DRAM process
  - » DRAM vendors offer same fast transistors + same number metal layers as good logic process? @ 10% - 30% higher cost per wafer?
- 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 Architecure 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

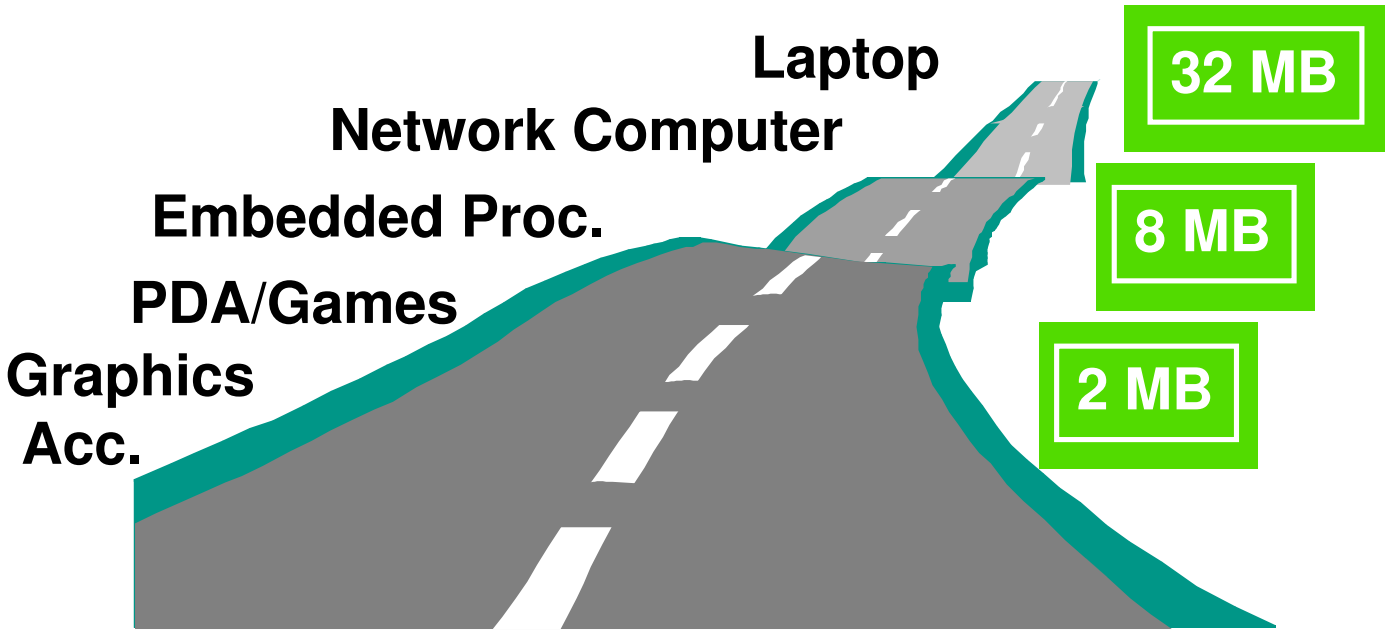
0.25 μm, Fast Logic IRAM: ≈500MHz
5 GFLOPS(64b)/40 GOPS(8b)/24MB



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 => 1st solve the unsolved (parallel processing)
  - » Gigabit DRAM => 128 MB; OK for many apps?
- Embedded apps leverage energy efficiency, adjustable mem. capacity, smaller board area => large alternative market to conventional computing (viable business model?)

Commercial IRAM highway is governed by memory per IRAM?



IRAM Challenges

- Chip
  - » Speed, area, power, yield in DRAM process?
  - » Good performance and reasonable power?
  - » BW/Latency oriented DRAM tradeoffs?
  - » Testing Time of IRAM vs DRAM vs μP?
- Architecture
  - » How to turn high memory bandwidth into performance for real applications?
  - » Extensible IRAM: Large pgm/data solution?

IRAM Conclusion

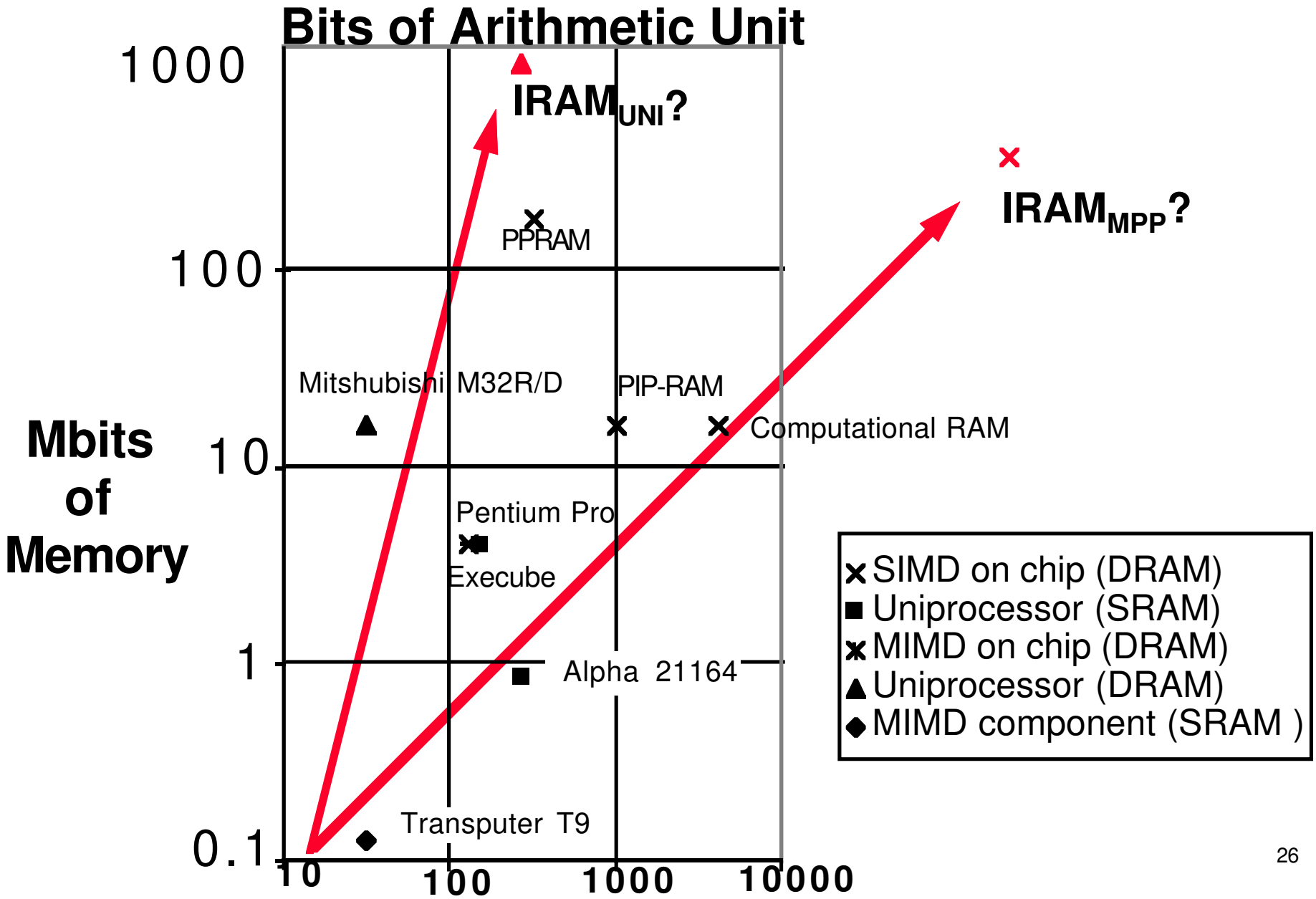
- IRAM potential in performance, energy, capacity, board area; challenges in yield, power, testing
- IRAM rewards creativity + manufacturing
- Potential shift in balance of power in DRAM/ microprocessor (μP) industry in 5-7 years?
  - μP-oriented vs. DRAM-oriented manufacturers:
  - who ships the most DRAM memory?
  - who ships the most μPs?

Interested in Helping?

- Spice parameters for advanced DRAM and Logic fabs in modern technology
- Fab of test chips
- Design/fab prototype
- Contact us if you're interested:
  `http://iram.cs.berkeley.edu/`
  `email: patterson@cs.berkeley.edu`
- Thanks for advice and/or support: DARPA, Intel, Neomagic, SGI/Cray, Sun Microsystems

Backup Slides

(The following slides are used to help answer questions)


Bits of Arithmetic Unit
1000
100
10
1
0.1
Mbits of Memory
10
100
1000
10000
IRAM_UNI?
IRAM_MPP?
PPRAM
Mitshubishi 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)

Why Vector? Isn't it dead?

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

- Single-chip CMOS microprocessor/IRAM
- Small % in future + scales as no. transitors increase
- IRAM = low latency, high bandwidth memory
- Multimedia apps (MMX) are vectorizable too
- Include modern, modest CPU => scalar performs OK-good