
High-Performance Architectures for Embedded Memory Systems

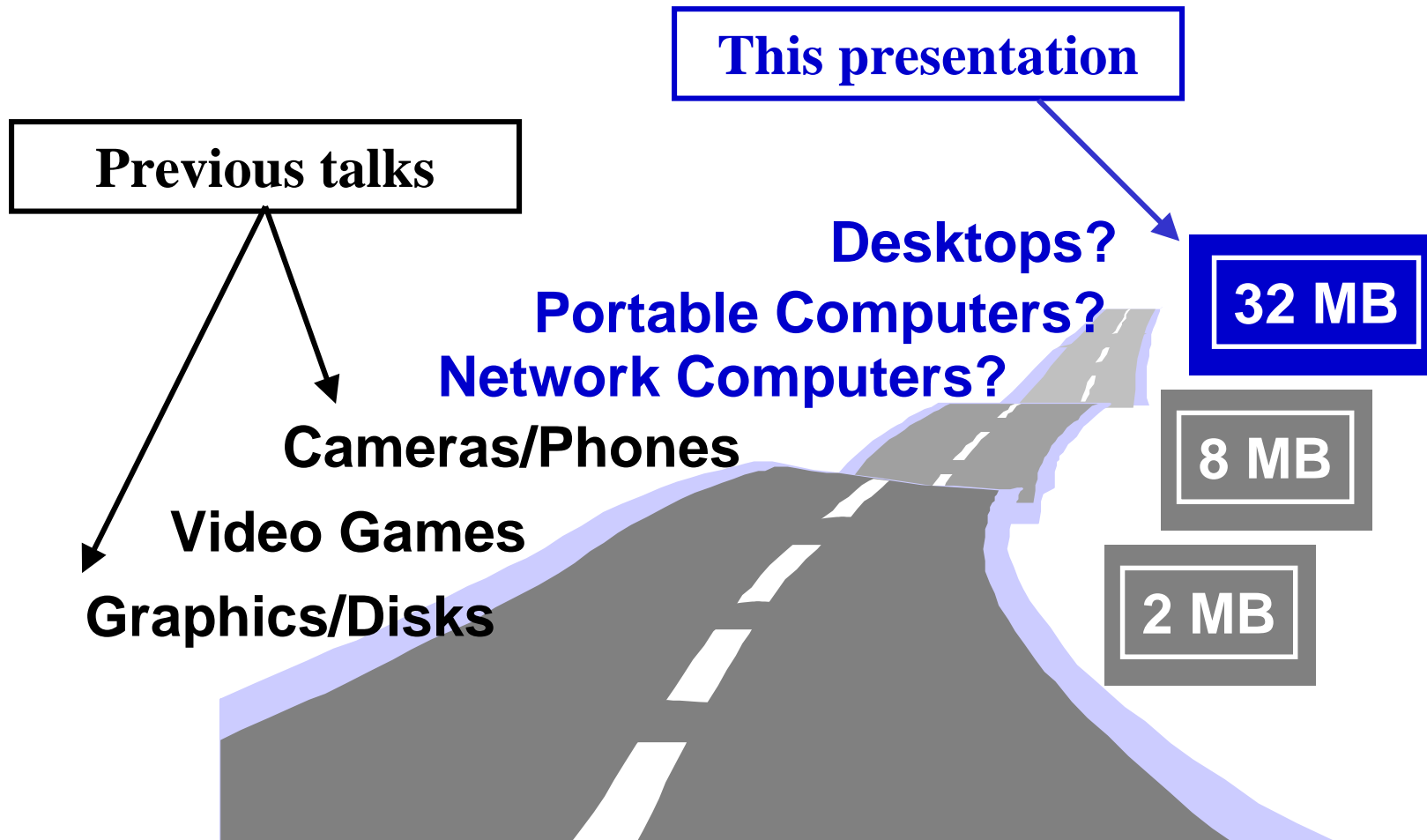


Christoforos E. Kozyrakis

Computer Science Division
University of California, Berkeley

`kozyraki@cs.berkeley.edu`
`http://iram.cs.berkeley.edu/~kozyraki`

Embedded DRAM systems roadmap



Outline

- Overview of general-purpose processors today
- Future processor applications & requirements
- Advantages and challenges of processor-DRAM integration
- Future microprocessor architectures
 - characteristics and features
 - compatibility and interaction with embedded DRAM technology
- Comparisons and conclusions

Current state-of-the-art processors

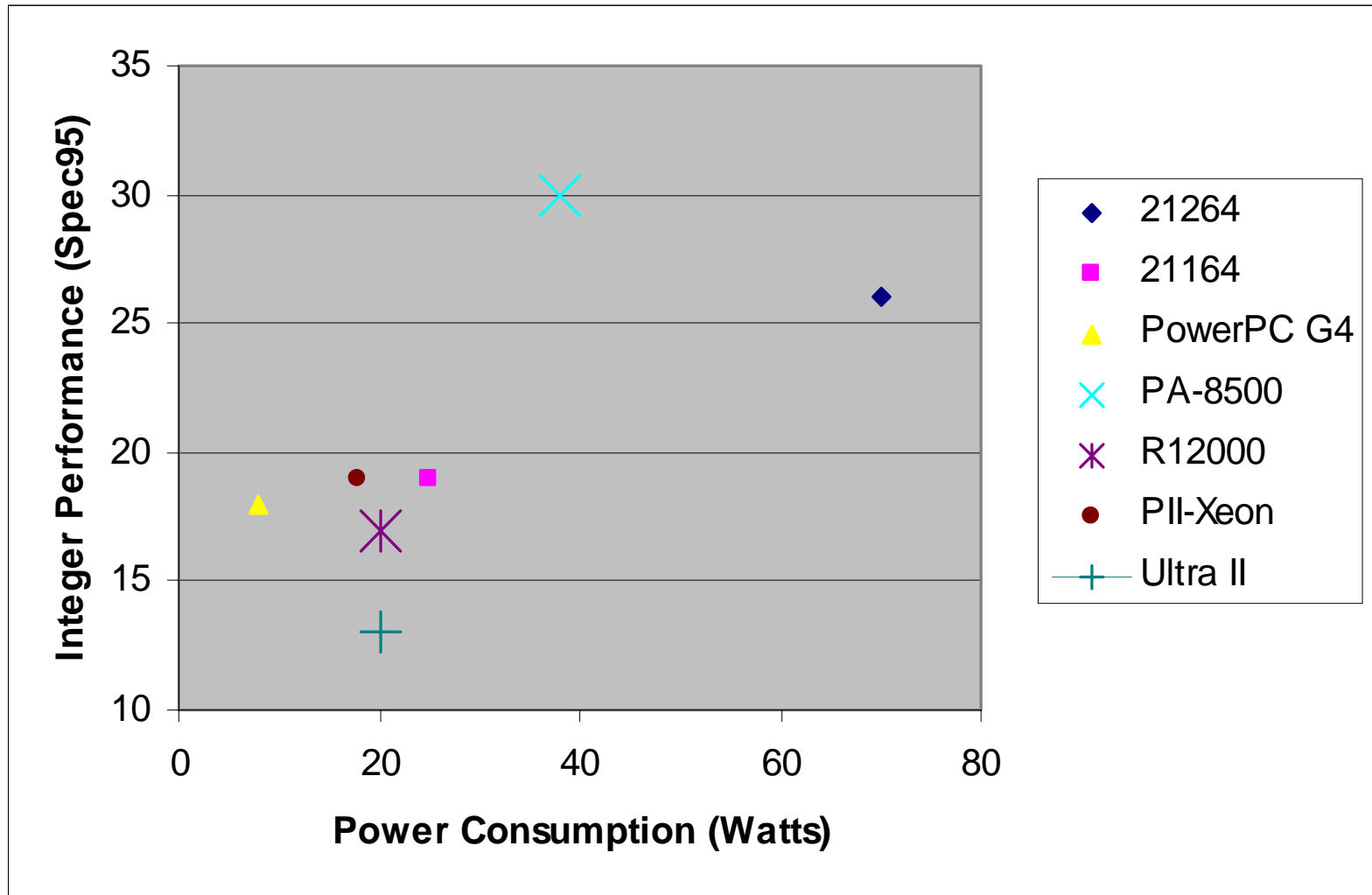
	High-Performance	Embedded
Data width	64 bit	32 or 64 bit
Issue width	3-6	1-2
Execution	out-of-order	in-order
Cache	large multi-level	small single-level
Additional features	parallel systems support SIMD extensions	integrated I/O controllers on-chip DRAM
Metrics	peak performance	price/performance MIPS/Watt code density
Frequency	300 to 600 MHz	40 to 400 MHz
Die area	200 to 300 sq. mm	10 to 100 sq. mm
Power	20 to 80 Watts	0.3 to 4 Watts
Examples	Alpha 21264, MIPS R12K, Pentium III, Sparc III	ARM-9, MIPS R5K, M32R, SH-4

Current microprocessor applications

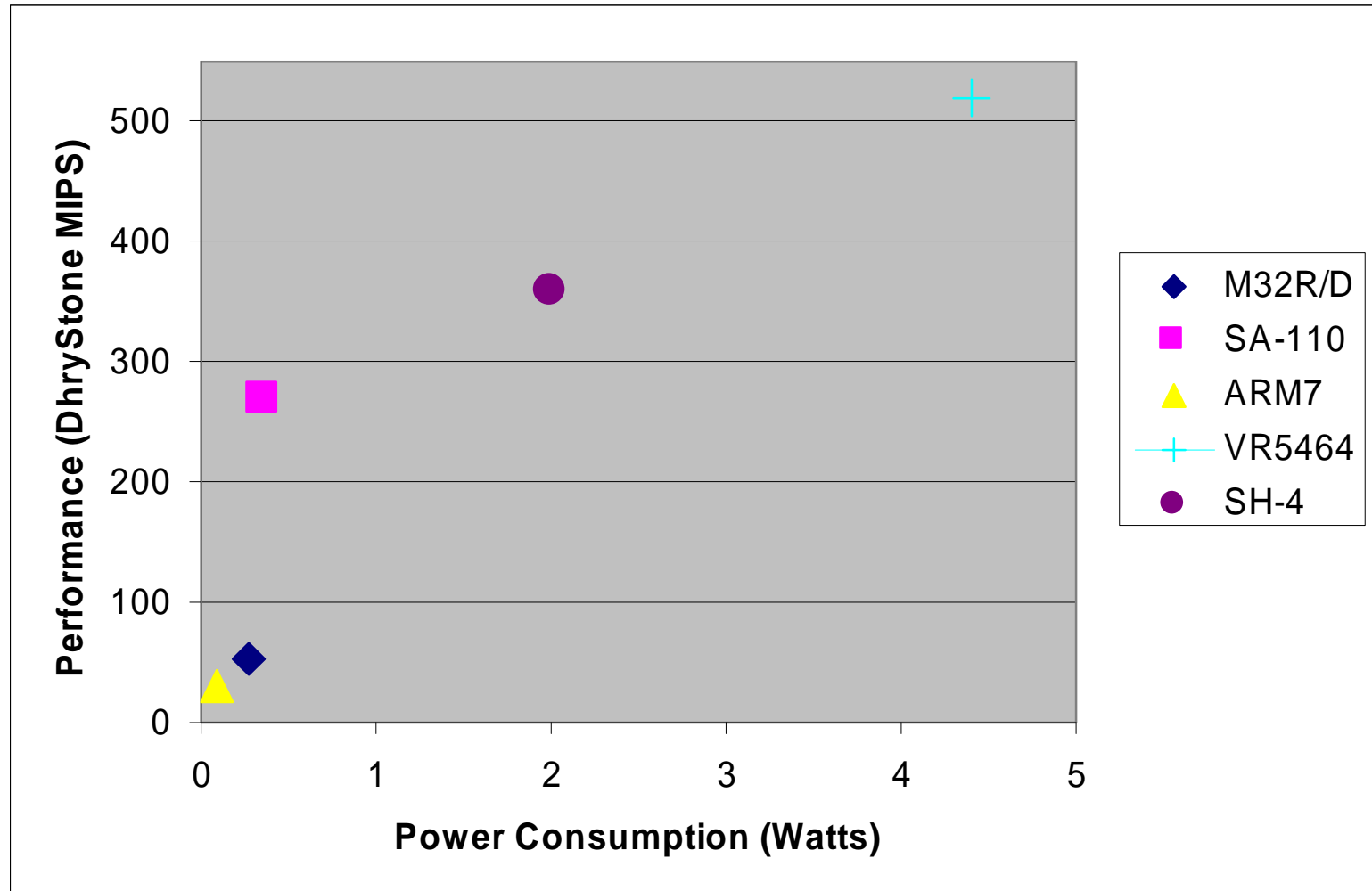
	High-Performance	Embedded
Major applications	<u>Desktop</u> technical workloads (CAD) office productivity tools	postscript printers digital cameras networking hardware PDAs disk drives
	<u>Server</u> file systems transaction processing decision support	
Benchmarks	SPEC95 (Int/FP) TPC-C/D (OLTP, decision support)	DhryStone



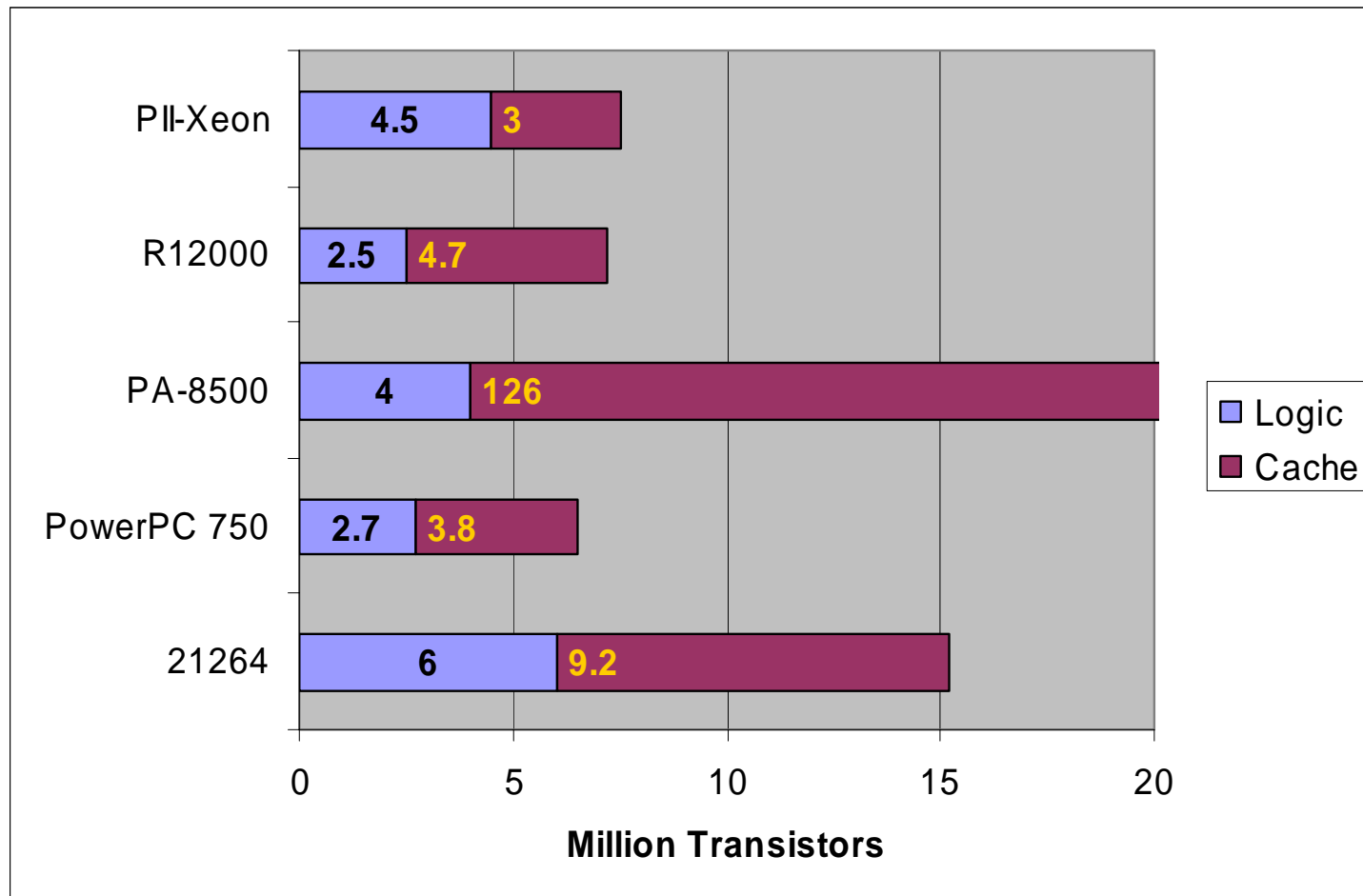
High-performance microprocessors: power/performance



Embedded microprocessors: power/performance



High-performance microprocessors: transistor use




Future microprocessor applications

Mobile multimedia computing

- A single device is: PDA, video game, cell phone, pager, GPS, tape recorder, radio, TV remote...
- Basic interfaces: voice (speech recognition) and image (image/video processing)
- Small size, battery operated devices
- Media processing functions are the basic workload



Requirements on microprocessors

- High performance for multimedia:
 - real-time performance guarantees 
 - support for continuous media data-types
 - fine-grain parallelism
 - coarse-grain parallelism
 - high instruction reference locality, code density
 - high memory bandwidth
- Low power/energy consumption
- Low design/verification complexity, scalable design
- Small size/chip count

Embedded DRAM advantages summary

- High memory bandwidth
- Low memory latency
 - slower than on-chip SRAM though
- Energy/power efficiency
- System size benefits
- Custom configuration
 - interface width, size, number of banks etc

Embedded DRAM challenges summary

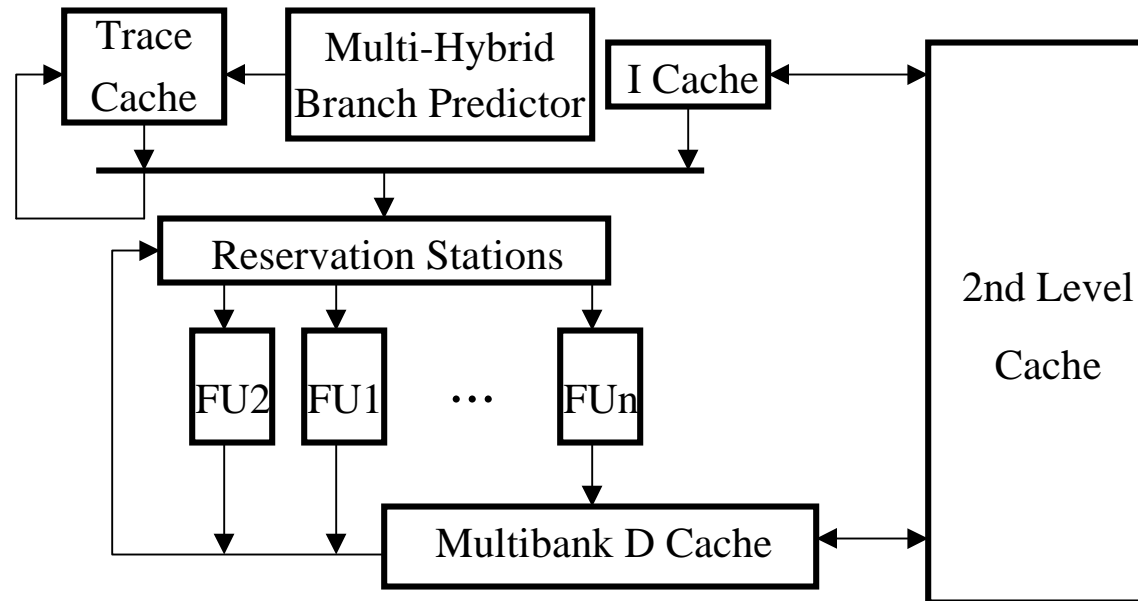
- eDRAM cost
 - wafer cost
 - cost per DRAM bit (density)
 - yield
 - cost of testing
- Performance, power and yield of logic components
- Complexity
- Optimal memory organization

Trends in high-performance architecture

- Advanced superscalar processors
- VLIW: Very long instruction word processors (IA-64/EPIC)
- Single chip multiprocessors
- Reconfigurable processors
- Vector microprocessors (Vector IRAM)


Advanced superscalar processors

- Scale up current designs to issue more instructions (16-32)



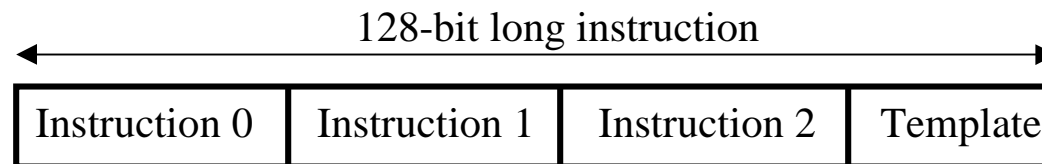
- Major features:
 - dynamic instruction scheduling in hardware, out-of order execution
 - branch/dependence/stride/data/trace prediction buffers
 - large multibank caches

Advanced superscalar processors (2)

- Advantages
 - dynamic scheduling exploits run-time info
 - software compatibility
 - high-performance for current desktop applications
- Disadvantages
 - relies on high-speed logic and fast, large caches
 - unpredictable performance (high misprediction cost)
 - limited media processing support (MMX-like units) 
 - high design/verification complexity
 - high power consumption due to extensive speculation
- eDRAM perspective
 - cannot fully utilize available eDRAM bandwidth
 - eDRAM “unfriendly” environment (power, complexity, size)
 - eDRAM for second-level cache?

VLIW processors (EPIC)

- Very long instruction word scheme



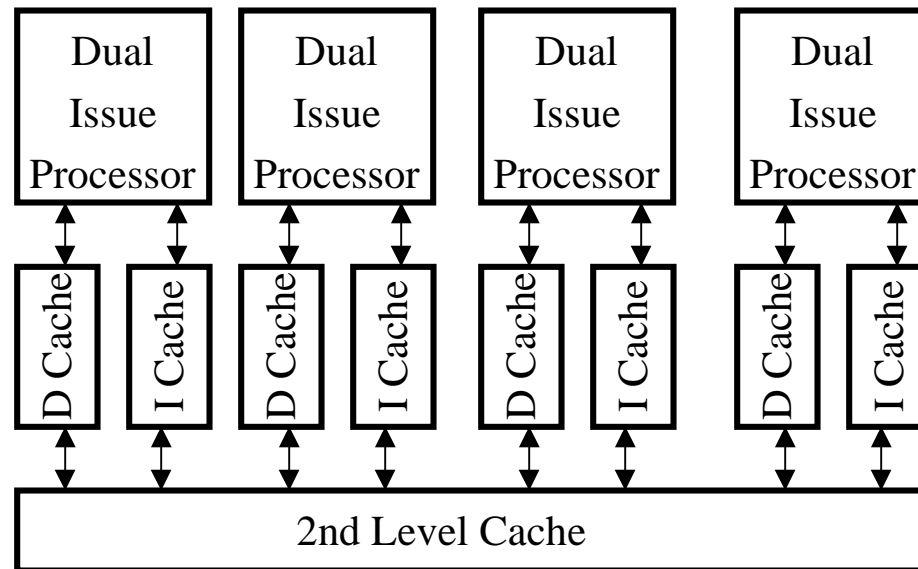
- Major features
 - instruction scheduling by compiler (dependence analysis, register renaming etc)
 - template carries dependence information
 - large number of registers
 - multiple functional units
 - software speculation and predicated (conditional) execution
 - cache based designs

VLIW processors (EPIC) (2)

- Advantages
 - simpler hardware
 - highly scalable
- Disadvantages
 - ability of compiler to optimize for performance (lack of run-time information)?
 - code size (loop unrolling, software pipelining)
 - software compatibility
 - limited media processing support (MMX-like units)
- eDRAM perspective
 - cannot fully utilize available eDRAM bandwidth
 - requires high-speed logic to make up for run-time information
 - eDRAM for second-level cache?

Single chip multiprocessors

- Place multiple processors on a single chip



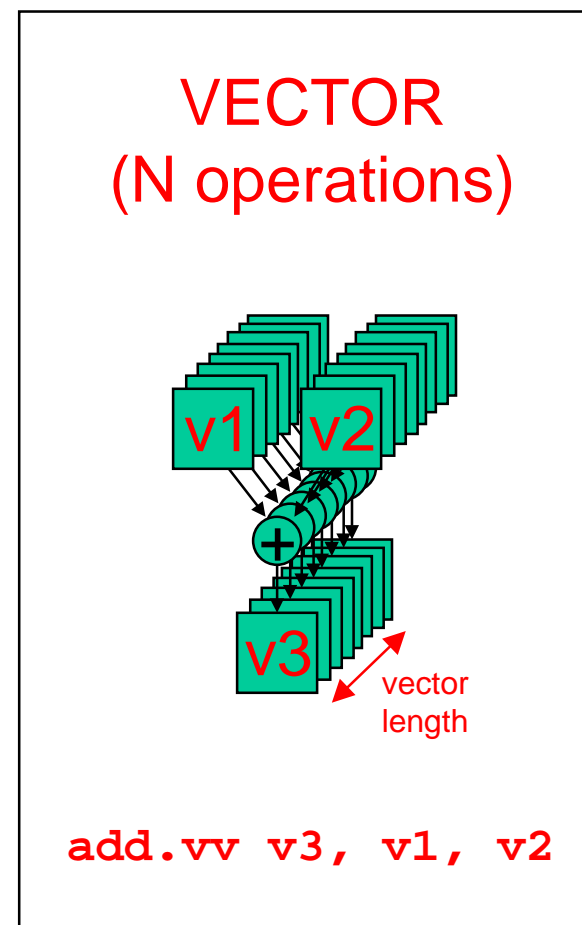
- Major features
 - symmetric multiprocessor system (shared memory system)
 - shared second-level cache
 - 4 to 8 uniprocessors, similar to current out-of-order designs

Single chip multiprocessors (2)

- Advantages
 - modular design
 - coarse-grained parallelism
- Disadvantages
 - difficulty of efficient parallel programming
 - lack of parallelizing compilers
 - high power consumption
 - complexity of shared-memory protocols
- eDRAM perspective
 - can utilize bandwidth of multi-bank eDRAM
 - inherent redundancy
 - multiprocessors require large amount of memory





Vector microprocessors

- Vector instructions
 - for $i=1$ to N : $v3[i]=v1[i]+v2[i]$
- Major features
 - vector coprocessor unit
 - instructions define operations on vectors (arrays) of data
 - vector register file
 - strided and indexed memory accesses
 - support for multiple data widths
 - support for DSP/fixed-point
 - conditional/speculative execution support through flag registers



Vector microprocessors (2)

- Advantages

- predictable performance: in-order model, no caches
- high performance for media processing 
- simple design: no complex issue/speculation logic
- low power/energy consumption 
- performance through parallel pipelines, not just clock frequency 
- scalable 
- small code size: single instruction loops

- Disadvantages

- cannot utilize random instruction-level or thread-level parallelism; just fine-grain parallelism
- poor performance for many current desktop applications
- requires high-bandwidth memory system

Vector processors and eDRAM

- Vector processors require multi-bank, high-bandwidth memory system:
 - multiple wide DRAM banks, crossbar interconnect
- Vector processors can tolerate DRAM latency
 - delayed or decoupled vector pipelines
- eDRAM friendly environment
 - low power, low complexity, modest clock frequencies
- eDRAM testing
 - use vector processor as BIST engine; 10x faster than scalar processors
- Logic redundancy
 - use a redundant vector pipeline

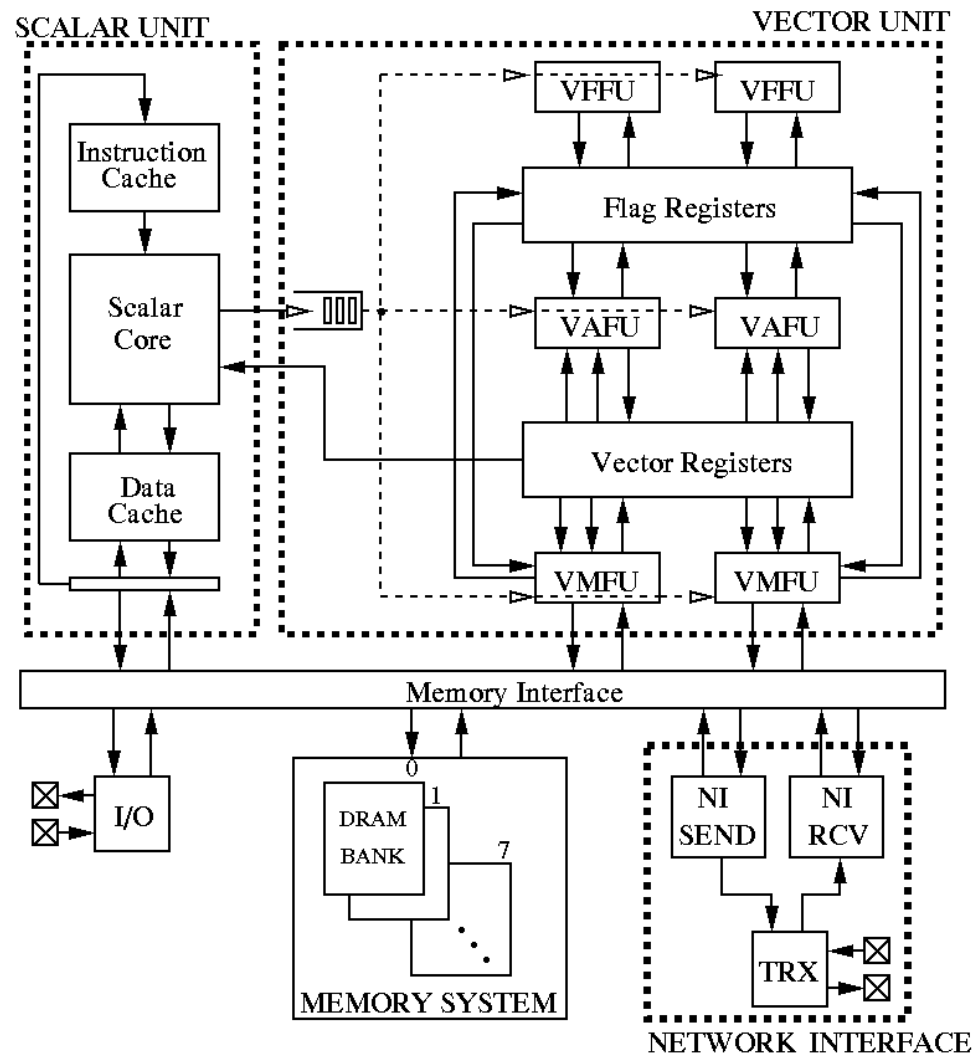
Architecture combinations

- Vector and SIMD can be combined with superscalar, VLIW and chip multiprocessor
 - Examples: x86 with media extensions, Philips Tri-media
- Goal: get the best of both worlds
 - superscalar or VLIW for irregular parallelism
 - vector or SIMD for regular (data) parallelism
- Open questions
 - what is the right processing power balance
 - how to achieve efficient interaction between the two components

Vector IRAM-1

- Scalar core
 - 2-way superscalar MIPS
 - 16KByte I/D caches
- Memory system
 - 24 MBytes DRAM
 - multi-bank memory system
 - 256-bit synchronous interface
 - crossbar interconnect for 25.6 GB/sec aggregate bandwidth
- Vector coprocessor
 - 8-KByte vector register file, 32 vector registers
 - supports 64b, 32b, 16b data
 - 2 arithmetic, 2 load/store, 2 flag processing units
 - 4 64-bit pipelines per functional unit
 - separate multi-ported TLB
 - fixed-point arithmetic support
 - no caches; pipeline enhanced to tolerate DRAM latency

Vector IRAM-1 Block Diagram



VIRAM-1 Technology Summary

Technology: **0.18 micron embedded DRAM-logic process**

Memory: **24 MBytes**

Die size: **400 mm²**

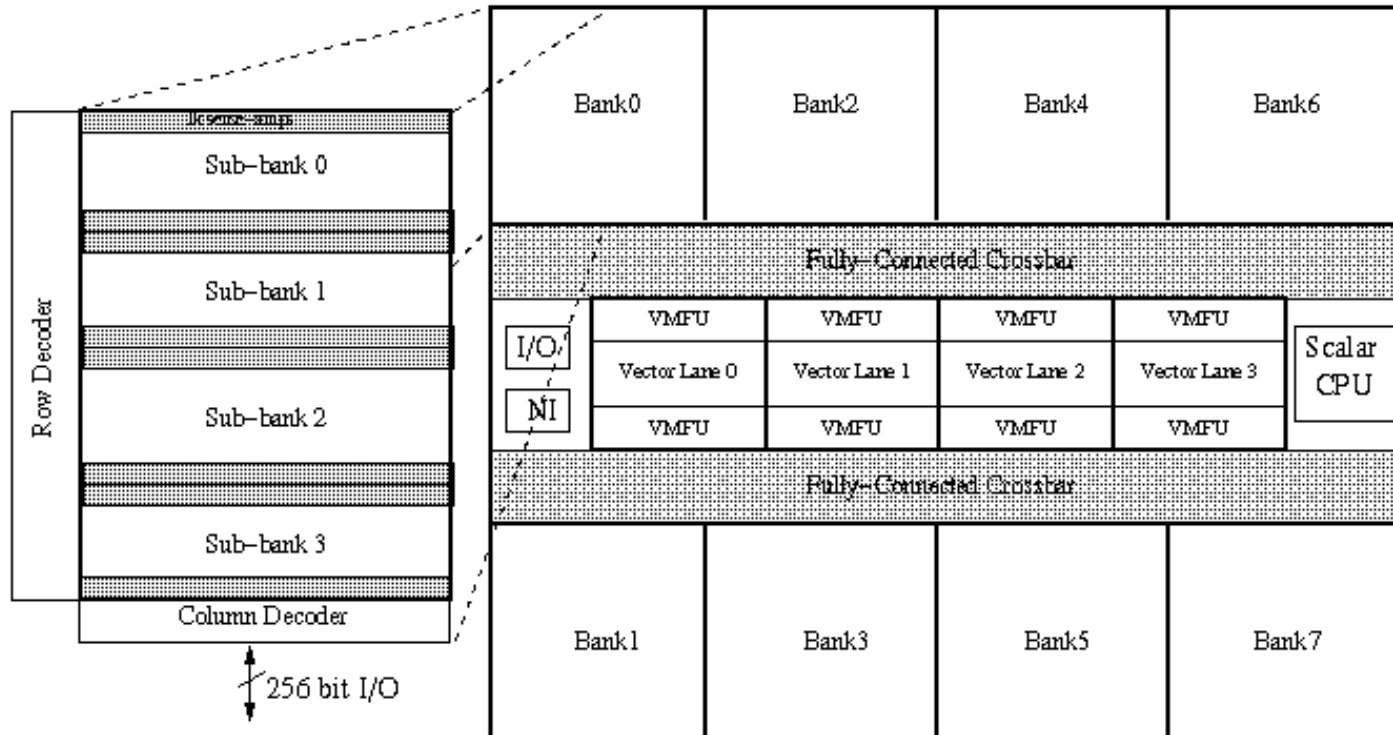
Vector pipelines: **4 64-bit (or 8 32-bit or 16 16-bit)**

Clock Frequency: **200MHz scalar, vector, DRAM**

Power: **~2 W**

Performance: **3.2 GFLOPS₃₂ – 6.4 GOPS₁₆**

VIRAM-1 Floorplan



Comparison: current desktop and servers

	SS	VLIW	CMP	RC	VIRAM
SPEC Int	■	■	□	□	■
SPEC FP	■	■	■	□	■
TPC (DB)	□	□	■	■	□
SW Effort	■	□	□	■	□
Design Scalability	□	□	□	□	□
Design Complexity	■	□	□	■	□

Legend: ■ strength, □ neutral, ■ weakness

Comparison: mobile multimedia computing

	SS	VLIW	CMP	RC	VIRAM
Real-time perf.	Red	White	White	White	Green
Cont. data support	White	White	White	White	Green
Energy/Power	Red	White	White	Red	Green
Code Size	White	Red	White	White	Green
Fine-grain parall.	White	White	White	Green	Green
Coarse-grain parall	White	White	Green	Green	White
Memory BW	White	White	White	White	Green
Design scalability	White	White	White	White	White
Design Complexity	Red	White	White	Green	White

Comparison: eDRAM perspective

	SS	VLIW	CMP	RC	VIRAM
BW utilization					
Latency tolerance					
Power consumption					
Need for fast logic					
DRAM testing					
Logic redundancy					
Design scalability					
Design complexity					

Summary (1/3)

- Future microprocessor applications
 - designed for mobile multimedia devices
 - media performance, energy efficiency and reduced complexity are major requirements
- Embedded DRAM advantages
 - high memory bandwidth
 - system-on-a-chip benefits
- Microprocessor architecture approaches
 - advanced superscalar
 - VLIW
 - single chip multiprocessor
 - vector microprocessor

Summary (2/3)

- Superscalar & VLIW
 - best with desktop/server applications
 - not good match for embedded DRAM technology
- Chip Multiprocessor & vector microprocessor
 - modular/scalable designs
 - can utilize eDRAM benefits and address challenges
- Vector microprocessor
 - high media performance
 - simple, energy efficient hardware
 - great match for eDRAM

Summary (3/3)

- Unlikely that eDRAM will make it in the desktop high-performance microprocessors (at least for a while...)
- Processor architectures developed for mobile media devices address many of the eDRAM challenges
- Cost and commercial success of eDRAM based processors remains to be seen...

References (1)

- C. Kozyrakis, D. Patterson, “A New Direction in Computer Architecture Research”, IEEE Computer, vol. 31, no. 11, November 1998

Computer Architecture

- D. Patterson, L. Hennessy, “Computer Organization and Design: The Hardware/Software Interface”, 2nd edition, 1997, Morgan Kaufmann
- L. Hennessy, D. Patterson, “Computer Architecture: A Quantitative Approach”, 2nd edition, 1995, Morgan Kaufmann

High-performance Processors

- R. E. Kessler, “The Alpha 21264 Microprocessor: Out-Of-Order Execution at 600 Mhz”, Hot Chips Conference Record, August 1998
- Gary Lauterbach, “UltraSPARC-III: A 600 MHz 64-bit Superscalar Processor for 1000-way Scalable Systems”, Hot Chips Conference Record, August 1998
- M. Choudhury et.al, “A 300MHz CMOS Microprocessor with Multi-Media Extensions”, Digest of Technical Papers, ISSCC, February 1997

References (2)

Embedded Processors

- M. Schlett, “Trends in Embedded Microprocessor Design”, IEEE Computer, vol. 31, no. 8, August 1998
- Toru Shimizu, “The M32Rx/D - A Single Chip Microcontroller With a 4MB Internal DRAM”, Hot Chips Conference Record, August 1998
- J. Choquette, “Genesis microprocessor”, Hot Chips Conference Record, August 1998
- F. Arakawa et.al, “SH4 RISC Multimedia Microprocessor”, IEEE Micro, vol. 18, no. 2, March 1998
- T. Litch et.al, “StrongARMing Portable Communications”, IEEE Micro, vol. 18, no. 2, March 1998
- L. Goudge, S. Segars, “Thumb: reducing the cost of 32-bit RISC performance in portable and consumer applications”, In the Digest of Papers, COMPCON '96, February 1996

Embedded DRAM

- D. Patterson et.al, “Intelligent RAMs”, Digest of Technical Papers, ISSCC, February 1997
- R. Fromm et.al., “The Energy Efficiency of IRAM Architectures”, The 24th International Symposium on Computer Architecture , June 1997

References (3)

- K. Murakami et.al, “Parallel Processing RAM Chip with 256Mb DRAM and Quad Processors”, Digest of Technical Papers, ISSCC, February 1997
- Tadaaki Yamauchi et.al., “The Hierarchical Multi-Bank DRAM: A High-Performance Architecture for Memory Integrated with Processors”, Proceedings of the 19th Conference on Advanced Research in VLSI, September 1997
- J. Dreibelbis et.al, “An ASIC Library Granular DRAM Macro with Built-In Self Test”, Digest of Technical Papers, ISSCC, February 1998
- T. Yabe et.al, “A Configurable DRAM Macro Design for 2112 Derivative Organizations”, Digest of Technical Papers, ISSCC, February 1998
- A. Saulsbury, A. Nowatzky, “Missing the memory wall: the case for processor/memory integration”, in proceedings of the 23rd Annual International Conference on Computer Architecture, May 1996
- T. Sunaga et.al, “A parallel processing chip with embedded DRAM macros. IEEE Journal of Solid-State Circuits, vol. 31, no.10, October 1996
- D. Elliott et.al, “Computational RAM: a memory-SIMD hybrid and its application to DSP”, Proceedings of the IEEE 1992 Custom Integrated Circuits Conference, May 1992

References (4)

- N. Bowman et.al, "Evaluation of Existing Architectures in IRAM Systems", Workshop on Mixing Logic and DRAM: Chips that Compute and Remember at ISCA '97, June 1997
- NEC Corp., "Virtual Channel Memory Technology", <http://www.nec.com>
- Miyano S. et.al: "A 1.6Gbyte/s Data Transfer Rate 8 Mb Embedded DRAM", IEEE Journal of Solid-State Circuits, v. 30, no. 11, November 1995

Microprocessor Architecture Trends

- T. Mudge, "Strategic Directions in Computer Architecture", ACM Computing Surveys, 28(4):671-678, December 1996
- J. Crawford, J. Huck, "Motivations and Design Approach for the IA-64 64-Bit Instruction Set Architecture", In the Proceedings of the Microprocessor Forum, October 1997
- Y.N. Patt et.al., "One Billion Transistors, One Uniprocessor, One Chip", IEEE Computer, 30(9):51-57, September 1997
- M. Lipasti, L.P. Shen, "Superspeculative Microarchitecture for Beyond AD 2000", IEEE Computer, 30(9):59-66, September 1997

References (5)

- J. Smith, S. Vajapeyam, “Trace Processors: Moving to Fourth Generation Microarchitectures”, IEEE Computer, 30(9):68-74, September 1997 S.J. Eggers et.al., “Simultaneous Multithreading: a Platform for Next-Generation Processors”. IEEE MICRO, 17(5):12-19, October 1997
- L. Hammond et.al., ”A Single-Chip Multiprocessor”. IEEE Computer, 30(9):79-85, September 1997
- E. Waingold et.al, “Baring It All to Software: Raw Machines”, IEEE Computer, 30(9):86-93, September 1997 S.J. Eggers et.al, “Simultaneous Multithreading: a Platform for Next-Generation Processors”, IEEE MICRO, 17(5):12-19, October 1997
- C.E. Kozyrakis et.al., “Scalable Processors in the Billion-Transistor Era: IRAM”, IEEE Computer, 30(9):75-78, September 1997
- K. Olukotun et.al., “Improving the Performance of Speculatively Parallel Applications on the Hydra CMP”, In the Proceedings of the 1999 ACM International Conference on Supercomputing, June 1999.
- R. Barua et.al., “Maps: A Compiler-Managed Memory System for Raw Machines”, In the Proceedings of the 26th International Symposium on Computer Architecture June 1999.
- S. Goldstein et.al., "PipeRench: a Coprocessor for Streaming Multimedia Acceleration”, In the Proceedings of the 26th International Symposium on Computer Architecture June 1999.

References (6)

General-purpose architectures and media-processing

- T. Conte et.al, “Challenges to Combining General-Purpose and Multimedia Processors”, IEEE Computer, vol. 30, no. 12, December 1997
- K. Diefendorff , P. Dubey, “How Multimedia Workloads Will Change Processor Design”, IEEE Computer, 30(9):43-45, September 1997