The Future of Microprocessor Architecture

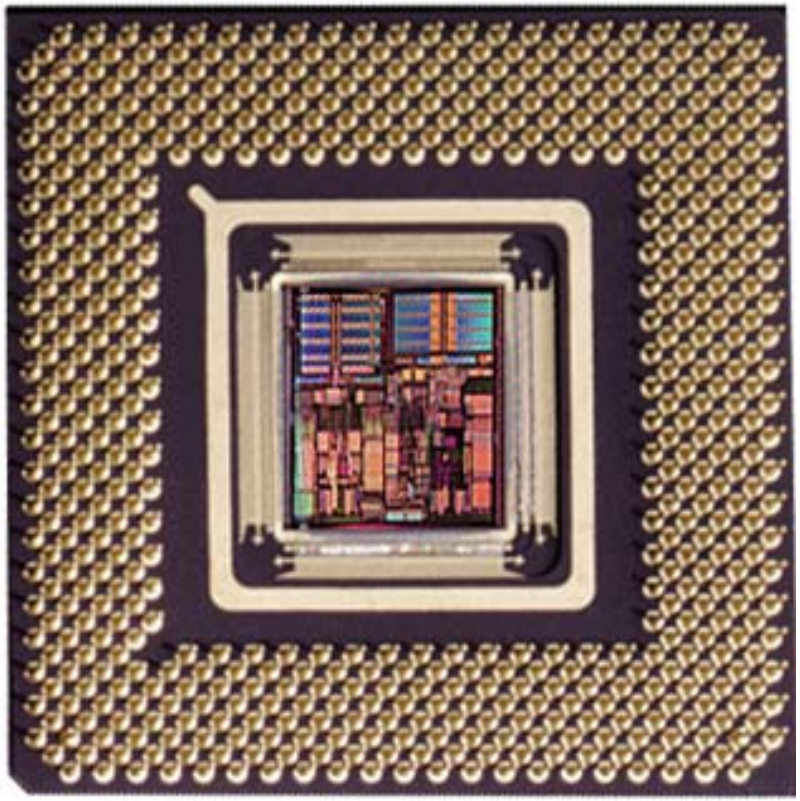
Donald Alpert
Stanford University
Intel Corporation

Outline

- Where are we?
- How did we get here?
- Where are we going?

Today: Alpha 21264

- 64-bit Address/Data
- Superscalar
- Out-of-Order Execution
- 256 TLB entries
- 128KB Cache
- Adaptive Branch Prediction
- 0.35 μm CMOS Process
- 15.2M Transistors
- 600 MHz



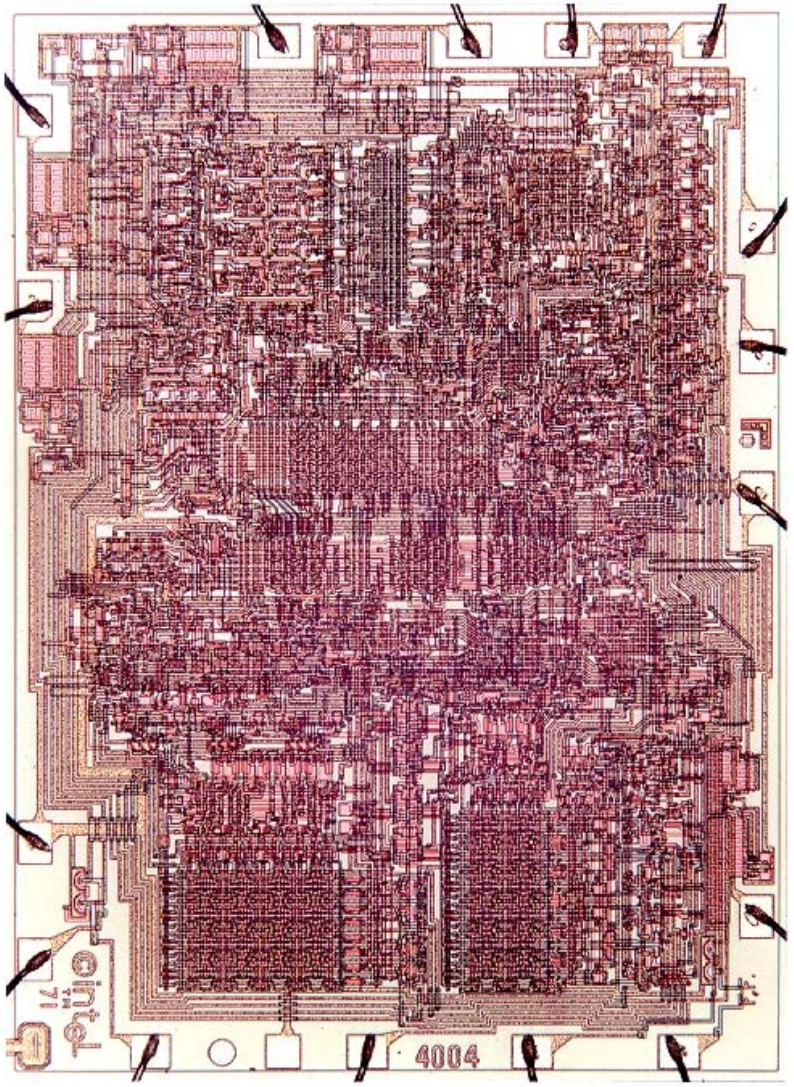
Source: Digital

History

- Technology
- Functionality
- Partitioning

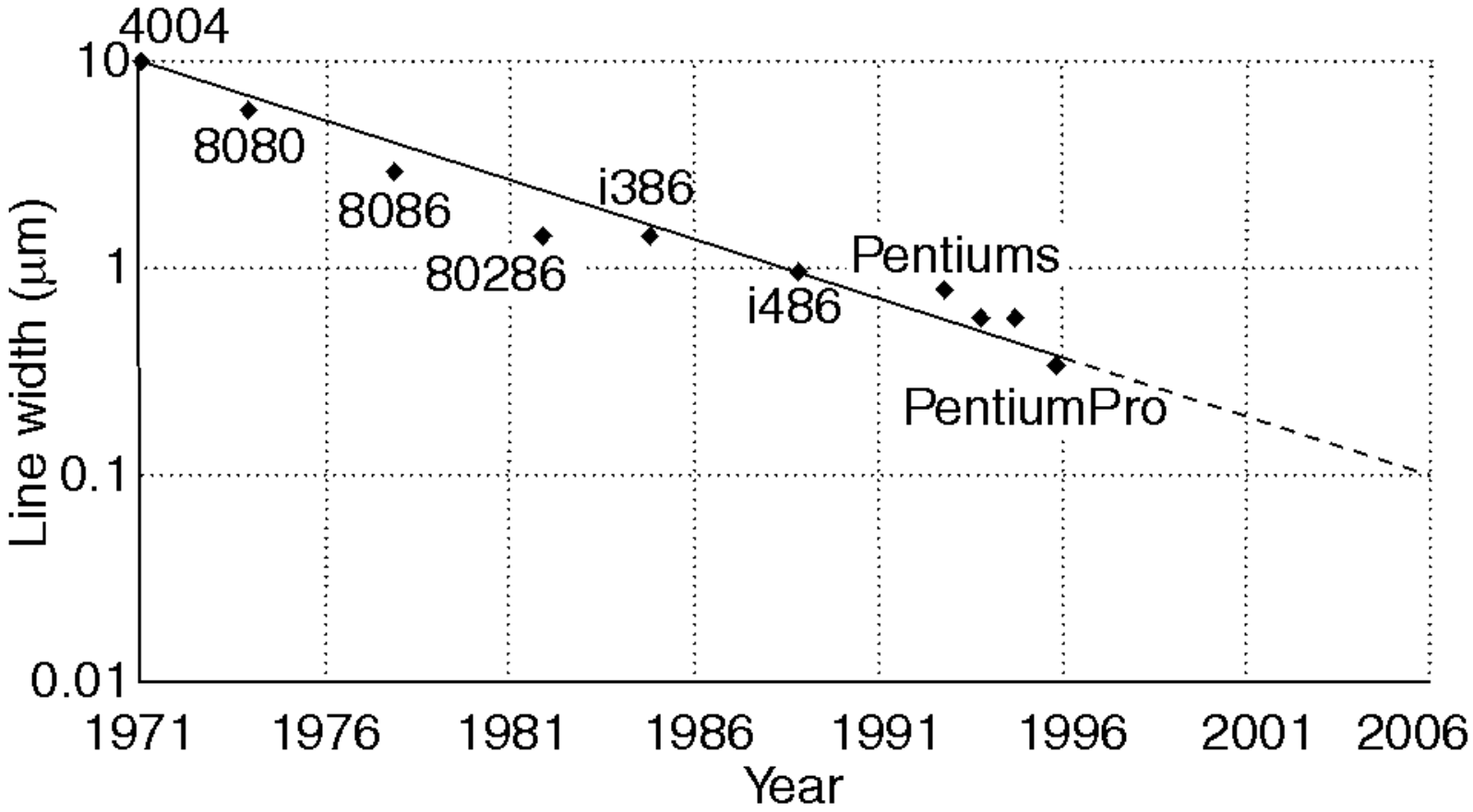
In the Beginning: Intel 4004

- 4-bit Data
- 12-bit Address
- 8 μm PMOS
- 2300 Transistors
- 750 KHz
- 1971

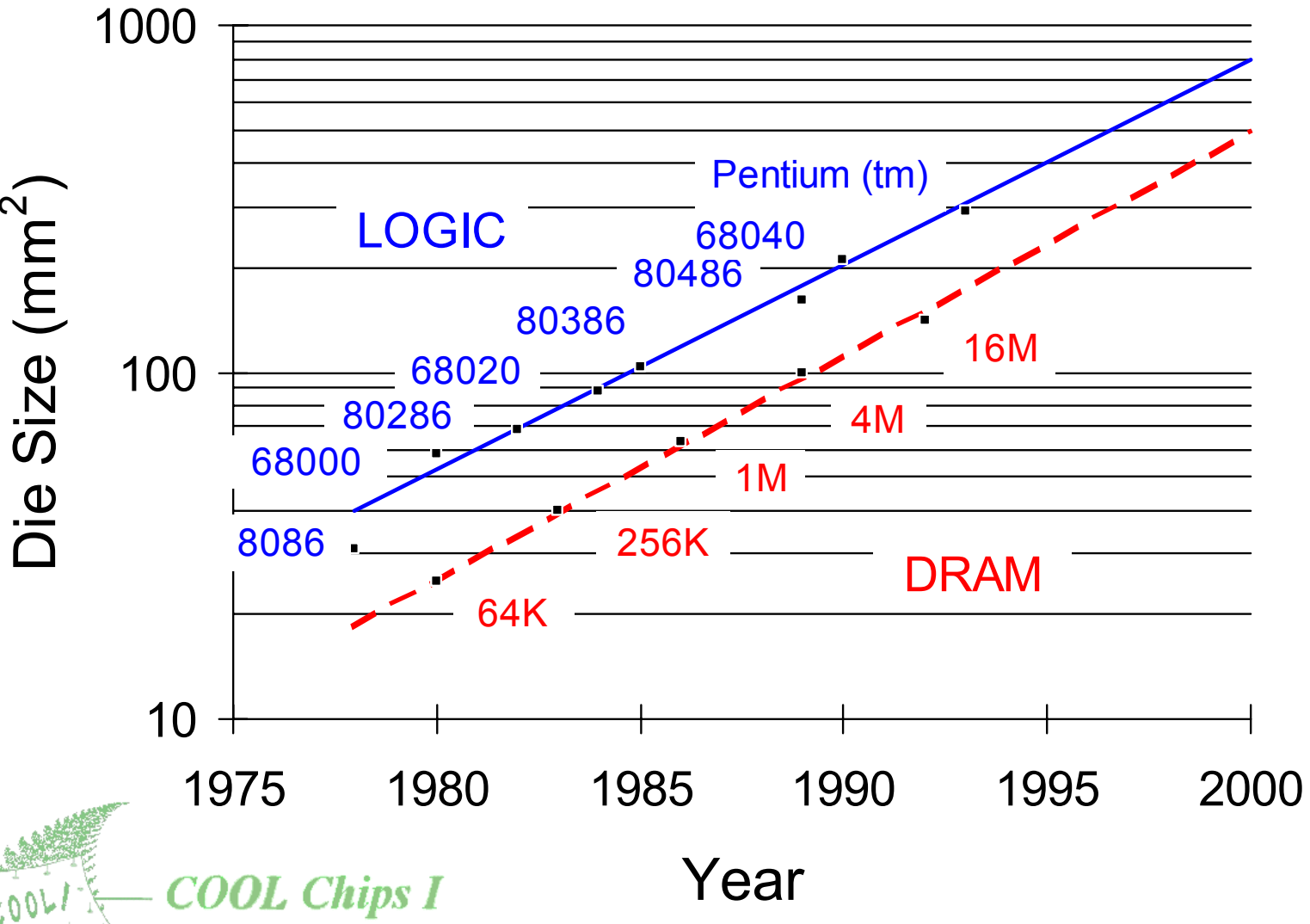


Source: Intel

Lithography

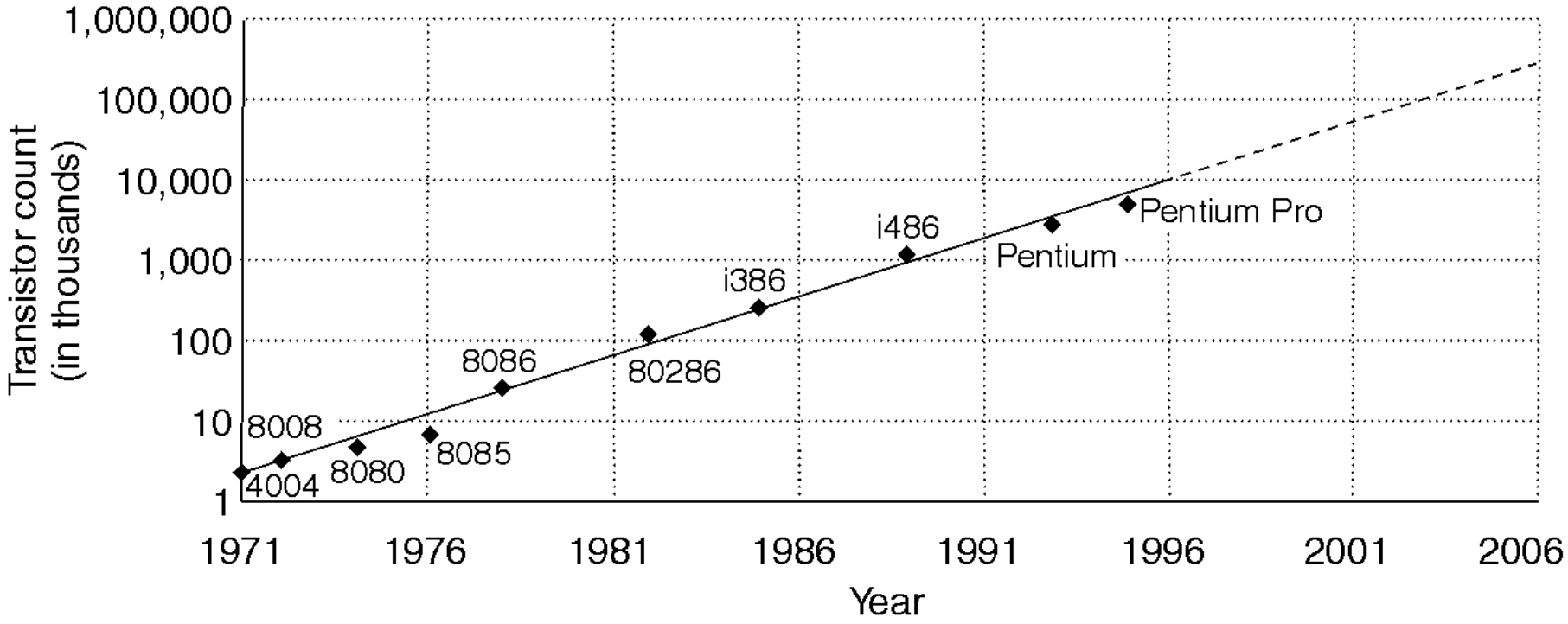


Die Size



Source: Intel

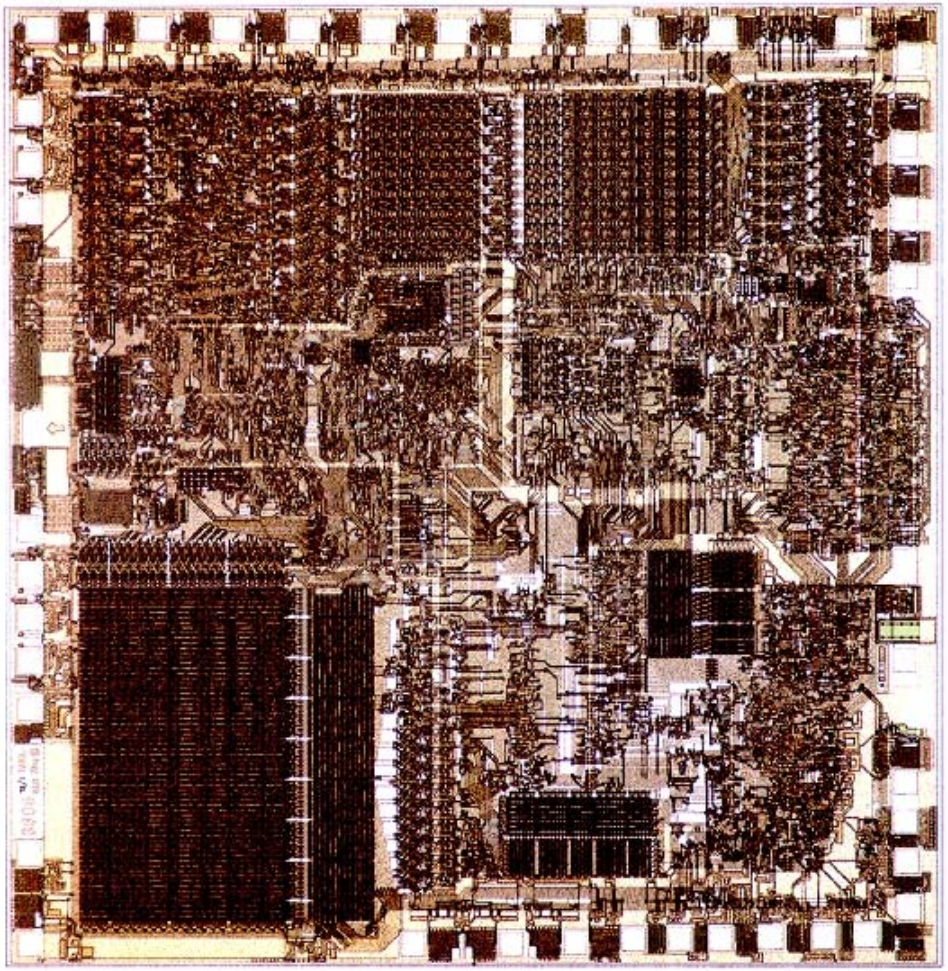
Transistor Count



Source: A. Yu, IEEE Micro 12/96

8-Bit Microprocessor Generation

- 8-bit Data
- 16-bit Address
- 6 μm NMOS
- 6K Transistors
- 2 MHz
- 1974

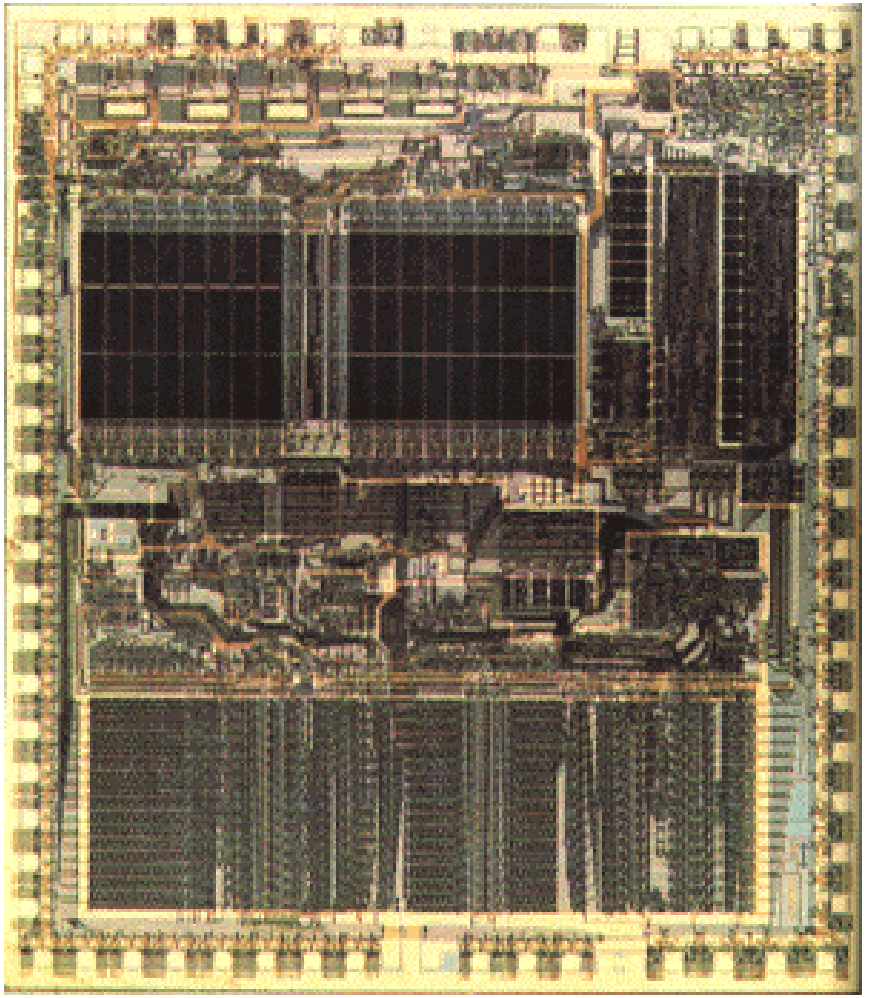


Intel 8080

Source: Intel

16-Bit Microprocessor Generation

- **Issues**
  - — Segmented vs. Linear Memory Addresses
  - — Registers
  - — Addressing Modes
  - — Floating-Point



Motorola 68000

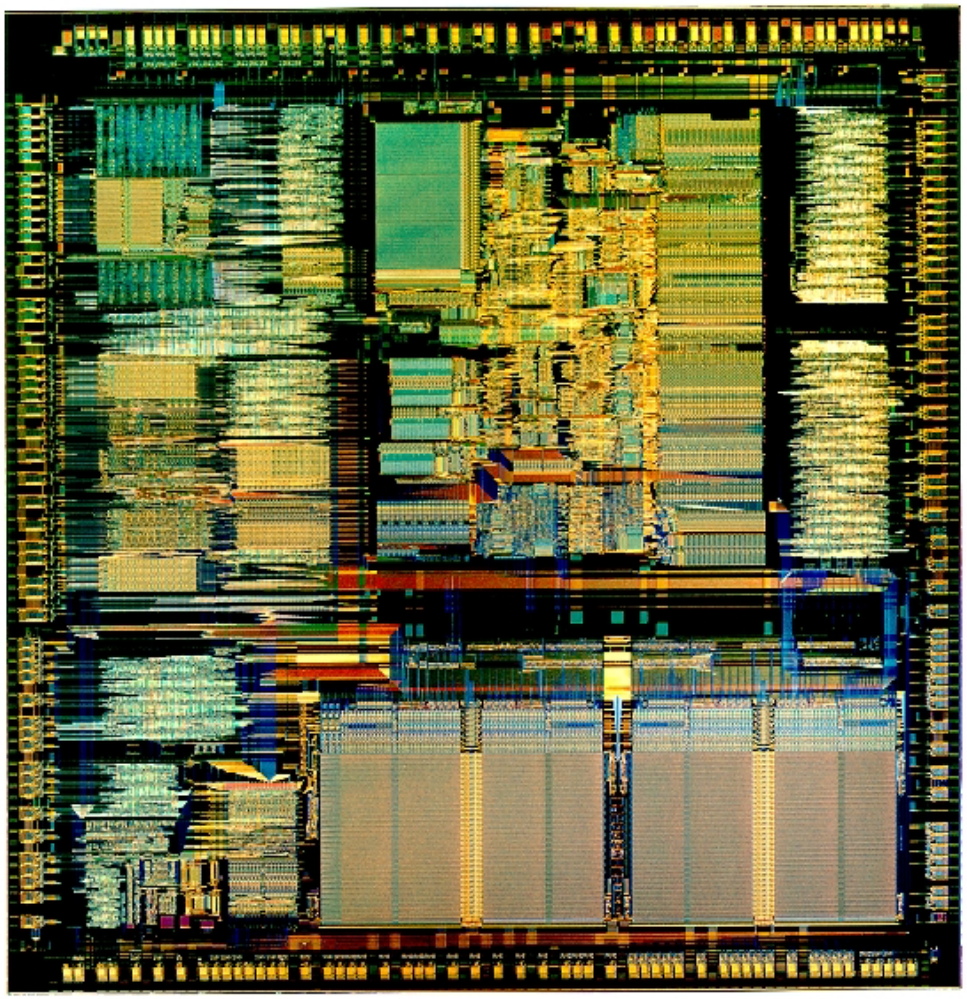
Photograph: Computer Museum

16-Bit Microprocessor Generation

	Intel 8086	Zilog Z8000	Motorola 68000
Integer Path	16-bit	16-Bit	16-Bit
Floating-Point	8087	No	No
Addresses	Segment (16)	Segment (16)	Linear (24)
OS Protection	No	Yes	Yes
Memory Mgt.	No	Segmented	No
Cache	No	No	No
Technology	3μm NMOS	4-6(?)μm NMOS	4μm NMOS
No. Transistors	29K	17.5K	68K
Frequency	5 MHz	4 MHz	8 MHz
Year	1978	1979	1979

32-Bit Microprocessor Generation

- **Issues**
  - — Cache
  - — TLB
  - — RISC vs. CISC



Intel386 CPU

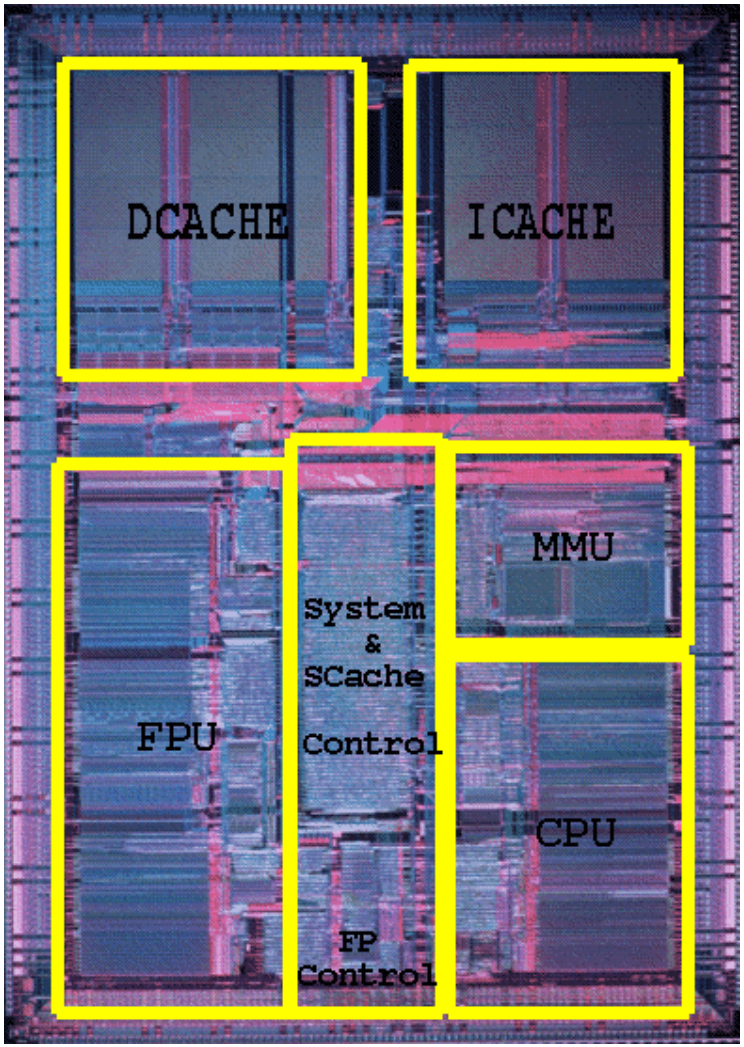
Source: Intel

32-Bit Microprocessor Generation

	Intel 80386	Motorola 68020	MIPS R2000
Integer Path	32-bit	32-Bit	32-Bit
Floating-Point	80387	68881	R2010
Addresses	Seg/Linear (32)	Linear (32)	Linear (32)
OS Protection	Yes	Yes	Yes
Memory Mgt.	32-entry TLB	68851	64-entry TLB
Cache	82385	256B	Controller
Technology	1.5μm CMOS	2μm CMOS	2μm CMOS
No. Transistors	275K	200K	100K
Frequency	16 MHz	16 MHz	16.7 MHz
Year	1985	1984	1986

Baseline Microprocessor

- **Full Functionality**
  - 32-bit Integer
  - 64-bit Floating-Point
  - Paged Virtual Memory (TLB)
- **Performance**
  - Full-Width Datapaths
  - Pipelined Function Units
  - 8-16KB Cache
- **Technology**
  - ~1.0 μm CMOS
  - ~1M Transistors



MIPS R4000

Source:SGI MIPS

Since Baseline Microprocessor

- **Technology**
  - 1.0 μm → 0.25 μm
  - 1M Tx → 10M Tx
- **Addresses/Integers**
  - 32b → 64b
- **Superscalar**
  - In-Order Execution
  - Out-of-Order Execution
- **Branch Prediction**
- **Cache**

Consistent Mistakes

- Underestimate Technology Improvement Rate
- Underestimate Complexity
- Underestimate Software Development Effort
- Underestimate Market Size

Where Are We going?

- What we know
- What we know that we don't know
- What we don't know that we don't know

Semiconductor Technology Roadmap

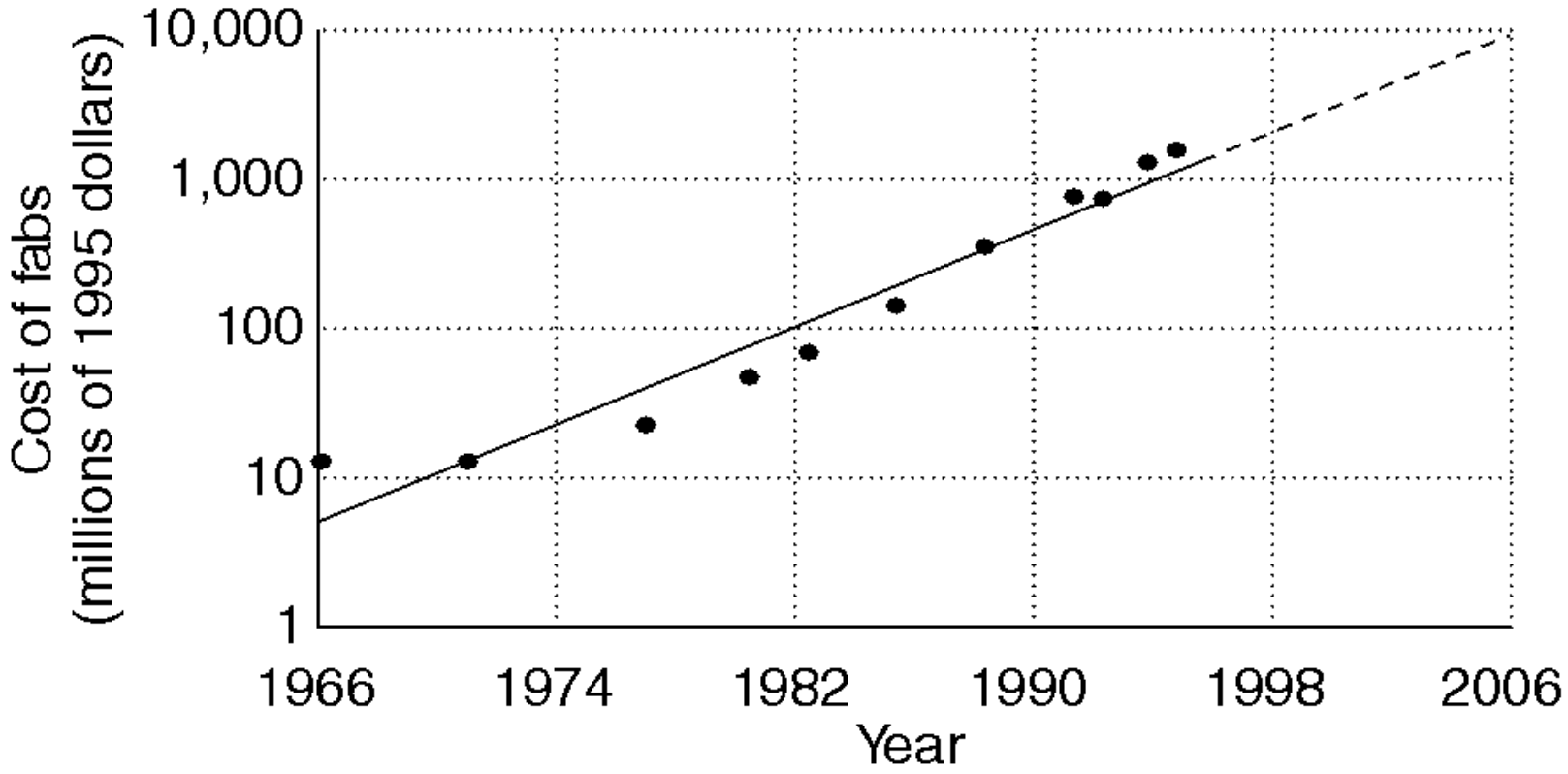
	1997	1999	2001	2003	2006	2009	2012
Lithography (nm)	250	180	150	130	100	70	50
Die Size (mm^2)	300	340	385	430	520	620	750
Transistors (M)	11	21	40	76	200	520	1400
Frequency (MHz)							
Local Clock	750	1250	1500	2100	3500	6000	10000
Cross-Chip Clock	750	1200	1400	1600	2000	2500	3000
Power (W)	70	90	110	130	160	170	175
Voltage (V)	1.8-2.5	1.5-1.8	1.2-1.5	1.2-1.5	0.9-1.2	0.6-0.9	0.5-0.6
I/O Pins	1450	2000	2400	3000	4000	5400	7300
Wiring Levels	6	6-7	7	7	7-8	8-9	9

Source: Semiconductor Industry Association

IA-64 and Merced™ CPU

- **IA-64**
  - — Joint 64-bit architecture definition by Intel and H-P
  - — Explicitly Parallel Instruction Computing (EPIC)
    - Encode independent instructions
    - 128 registers
    - Predication
    - Speculation
- **Merced CPU**
  - — First IA64 implementation
  - — 0.18 μm technology
  - — 1999 Production

Fabrication Facility Costs



Moore's Second Law: Fab Costs Grow 40% Per Year

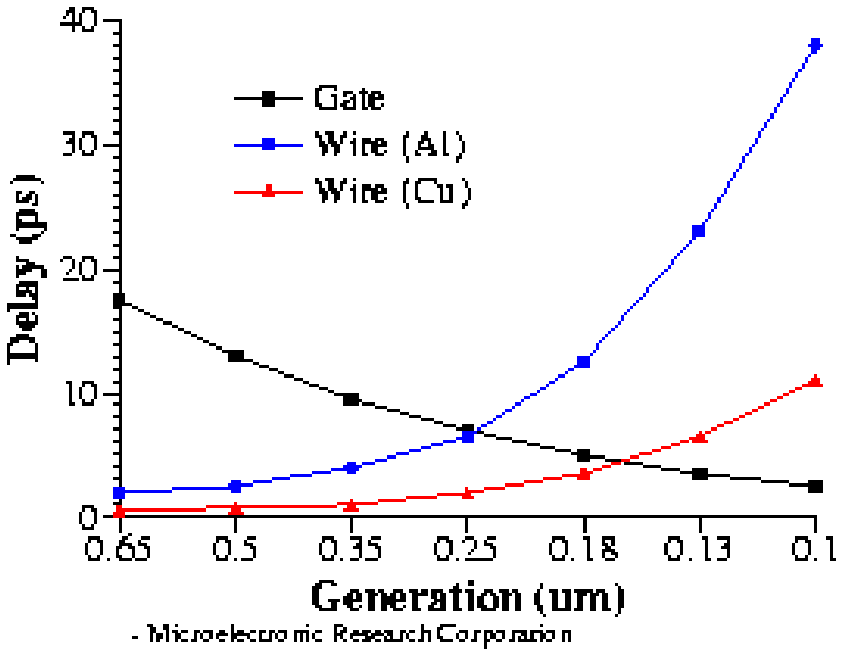
Source: A. Yu, IEEE Micro 12/96

Known Challenges

- Interconnect
- Power
- Reliability
- Verification
- Mixed-Signal

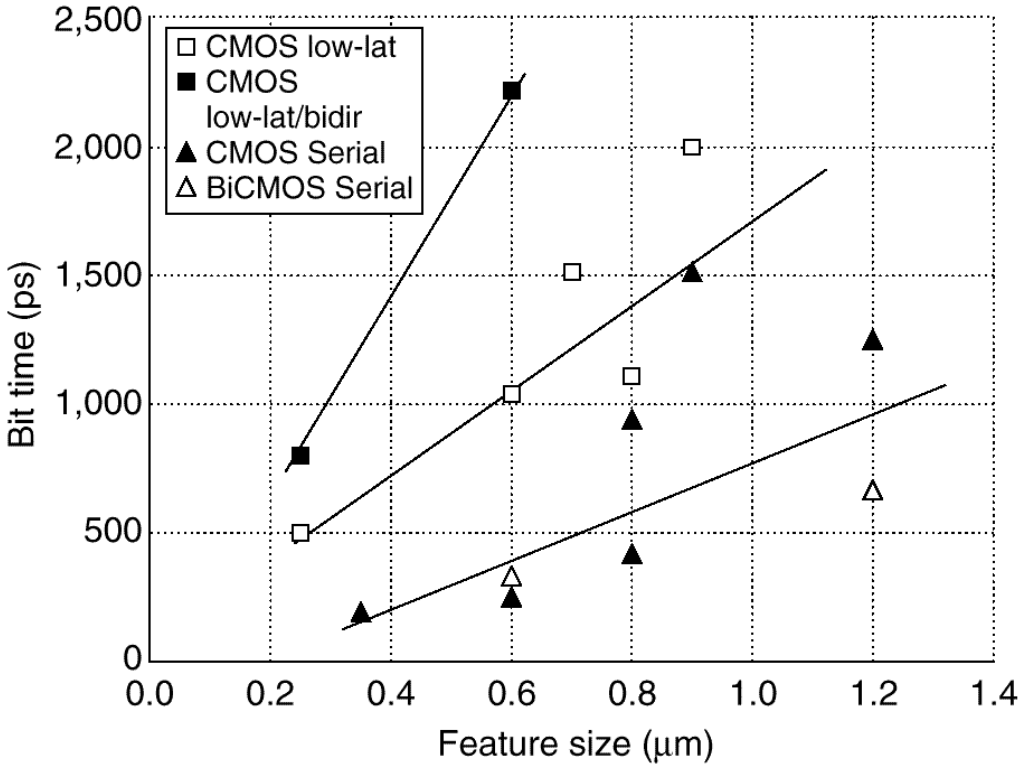
Wire Delay Is Increasing

- **Gate delay decreasing 25% per generation**
- **Wire delay increasing 100% per generation**
- **Communicate across a chip**
  - 1 clock at 400 MHz in 0.35μm
  - 12.4 clocks at 1 GHz in 0.1μm



Off-Chip Data Bandwidth Is Scaling

- Achievable bit times scale with circuit speed
- Transceiver fits in the area of a (large) pad driver
- Still may need to increase number of I/O signals each generation to match logic integration

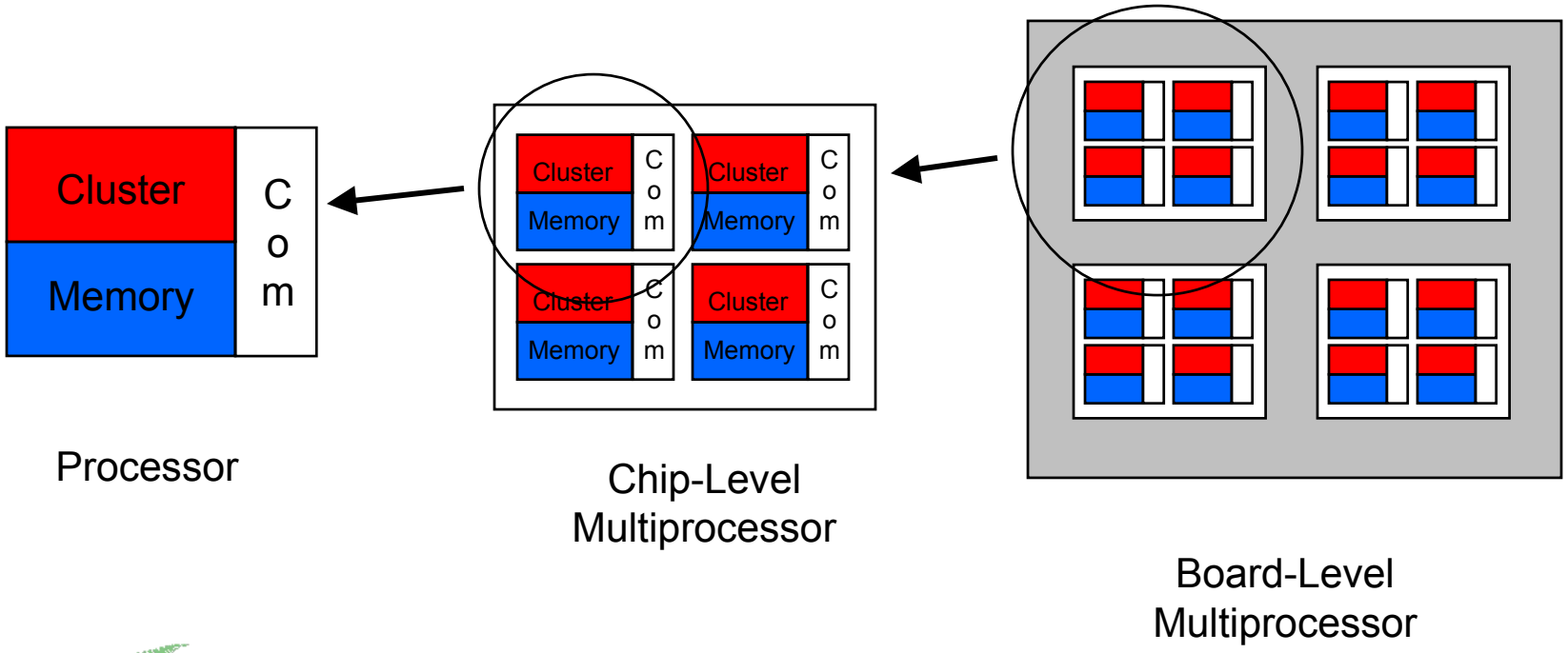


Source: M. Horowitz, IEEE Micro 1/98 and B. Dally

Scalable Architecture for ULSI

- **Processor**
  - — Core cluster of computational units and registers
  - — Memory
  - — Inter-processor communication unit
- **Technology Properties**
  - — Local interconnect for highest-frequency cluster
  - — Shrink and replicate processors for higher integration
- **Programming Properties**
  - — Replicate chips of multiprocessors for higher performance
  - — Consistent latencies in clocks across generations

Scalable Architecture for ULSI

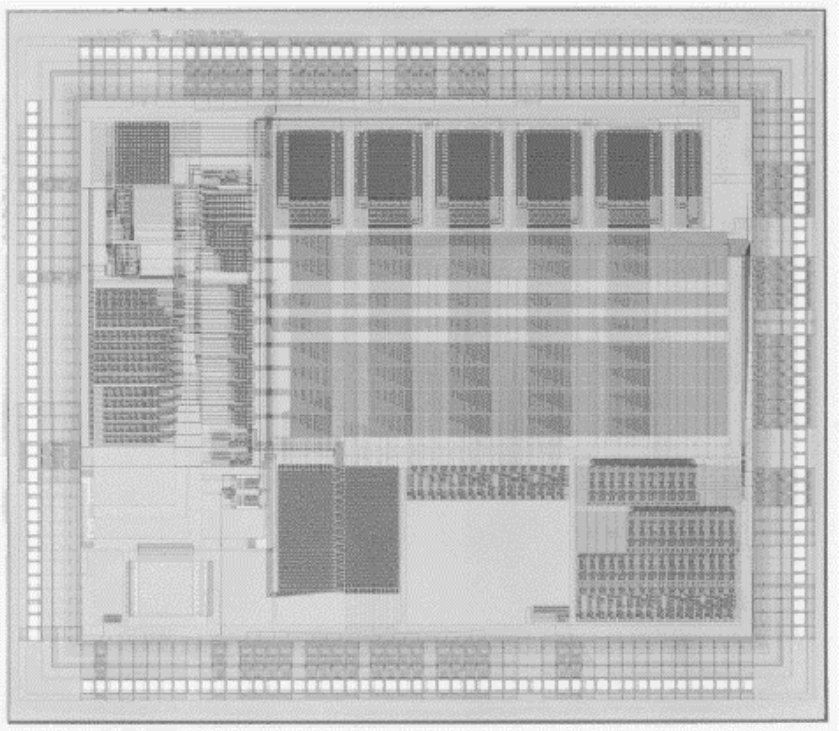


Microprocessor Architecture Research

- Wave Pipelining
- Multithreaded Processors
- Single-Chip Multiprocessors
- Vector/Stream Processors
- Intelligent RAM
- Reconfigurable Computing

Wave Pipelining

- **Sub-Nanosecond Arithmetic Processor (SNAP)**
  - Prof. Mike Flynn at Stanford
- **Wave Pipelining**
  - Uses minimum propagation delay (T_{min}) to store data in combinational logic paths
  - Conventional pipeline limited by maximum delay path (T_{max})
  - Wave pipeline limited by difference in delay (T_{max} - T_{min})
  - Potential 2-3X performance improvement in CMOS with comparable cost to conventional pipelining



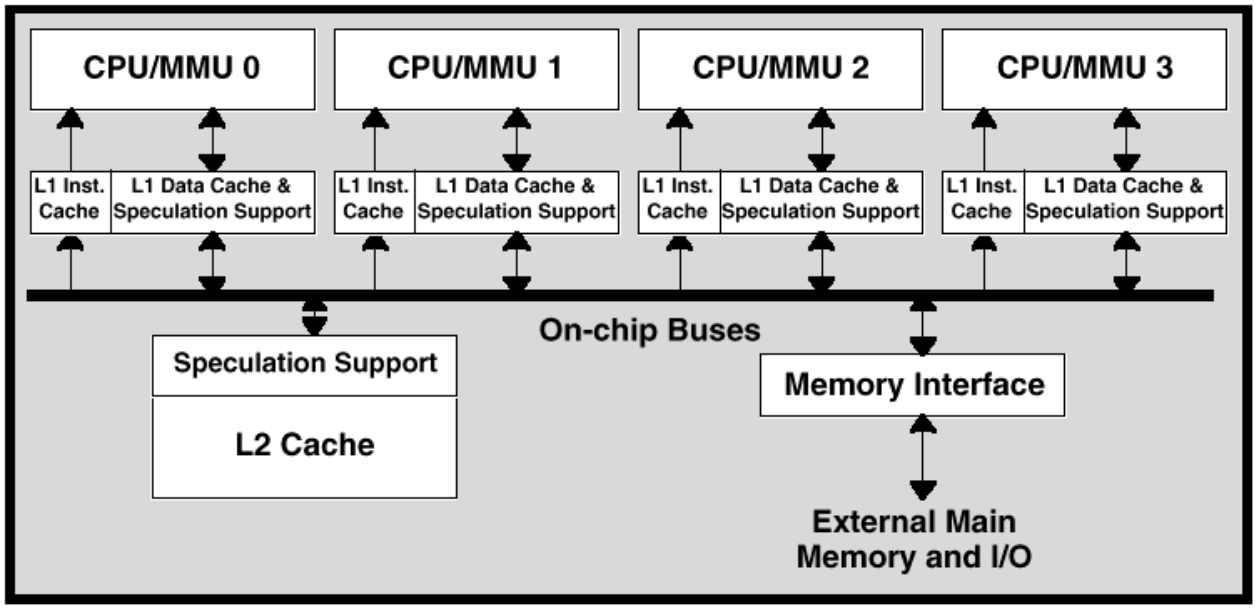
CMOS Wave-Pipelined Vector Unit

Multithreaded Processors

- **Simultaneous Multithreading (SMT) Processors**
  - — Prof. Susan Eggers et al at University of Washington
  - — Targets fine-grain multithreaded applications/workloads
- **Based on a Dynamic, Superscalar Processor**
  - — Add IDs for multiple (8) threads to registers/structures
  - — Function units are scheduled dynamically with data-ready instructions from multiple threads
- **Multi-ported Instruction Cache**
  - — Fetch from two threads simultaneously
  - — Priority to threads with fewest instructions in pipe
- **Potential 2X Performance Improvement**
  - — For incremental cost vs. conventional superscalar

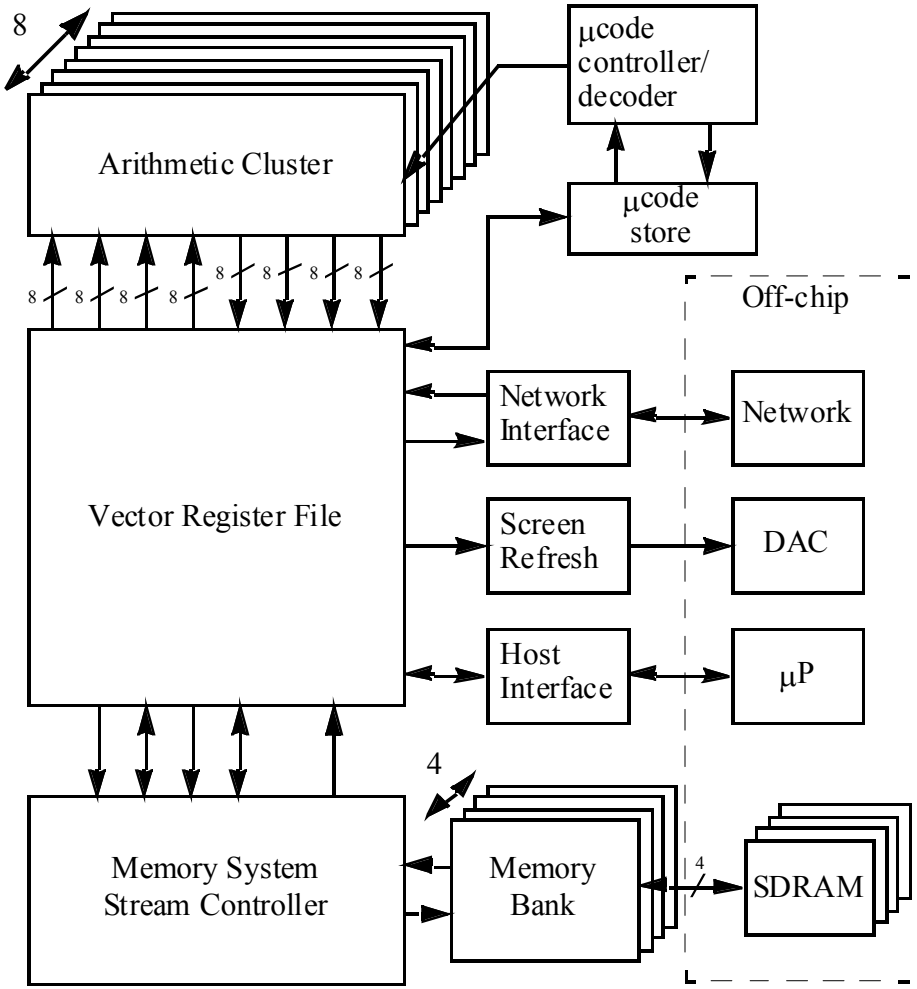
Single-Chip Multiprocessors

- **Hydra Project**
  - Prof. Kunle Olukotun at Stanford
  - Targets thread-level parallelism
- **4 CPUs on a Chip**
- **3-Level Cache Hierarchy**
- **Parallelizing Compiler Technology**



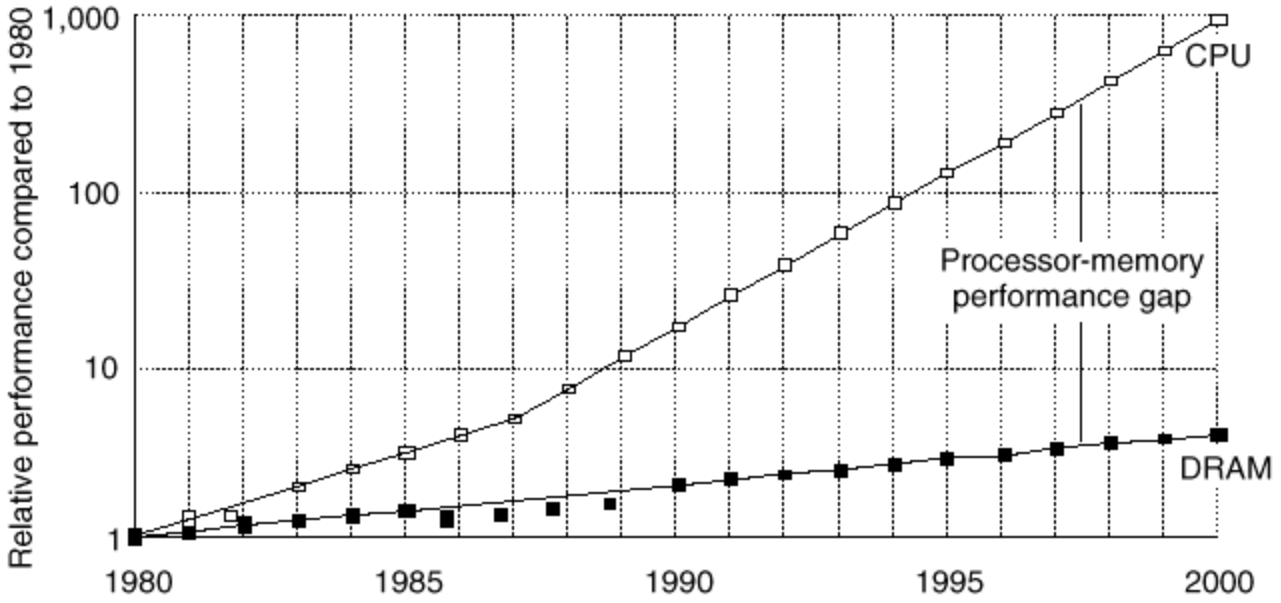
Vector/Stream Processors

- **Imagine Project**
  - Prof. Bill Dally at Stanford
  - Targets Graphics and Signal Processing
- **Arithmetic Clusters (8)**
  - Multiple interconnected-ALUs
  - Local registers
  - Statically scheduled operations and bus usage
- **Memory Streams**
  - Arrays of multimedia structures
  - Multiple SDRAM banks
  - Vector register file
    - 16K words
    - 18 streams



Intelligent RAM

- **IRAM**
  - — Prof. Dave Patterson at U .C. Berkeley
  - — Target latency/bandwidth gap between CPU and DRAM
- **Integrate DRAM with Conventional CPU**
- **Specialized Processor Exploits On-chip DRAM Bandwidth**



Source: D. Patterson, IEEE Micro 3/97

Reconfigurable Computing

- **Adaptive Computing Systems**
  - DARPA program
  - Target FPGAs for high-performance programmable HW
- **Improved Performance Over Programming SW**
  - 10X over DSP
  - 100X over general-purpose microprocessor
- **Cost 2X Over ASIC at Comparable Performance**
- **Potential Applications**
  - Pattern matching (image recognition)
  - Encryption
  - Signal processing

Predictions for Cool Chips X

- **No "Cool" Technology**
  - — But power managed at all design levels
- **Systems on Chips**
  - — Integrated application solutions
  - — Majority of transistors for memory
  - — Multiple, heterogeneous processors
  - — Mixed-signal applications
  - — On-chip bus standards
  - — On-chip interconnection networks
- **Unlikely**
  - — Optical interconnect
  - — Reconfigurable computing