EE382N (22): Computer Architecture - Parallelism and Locality
Fall 2009

Lecture 3 – Locality Mechanisms

Mattan Erez

The University of Texas at Austin

Announcement

- Please sign up for scribing
- Project presentations in lieu of exam — held at same time and place of exam
- Reading assignment can be done in groups
- Start forming groups for lab 1
  - Remember that you will not be partnering with same people for labs 2 and 3 (OK for project and reading)
- Reading homework assignments encourage, but not required, in groups
  - Writeups should be "talking points", not essays

- Questions on procedures, requirements, …?

Next Three Lectures:

- Locality mechanisms in a CPU – traditional view
  - What?
  - Why?
  - How?
- Another view of locality mechanisms
  - Why?
- Exploiting locality in a CPU
  - Registers
  - Cache-aware
  - Cache-oblivious

A CPU-Based Computer

- 360N!
- Review on board

A CPU-Based Computer

- CPU
  - ALUs
  - Registers
  - Reservation stations
  - Bypass networks
  - Caches

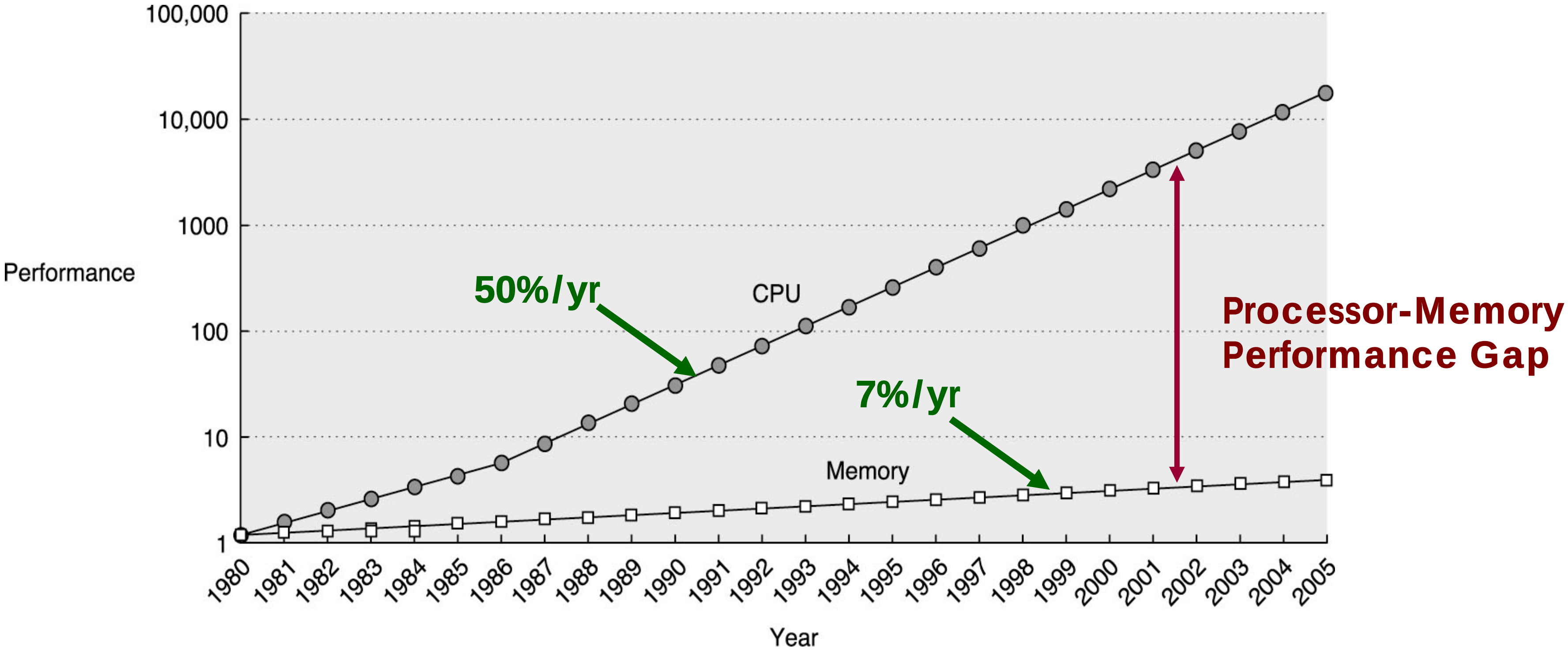
locality

- Main memory
- I/O
  - Disk
  - Network
  - Display
  - User input devices

Registers

- CPUs
  - First form of storage – not locality at all really
  - Then used for interfaces (buffers), compact encoding, and convenience
    - Accumulator, instruction pointer, status registers, …
  - Later (when memories got slower than ALUs)
    - Shorter access latency benefit
- ASICs
  - Buffering
  - Latency
  - Power
  - Bandwidth
    - Many “parallel” registers

Processor-Memory "Performance" Gap



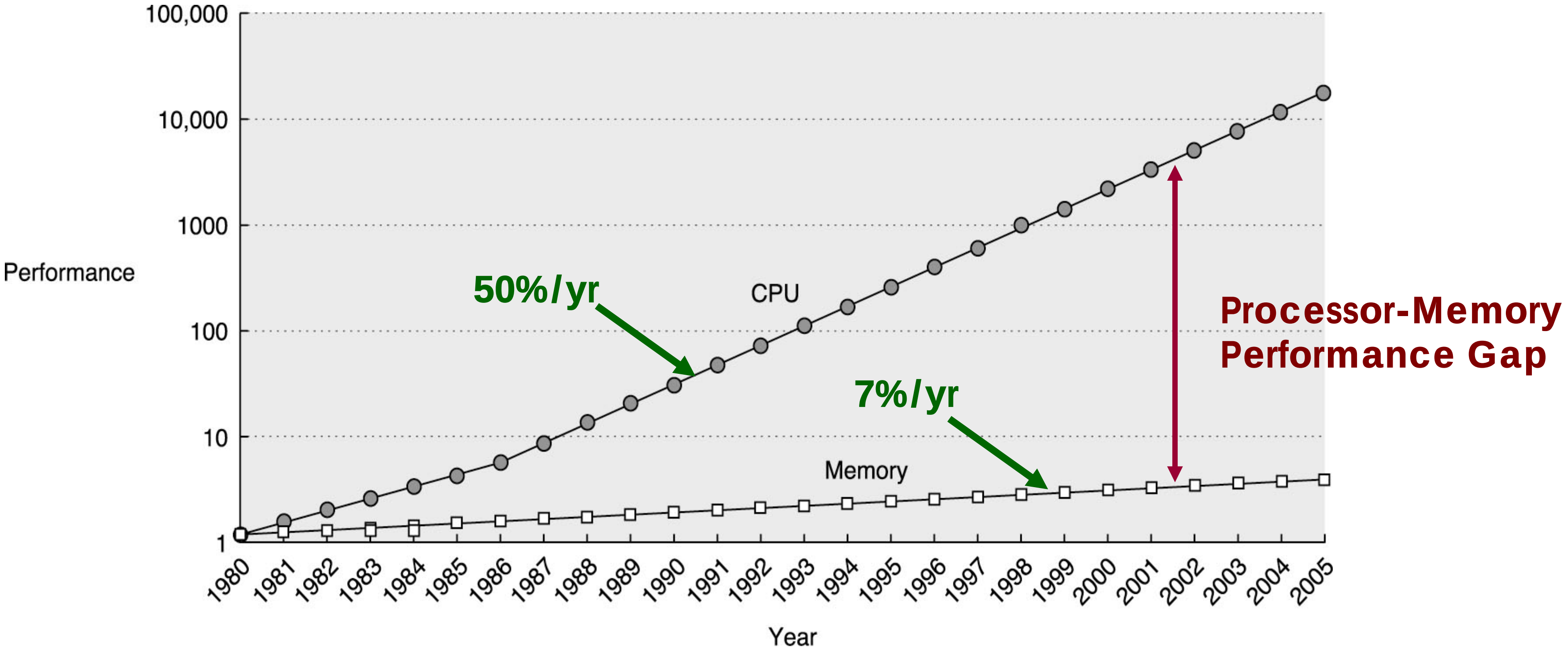
50%/yr

7%/yr

Processor-Memory Performance Gap

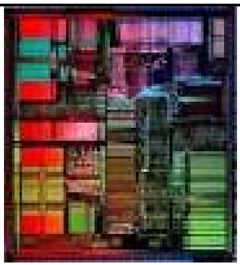
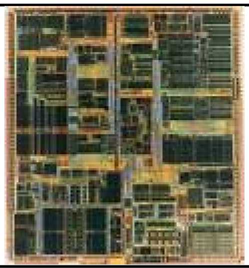
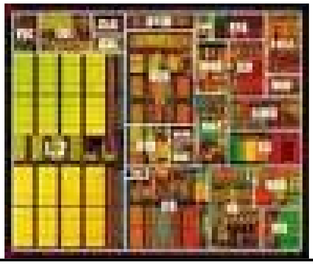
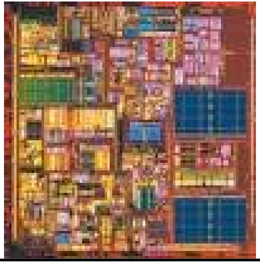
Hennessy and Patterson,
Computer Architecture –
A Quantitative Approach
(2003)

Processor-Memory "Latency" Gap



Hennessy and Patterson,
Computer Architecture –
A Quantitative Approach
(2003)

Locality Improves Latency

	Intel Pentium	Pentium II	Pentium III	Pentium IV
				
Technology	.80µm .25µm	.28µm .25µm	.25µm .13µm	.18µm 65nm
Frequency	75 MHz 300 MHz	233 MHz 533 MHz	500 MHz 1.4 GHz	1.5 GHz 3.8 GHz
Register access	1	1	1	1
L1 access	1	3	2	4
L2 access		18	5	12
Memory access	10 – 20	20 - 30	30 - 50	30 - 400

Locality Improves Power

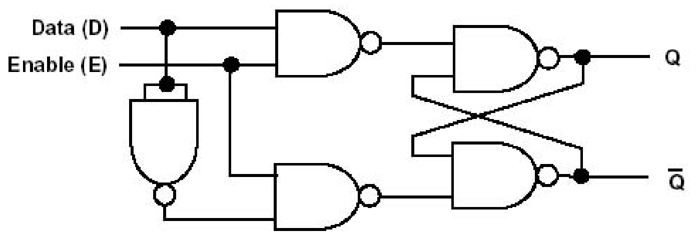
Operation	65nm	32nm	16nm
64b FP Operation	38pJ	12.5pJ	3.8pJ
Read 64b from 16KB Cache	17.5pJ	5.3pJ	2pJ
Transfer 64b across chip (10mm, Rep.)	179pJ	179pJ	179pJ
Transfer 64b across chip (10mm, Cap.)	18pJ	18pJ	18pJ
Transfer 64b off chip	154pJ	115pJ	100pJ

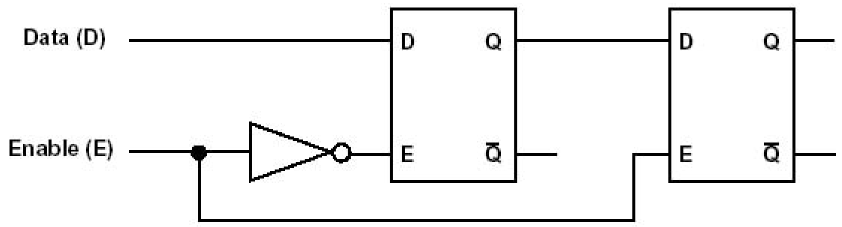
Locality Improves Bandwidth

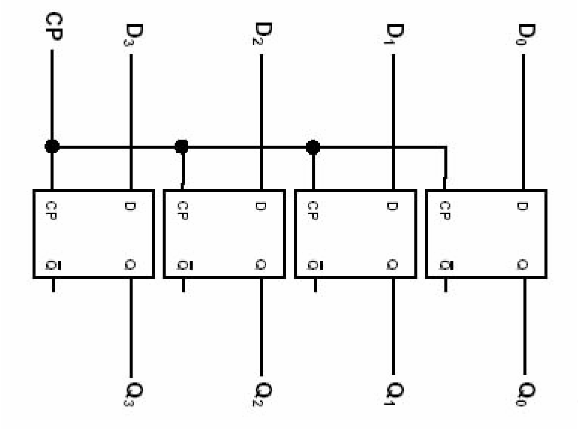
- Mostly a matter of wire density
  - Min wire pitch (χ) is ½ intermediate and ¼ global pitches
  - Vias and repeaters restrict routing and add area
- Rules of thumb for wires
  - Latency directly proportional to distance
  - BW inversely proportional to distance
  - Power directly proportional to distance + step function
- More on wires next lecture

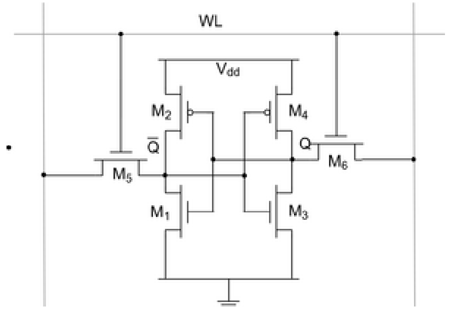
Register or Memory?

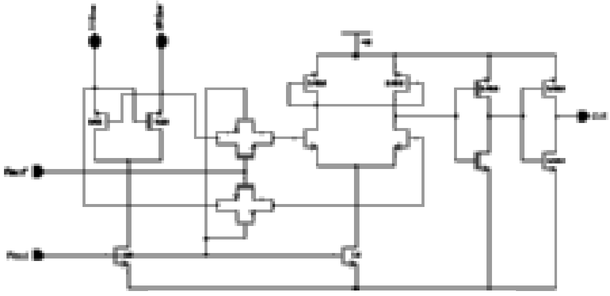
- Defined by implementation?
  - Registers = latches memory = SRAM/DRAM ?

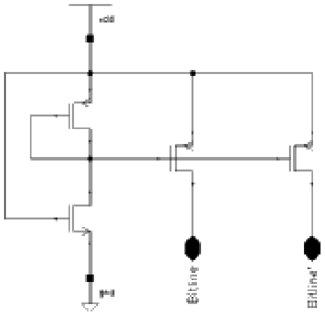












Register or Memory?

- Defined by implementation?
  - Registers = latches memory = SRAM/DRAM ?
  - Why use these structures and when?
- Latches
  - Low latency
  - High frequency
  - High power
  - Large area
- SRAM
  - Higher latency
  - Lower Frequency
  - Tricky to design
  - Small bit area
  - Amortized periphery area

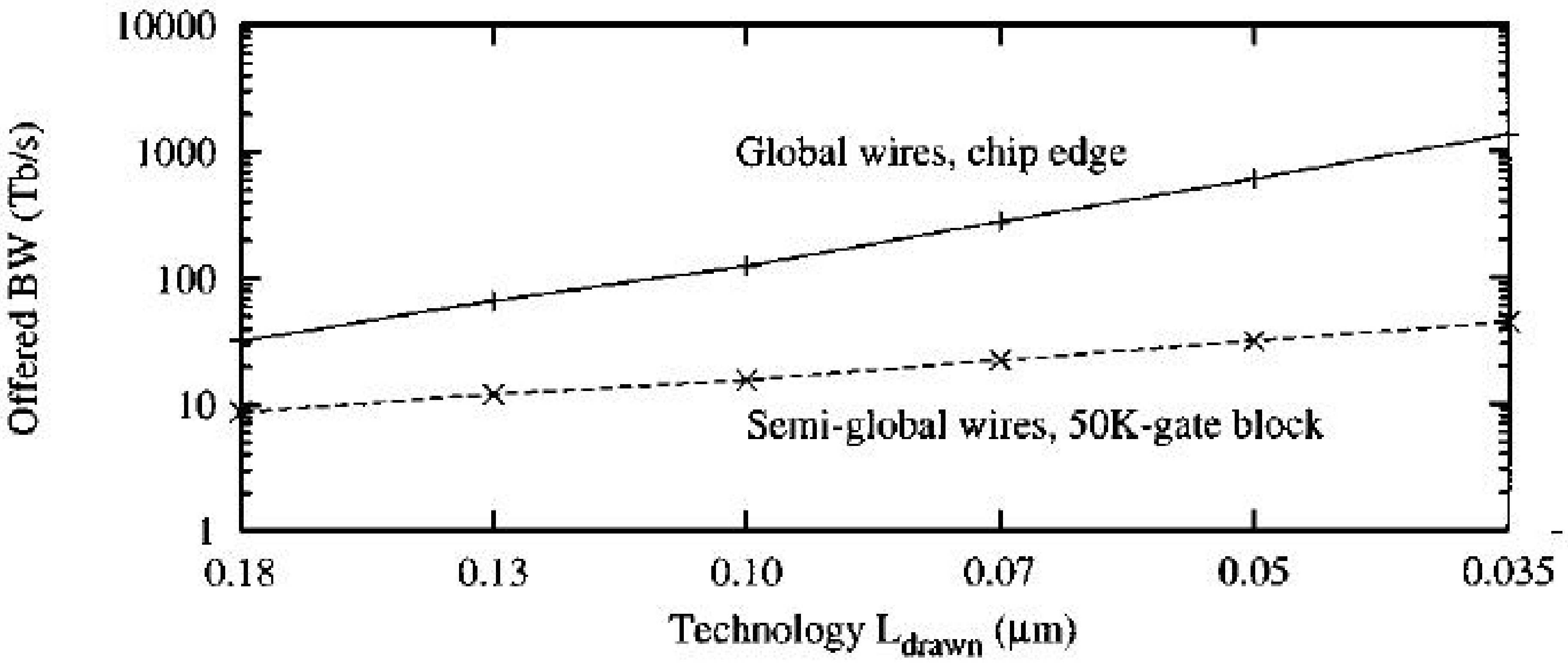
Register or Memory?

- Defined by implementation?
  - Registers = latches   memory = SRAM/DRAM ?
  - Why use these structures and when?
- Latches only used for pipeline registers today!
- Defined by use
  - Registers = small and fast
  - Memory = larger and slower

Board Time

- Wires
- Caches
  - What?
  - Why?
  - How?
- Exploiting locality in CPUs
  - Cache-aware programming

Future of Wires BW Estimates


Offered BW (Tb/s)
Global wires, chip edge
Semi-global wires, 50K-gate block
Technology L_{drawn} (µm)