ESE532: System-on-a-Chip Architecture

Day 7: September 23, 2019

Pipelining

Previously

- Pipelining in the large
  - Not just for gate-level circuits
- Throughput and Latency
- Form of parallelism

Today

Pipelining details (for gates, primitive ops)

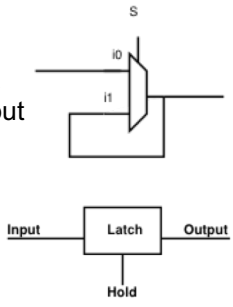
- Systematic Approach
- Justify Operator and Interconnect Pipelining
- Loop Bodies
- Cycles in the Dataflow Graph
- C-slow [bonus if have time]

Message

- Pipelining is an efficient way to reuse hardware to perform the **same** set of operations at high throughput

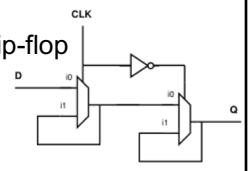
Latch

- Element that can hold a previous value of an input



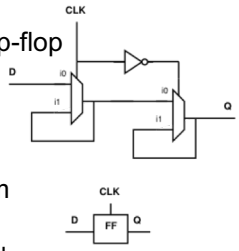
Register

- Use a pair to create a flip-flop
  - Also call register
- What happens when
  - CLK is low (0) ?
  - CLK is high (1) ?
  - CLK transitions from 0 to 1?



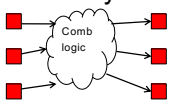
Register

- Use a pair to create a flip-flop
  - Also call register
- Sample D input on 0→1 transition of clock (CLK)
- Never an open path from D→Q
  - One of the mux latches always in hold state



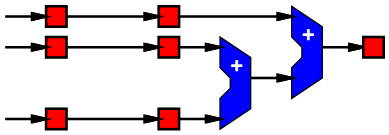
Synchronous Circuit Discipline

- Registers that sample inputs at clock edge and hold value throughout clock period
- Compute from registers-to-registers
- Cycle time large enough for longest logic path between registers
- Min cycle = Max path delay between registers



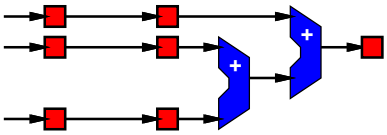
Preclass 1

- Delay between registers as shown?



Preclass 1

- Move registers so can clock at adder delay?

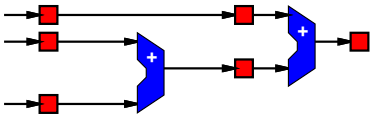


Pipeline Reuse

- Lower delay between clocks
  - Higher clock rate
  - Higher potential throughput
  - Faster we reuse our logic
  - More capacity get out of design
    - Assuming registers cheap in area and time overhead
      - T_{setup}, $T_{clk->q}$ ~ 20ps, T_{add} ~ 500ps
      - Registers ~ 10 transistors/bit
      - Adder ~ 40—50 transistors/bit

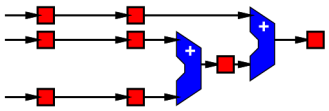
Note Registers on Links

- Some links end up with multiple registers.
- Why?

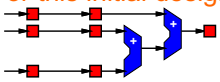


Preclass 2: What Happens?

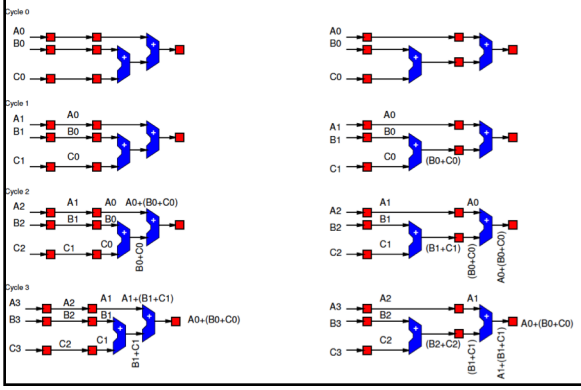
- What would be wrong with this pipelining?



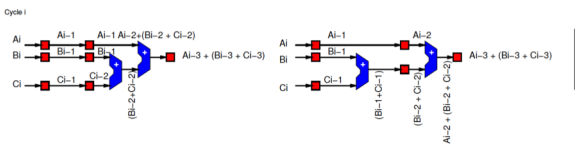
- For this initial design:



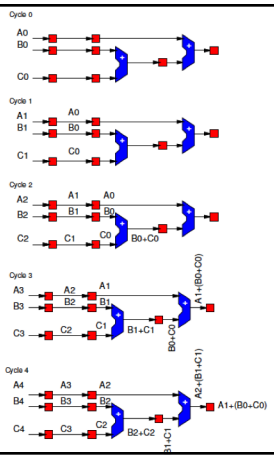
Behavior



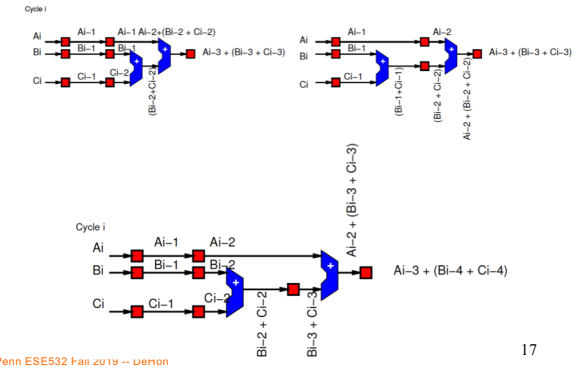
Equations



Behavior Add Register



Equations

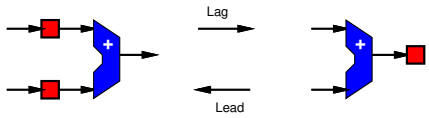


Consistent Pipelining

- Makes sure a consistent input set arrives at each gate/operator
  - Don't get mixing between input sets

Legal Register Moves

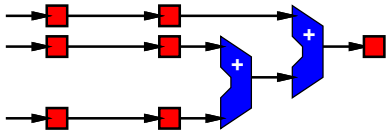
- Retiming Lag/Lead



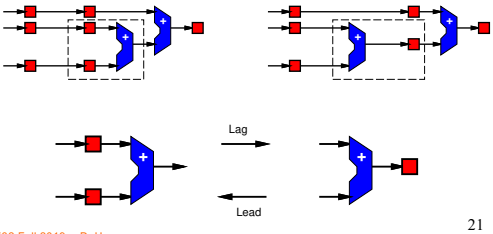
- Lag: remove register every input
  add register every output
- Lead: remove register every output
  add register every input

Preclass 1

- Retime using Lead/Lag



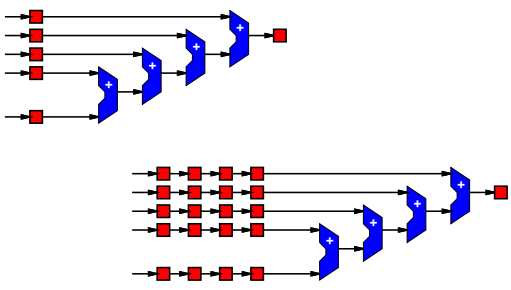
Preclass 1 (revisited)



Add Registers and Move

- If we're willing to add pipeline delay
  - Add any number of pipeline registers at input
  - Move registers into circuit to reduce cycle time
    - Reduce max delay between registers

Add Registers at Input



Add Register and Retime

- Add chain of registers on every input
- Retime registers into circuit
  - Minimizing delay between registers

Add Registers and Retime

- Lets us think about behavior
  - What the pipelining is doing to cycles of delay
- Separate from details of how redistribute registers
- Behavioral equivalence between the registers-at-front and properly retimed version of circuit

Justify Pipelining

(or composing pipelined operators)

Handling Pipelined Operators

- Given a pipelined operator
  - (or a pipelined interconnect)
- Discipline of picking a frequency target and designing everything for that
  - May be necessary to pipeline operator since its delay is too high
- Due to hierarchy
  - Pipelined this operator and now want to use it as a building block

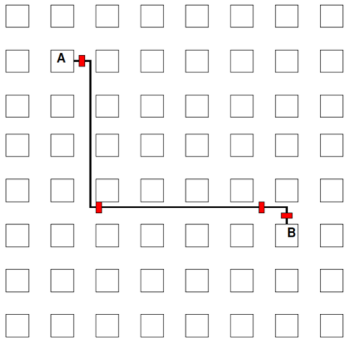
Examples

- Run at 500MHz
- Floating-point unit that takes 9ns
  - Can pipeline into 5, 2ns stages
- Multiplier that takes 6ns
- Memory can access in 2ns
  - Only if registers on address/inputs and output
  - i.e. exist in own clock stage

Interconnect Delay

- Chips >> Clock Cycles
- May have chip 100s of Operators wide
- May only be able to reach across 10 operators in a 2ns cycle
- Must pipeline long interconnect links

Interconnect Example

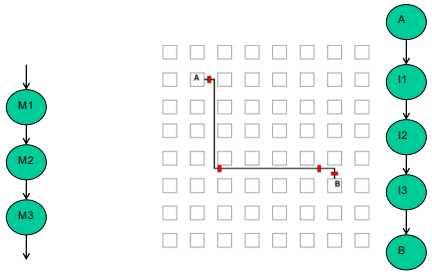


Methodology: Pipelined Operator Graph

- Start with logical, unpipelined graph
- Treat each pipelined operator as a set of unit-delay operators of mandatory depth
- Treat each interconnect pipeline stage as a unit-delay buffer
- Add registers at input
- Retime into graph

Model

- 3-stage Multiplier
- Interconnect Delay

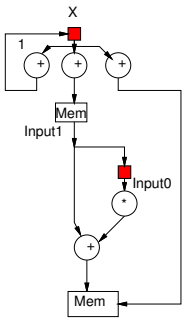


Pipeline Loop

(and use for justify pipeline example)

Preclass 4

- Logical (unpipelined) dataflow graph for loop body

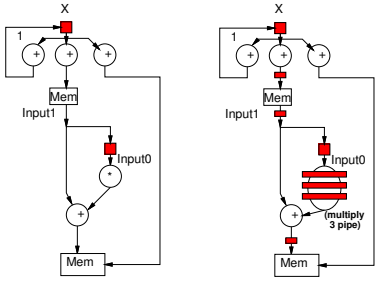


Example Operators

- Operator and Interconnect delays
  - Multiplier 3 cycles
  - Reading from Input array
    - Memory op is cycle after computing address
    - Takes one cycle delay bring data back to multiplier

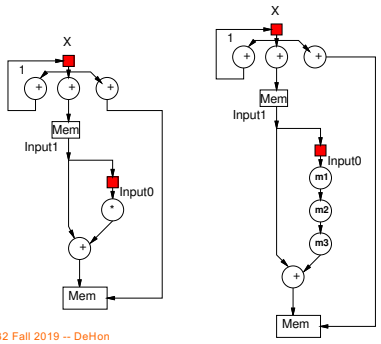
Illustrate Need

- What happens if just use graph as is (with operators pipelined as required)?



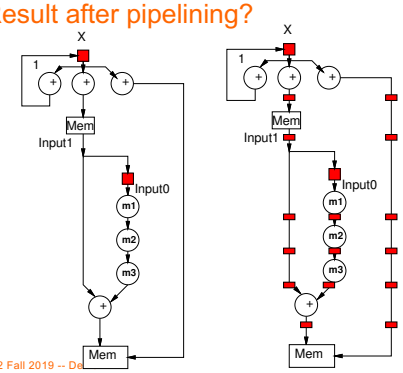
Model Graph

- Revised graph for modeling



Pipeline Graph

- Result after pipelining?



Pipelining Lesson

- Can always pipeline an **acyclic** graph to fixed frequency target
  - fixed pipelining of primitive operators
  - Pipeline interconnect delays
- Need to keep track of registers to balance paths
  - So see consistent delays to operators

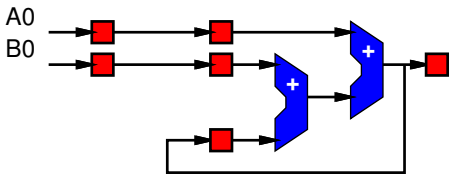
Graph Cycles

Watch: Clock cycle
Cycle time

Cycle in Graph

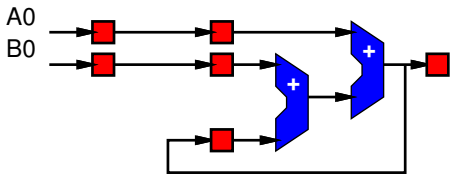
Preclass 3

- Can we retime to reduce cycle time



Retiming Limits?

- What prevents us from retiming?

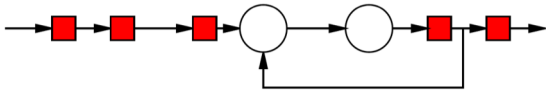


(Graph) Cycle Observation

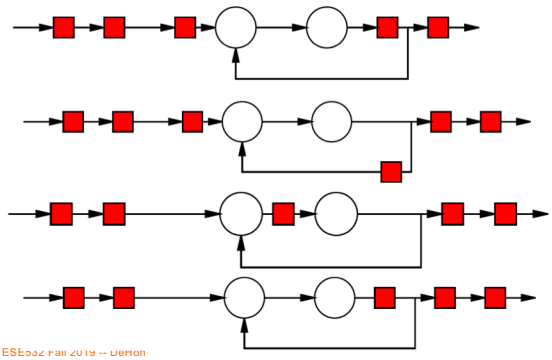
- Retiming does not allow us to change the *number of registers inside a cycle*.
- Limit to cycle time
  - Max delay in cycle / Registers in cycle
- Pipelining doesn't help inside cycle
  - Cannot push registers into cycle

Simple Cycle

- Delay of cycle?
- Registers in cycle?
- What happens to cycle if try to apply lead/lag?



Retiming

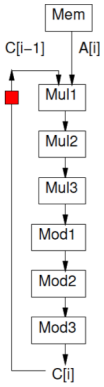


Loop

- Consider
  - [multiply and mod each take 3 cycles]
- For (i=0;i<N;i++)
  C[i]=(C[i-1]*A[i])%N;

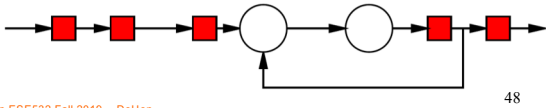
Loop

- For (i=0;i<N;i++)
  C[i]=(C[i-1]*A[i])%N;



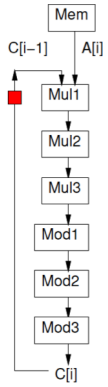
Initiation Interval (II)

- Cyclic dependencies can limit throughput
- Due to data dependent cycles in graph,
  - May not be able to initiate a new computation on every clock cycle
- II – clock cycles (delay) before can initiate
- Throughput = 1/II



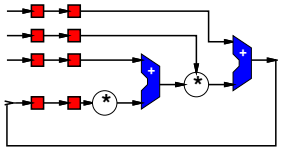
Loop

- For (i=0;i<N;i++)
  C[i]=(C[i-1]*A[i])%N;
- Initiation Interval?



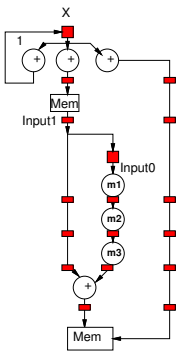
Initial Interval

- Delay Around graph cycle?
  - Assume multiply 3, add 1
- Registers in graph cycle?
- Retiming clock cycle bound = II ?
- Achievable?



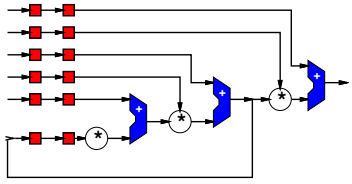
II and Latency

- Actually is a cycle
  - II?
  - Latency?



II and Latency

- II? (assume willing to pipeline inputs)
- Latency?



C-Slow

(if time permits)

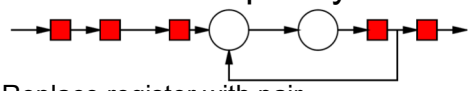
Problem

- Pipelining cannot push registers into cycle
- Graph cycles can prevent running at full pipeline target (maximum frequency)
- If not reusing operators at full pipeline target are underutilizing resources
- Can we use the resources for something?

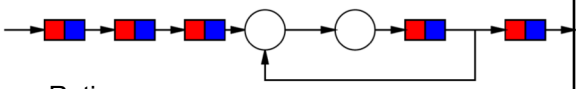
C-Slow

- **Observation:** if we have data-level parallelism, can use to solve independent problems on same hardware
- **Transformation:** make C copies of each register
- **Guarantee:** C computations operate independently
  - Do not interact with each other

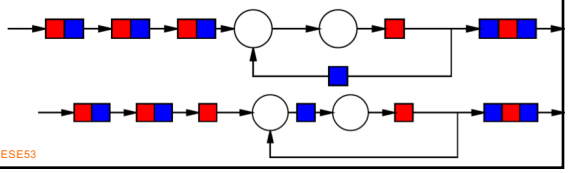
2-Slow Simple Cycle



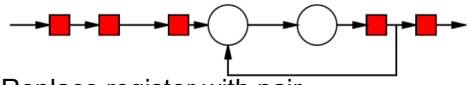
- Replace register with pair



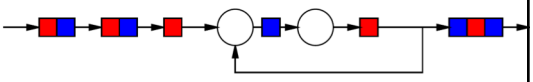
- Retime



2-Slow Simple Cycle



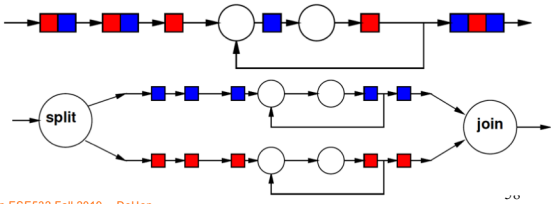
- Replace register with pair
- Retime



- Observe independence of red/blue computations

Equivalence

- The 2-slow operator is equivalent to two data parallel operators running at half the speed



Automation

- No mainstream tool today will perform C-slow transformation for you automatically
- Synthesis tools will retime registers

Lesson

- Cyclic dependencies limit throughput on single task or data stream
  - Cycle-length / registers-in-cycle
- Can use on C independent (data parallel) tasks

Big Ideas

- Pipeline computations to reuse hardware and maximize computational capacity
- Can compose pipelined operators and accommodate fixed-frequency target
  - Be careful with data retiming
- Cycles limit pipelining on single stream
- C-slow to share hardware among multiple, data-parallel streams

Admin

- Reading for Day 8 on web
- HW4 due Friday
  - HW4 lighter, HW5 heavier
  - Suggest planning to get early start on HW5