ESE532: System-on-a-Chip Architecture

Day 2: September 9, 2020

Analysis, Metrics, and Bottlenecks

Work Preclass
Lecture start 10:35am
(preclass, feedback linked to syllabus)

Today: Analysis

- How do we quickly estimate what's possible?
  - Before developing a complete solution
    - less effort than developing complete solution
- How should we attack the problem?
  - Achieve the performance, energy goals?
- When we don't like the performance we're getting, how do we understand it?
- Where should we spend our time?

Today: Analysis

- Part 1: Key Terms and Concepts
  - Throughput
  - Latency
  - Bottleneck
- Part 2: Broader view
  - Bottleneck
  - Computation as a Graph, Sequence
  - Critical Path
- Part 3: Time and Space
- Part 4: Limits
  - Resource Bound
  - And Critical Path Bound
  - 90/10 Rule (time permitting)

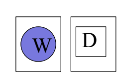
Message for Day

- Identify the **Bottleneck**
  - May be in compute, I/O, memory, data movement
- Focus and reduce/remove bottleneck
  - More efficient use of resources
  - More resources
- Repeat

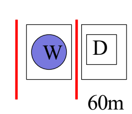
Latency vs. Throughput

- **Latency:** Delay from inputs to output(s)
- **Throughput:** Rate at which can produce new set of outputs
  - (alternately, can introduce new set of inputs)

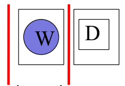
Preclass Washer/Dryer Example



- 10 shirt capacity
- 1 Washer Takes 30 minutes
- 1 Dryer Takes 60 minutes
- How long to do one load of wash?
  - → Wash latency
- Cleaning Throughput?



Pipeline Concurrency



- Break up the computation graph into stages
  - Allowing us to
    - reuse resources for new inputs (data),
    - while older data is still working its way through the graph
      - Before it has exited graph
  - Throughput > (1/Latency)
- Relate liquid in pipe
  - Doesn't wait for first drop of liquid to exit far end of pipe before accepting second drop

Penn ESE532 Spring 2020 -- DeHon

7

Escalator

Image Source: https://commons.wikimedia.org/wiki/File:Tanforan_Target_escalator_1.JPG

Penn ESE532 Spring 2020 -- DeHon

8

Escalator

- Moves 2 ft/second
- Assume for simplicity one person can step on escalator each second
- Escalator travels 30 feet (vertical and horizontal)
- Latency of escalator trip?
- Throughput of escalator: people/hour ?

Penn ESE532 Spring 2020 -- DeHon

9

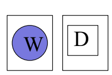
Bottleneck

- What is the rate limiting item?
  - Resource, computation, ….

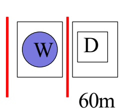
Penn ESE532 Spring 2020 -- DeHon

10

Preclass Washer/Dryer Example



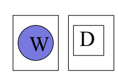
- 1 Washer Takes 30 minutes
  - Isolated throughput 20 shirts/hour
- 1 Dryer Takes 60 minutes
  - Isolated throughput 10 shirts/hour
- Where is bottleneck in our cleaning system?



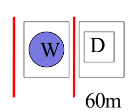
Penn ESE532 Spring 2020 -- DeHon

11

Preclass Washer/Dryer Example



- 1 Washer $500
  - Isolated throughput 20 shirts/hour
- 1 Dryer $500
  - Isolated throughput 10 shirts/hour
- How do we increase throughput with $500 investment

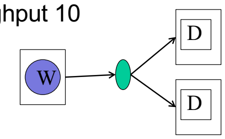


Penn ESE532 Spring 2020 -- DeHon

12

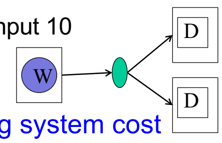
Preclass
Washer/Dryer Example

- 1 Washer $500
  - Isolated throughput 20 shirts/hour
- 2 Dryers $500
  - Isolated single dryer throughput 10 shirts/hour
- Latency?
- Throughput?

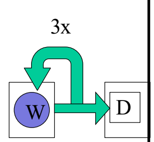


Preclass
Washer/Dryer Example

- 1 Washer $500
  - Isolated throughput 20 shirts/hour
- 2 Dryers $500
  - Isolated single dryer throughput 10 shirts/hour
- Able to double the throughput without doubling system cost



Preclass
Stain Example



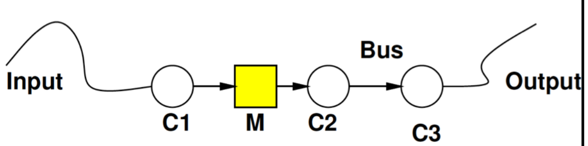
- 1 Washer Takes 30 minutes
  - Isolated throughput 20 shirts/hour
- 1 Dryer Takes 60 minutes
  - Isolated throughput 10 shirts/hour
- Shirt need 3 wash cycles
- Latency?
- Throughput?
  - (assuming reuse single washer)

Beyond Computation

(Part 2: Broader View)

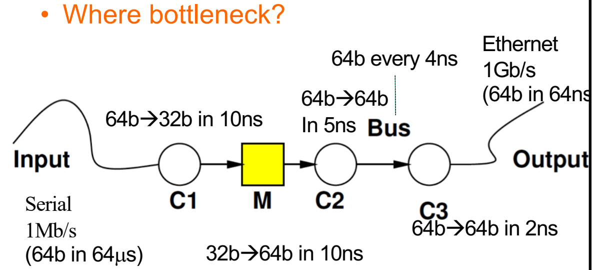
Bottleneck

- May be anywhere in path
  - I/O, compute, memory, data movement



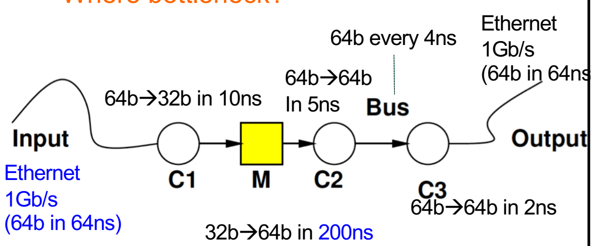
Bottleneck

- Where bottleneck?



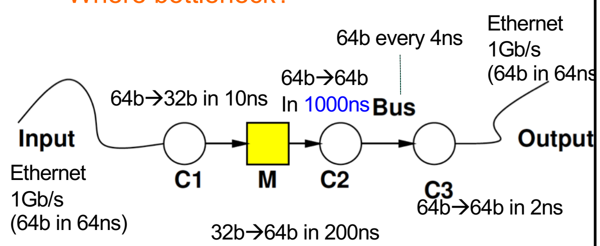
Bottleneck

- Where bottleneck?



Bottleneck

- Where bottleneck?



Feasibility / Limits

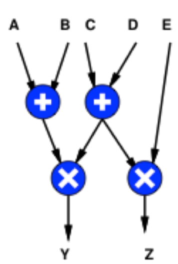
- First things to understand
  - Obvious limits in system?
- Impossible?
- Which aspects will demand efficient mapping?
- Where might there be spare capacity

Generalizing

(to more general task graphs)

Computation as Graph

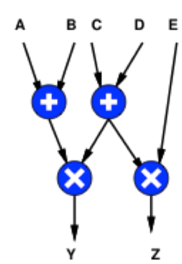
- Shown "simple" graphs (pipelines) so far
- Y=(A+B)*(C+D)
- Z=(C+D)*E



Note: HW2 ask you to draw a dataflow graph.
Here's an example…more to come.

Computation as Graph

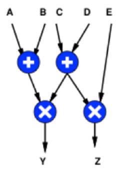
- Nodes have multiple input/output edges
- Edges may fanout
  - Results go to multiple successors



Computation as Sequence

- Shown “simple” graphs (pipelines) so far
- Y=(A+B)*(C+D)
- Z=(C+D)*E

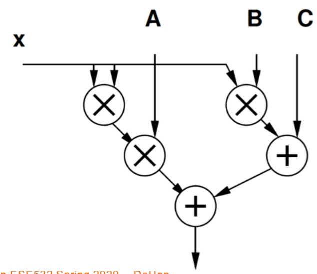
T1=A+B
T2=C+D
Y=T1*T2
Z=T1*E



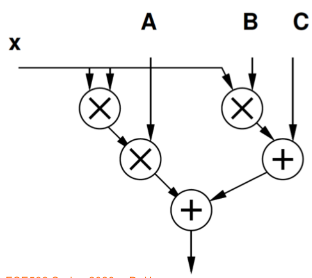
Computation as Graph

- $Y=Ax^2+Bx+C$

T1=x*x
T2=A*T1
T3=B*x
T4=T2+T3
Y=C+T4



Computation as Graph

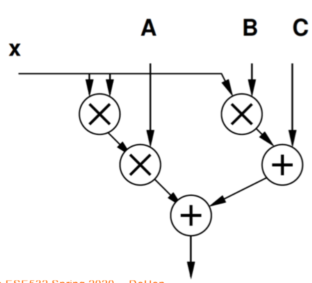


- Latency multiply = 1
- Latency add = 1/3
- Latency from B to output?
- Latency from x to output?
  - Through Ax^2 ?
  - Through Bx ?

Delay in Graphs

- **Observe:** There are multiple paths from inputs to outputs
- Need to complete all of them to produce outputs
- Limited by longest path
- **Critical path:** longest path in the graph

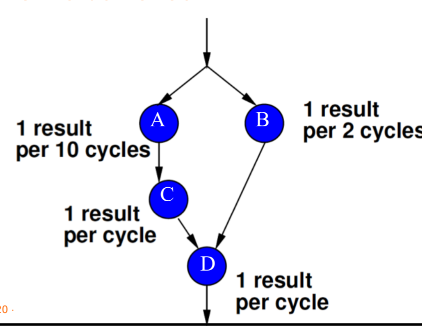
Computation as Graph



- Latency multiply = 1
- Latency add = 1/3
- Critical Path?

Bottleneck

- Where is the bottleneck?



Time and Space

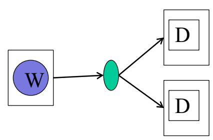
(Part 3)

Space

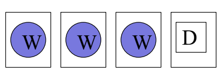
- “Space” is an abstract term for physical resources
  - On VLSI chip: Area – mm^2 of silicon
  - On our FPGA: # of LUTs used
  - More abstractly: # of Adders, multipliers
  - Laundry example
    - $$ to spend on laundry equipment
    - Physical space (sq. ft) in laundry room

Space-Time

- In general, we can spend resources to reduce time
  - Increase throughput

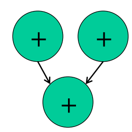


Three wash stain removal case

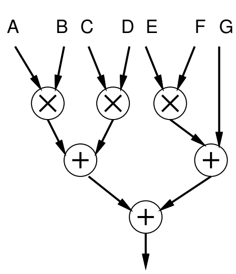


Space Time

- Computation
  - A=x0+x1
  - B=x2+x3
  - C=A+B
- Adder takes one cycle
- Latency on one adder?
- Latency on 3 adders?



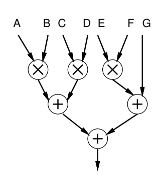
Computation as Graph



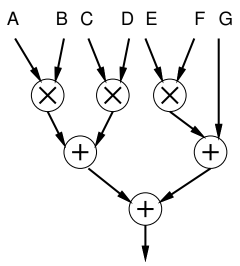
- Latency multiply = 1
- Space multiply = 3
- Latency add = 1
- Space add = 1
- Latency and Space
  - 3 mul, 2 add

Schedule 3 mul, 2 add

Cycle	Mul	Mul	Mul	Add	Add
0	A*B	C*D	E*F		
1				A*B+C*D	E*F+G
2				(A*B+C*)D +(E*F+G)	



Computation as Graph

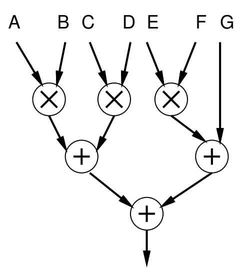


- Latency multiply = 1
- Space multiply = 3
- Latency add = 1
- Space add = 1
- Latency and Space
  – 2 mul, 1 add

Schedule 2 mul, 1 add

Cycle	Mul	Mul	Add
0	A*B	C*D	
1	E*F		(A*B+C*D)
2			E*F+G
3			(A*B+C*D)+ (E*F+G)

Computation as Graph

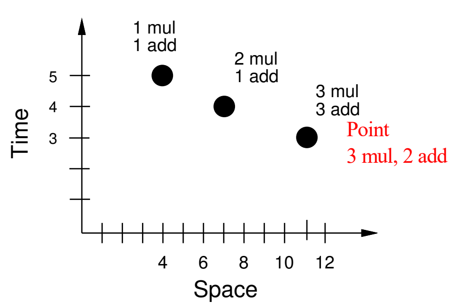


- Latency multiply = 1
- Space multiply = 3
- Latency add = 1
- Space add = 1
- Latency and Space
  – 1 mul, 1 add

Schedule 1 mul, 1 add

Cycle	Mul	Add
0	A*B	
1	C*D	
2	E*F	A*B+C*D
3		E*F+G
4		(A*B+C*D)+(E*F+G)

Space-Time Graph



Two Bounds

Part 4: Limits

(still in Time and Space)

Problem

- Coming up with an exact time count can be hard (human/computer time consuming)
  - Technically a hard problem
    - NP-Complete: no known non-exponential solution
- Requires reasoning about structure of graph
- Would be nice to have a quick (easy) answer on what is feasible
  - …and what is not feasible → impossible.

Bounds

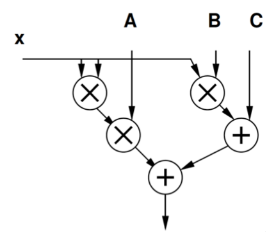
- Establish the feasible range
  - Must be larger than LB
  - Must be smaller than UB
  - Solution will be between LB and UB
  - $LB \leq ActualTime \leq UB$
- Bounds in sports
  - Ball landing in-bounds or out-of bounds

Bounds

- Quick **lower** bounds (LB) can estimate
  - $LB \leq ActualTime$
- Two:
  - CP: Critical Path
    - Sometimes call it "Latency Bound"
  - RB: Resource Capacity Bound
    - Sometimes call it "Throughput Bound" or "Compute Bound"

Critical Path Lower Bound

- Critical path assuming infinite resources
- Certainly cannot finish any faster than that
- CP$\leq ActualTime$
- Ignores resource limits



Resource Capacity Lower Bound

- Sum up all capacity required per resource: $\text{TotalOps} = \sum Ops$
  - E.g. number of multiplications, additions, memory lookups
- Divide by total resource (for type)
  - E.g., number of multipliers, adders, memory ports
  - $RB = \lceil TotalOps/Operators \rceil \leq ActualTime$
- Lower bound on compute
  - (best can do is pack all use densely)
  - Ignores data dependency constraints

Multiple Resource Types

- $RB = Max(\lceil TotalOps_1/Operators_1 \rceil, \lceil TotalOps_2/Operators_2 \rceil, \ldots.) \leq ActualTime$
- Combine Critical Path Lower Bound

$$Max(CP, \lceil TotalOps_1/Operators_1 \rceil, \lceil TotalOps_2/Operators_2 \rceil, \ldots.) \leq ActualTime$$

For Single Resource Type

- (and no communication time…)
- Can use to get upper bound:
- $ActualTime \leq CP + RB$
- Together:
- $Max(CP, RB) \leq ActualTime \leq CP + RB$

Washer-Dryer Bounds

- Task: wash & dry 30 shirts
- Washer: 10 shirts/30 min.
- Dryer: 10 shirts/60 min.
- 2 Washers, 2 Dryers
- W-->D CP=90 minutes

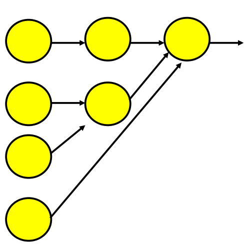
Washer-Dryer Bounds

- Task: wash & dry 30 shirts CP=90 min
- Washer: 10 shirts/30 min.
- Dryer: 10 shirts/60 min.
- 2 Washers, 2 Dryers
- Washer Bound:
  - WB= ⌈30 shirts/2 washers×10 shirts/washer⌉ × 30 min = 60 min
- Dryer
  - DB=⌈(30 shirts/2 dryers × 10 shirts/dryer)⌉× 60 min = 120 minutes

Washer-Dryer Bounds

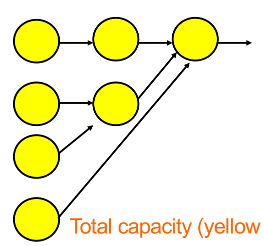
- Task: wash & dry 30 shirts CP=90 min
- Washer: 10 shirts/30 min.
- Dryer: 10 shirts/60 min.
- 2 Washers, 2 Dryers RB=120 min
- Max(90,60,120)=120≤ $TaskTime$
- TaskTime: 150
  - 0 start 2 washes
  - 30 start 2 dryers; start 1 wash
  - 90 finish 2 dryers; start last dryer load
  - 150 finish last dryer load

Example



Critical Path

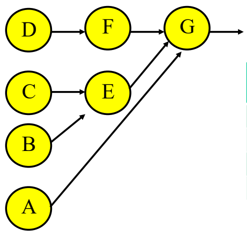
Example



Total capacity (yellow circle evaluations) needed?

Resource Bound (2 resources)?

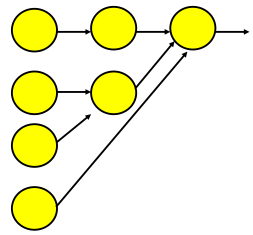
Example



Cycle	Resource 1	Resource 2
0	A	B
1	C	D
2	E	F
3	G	

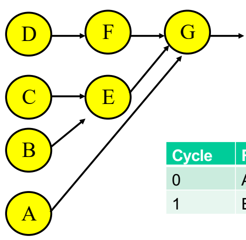
Resource Bound (2 resources)?

Example



Resource Bound (4 resources)?

Example



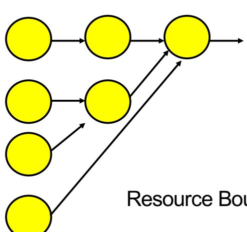
Cycle	R1	R2	R3	R4
0	A	B	C	D
1	E	F	G	

Legal Schedule?

Resource Capacity Lower Bound

- Sum up all capacity required per resource: $\text{TotalOps} = \sum Ops$
  - E.g. number of multiplications, additions, memory lookups
- Divide by total resource (for type)
  - E.g., number of multipliers, adders, memory ports
  - $RB = \lceil TotalOps/Operators \rceil \leq ActualTime$
- Lower bound on compute
  - (best can do is pack all use densely)
  - Ignores data dependency constraints

Example



Critical Path 3

Resource Bound (2 resources) 7/2=4

Resource Bound (4 resources) 7/4=2

Either one (CP,RB) can be limit. Check both.
In general, independent → relation depends on task.

What are the telling us

- If CP<RB
  - Adding resources (space) may be effective at reducing latency
- If RB<CP
  - Adding resources (space) will not reduce latency

90/10 Rule (of Thumb)

- Observation that code is not used uniformly
- 90% of the time is spent in 10% of the code
- Knuth: 50% of the time in 2% of the code
- Implications
  - There will typically be a bottleneck
  - We don't need to optimize everything
  - We don't need to uniformly replicate space to achieve speedup
  - Not everything needs to be accelerated

Big Ideas

- Identify the Bottleneck
  - May be in compute, I/O, memory ,data movement
- Focus and reduce/remove bottleneck
  - More efficient use of resources
  - More resources

Admin

- Reading for Day 3 on web
- HW1 due Friday
- HW2 out
  - Partner assignment (will add tonight) (see canvas)
- Remember feedback
- Remaining Questions?