Today

- Retiming Demand
  - Folded Computation
  - Logical Pipelining
  - Physical Pipelining

- Retiming Supply
  - Technology
  - Structures
  - Hierarchy

Retiming Demand

Image Processing

- Many operations can be described in terms of 2D window filters
  - Compute value as a weighted sum of the neighborhood
  - $\text{Out}[x][y] = in[x][y] \ast c[1][1] + in[x-1][y] \ast c[0][1] + \ldots$

- Blurring, edge and feature detection, object recognition and tracking, motion estimation, VLSI design rule checking

Preclass 2

- Describes window computation:
  1: for $x=0$ to $N-1$
  2: for $y=0$ to $N-1$
  3: $\text{out}[x][y]=0$
  4: for $wx=0$ to $W-1$
  5: for $wy=0$ to $W-1$
  6: $\text{out}[x][y] += in[x+wx][y+wy] \ast c[wx][wy]$

- How many times is each $in[x][y]$ used?
  1: for $x=0$ to $N-1$
  2: for $y=0$ to $N-1$
  3: $\text{out}[x][y]=0$
  4: for $wx=0$ to $W-1$
  5: for $wy=0$ to $W-1$
  6: $\text{out}[x][y] += in[x+wx][y+wy] \ast c[wx][wy]$

Preclass 2

• Sequentialized on one multiplier,
  – Distance between in[x][y] uses?
1: for x=0 to N-1
2:  for y=0 to N-1
3:   out[x][y]=0;
4:  for wx=0 to W-1
5:   for wy=0 to W-1
6:     out[x][y]+=in[x+wx][y+wy]*c[wx][wy]

Parallel Scaling/Spatial Sum

• Compute entire window at once
  – More hardware, fewer cycles

Fully Spatial

• What if we gave each pixel its own processor?
Flop Experiment #1

- Pipeline/C-slow/retime to single LUT delay per cycle
  - MCNC benchmarks to 256 4-LUTs
  - no interconnect accounting

<table>
<thead>
<tr>
<th>Number of Registers</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage</td>
<td>72</td>
<td>16</td>
<td>4.5</td>
<td>2.2</td>
<td>1.3</td>
<td>0.96</td>
<td>1.2</td>
<td>0.46</td>
<td>0.12</td>
<td>0.11</td>
</tr>
</tbody>
</table>

- average 1.7 registers/LUT (some circuits 2--7)

Long Interconnect Path

- What happens if one of these links ends up on a long interconnect path?

Pipeline Interconnect Path

- To avoid cycle being limited by longest interconnect
  - Pipeline network

Chips >> Cycles

- Chips growing
- Gate delays shrinking
- Wire delays aren’t scaling down
  -Will take many cycles to cross chip

Clock Cycle Radius

- Radius of logic can reach in one cycle (45 nm)
  - Radius 10 (preclass 20: $L_{seg}=5 \rightarrow 50\text{ps}$)
    - Few hundred PEs
  - Chip side 600-700 PE
    - 400-500 thousand PEs
  - 100s of cycles to cross

Pipelined Interconnect

- In what cases is this convenient?
Pipelined Interconnect

• When might it pose a challenge?

Long Interconnect Path

• What happens here?

Long Interconnect Path

• Adds pipeline delays
• May force further pipelining of logic to balance out paths
• More registers

Flop Experiment #1

• Pipeline/C-slow/retime to single LUT delay per cycle
– MCNC benchmarks to 256 4-LUTs
– no interconnect accounting

Number of Registers | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
Percentage           | 72 | 16 | 4.5 | 2.2 | 1.3 | 0.96 | 1.2 | 0.46 | 0.12 | 0.11 |
– average 1.7 registers/LUT (some circuits 2–7)

Reminder

Flop Experiment #2

• Pipeline and retime to HSRA cycle
  – place on HSRA
  – single LUT or interconnect timing domain
  – same MCNC benchmarks

Number of Registers | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | >10 |
Percentage           | 60 | 6.9 | 5.9 | 3.8 | 4.3 | 2.7 | 2.6 | 1.9 | 1.5 | 1.2 | 9.2 |
– average 4.7 registers/LUT

Retiming Requirements

• Retiming requirement depends on parallelism and performance
• Even with a given amount of parallelism
  – Will have a distribution of retiming requirements
    – May differ from task to task
  – May vary independently from compute/interconnect requirements

• Another balance issue to watch
  – Balance with compute, interconnect
• Need a canonical way to measure
  – Like Rent?

[Tsu et al., FPGA 1999]
Retiming Supply

Optional Output

- Flip-flop (optionally) on output
- flip-flop: 4-5Kλ²
- Switch to select: ~ 5Kλ²
- Area: 1 LUT (800K → 1Mλ²/LUT)
- Bandwidth: 1b/cycle

Output

- Single Output
  - Ok, if don’t need other timings of signal
- Multiple Output
  - more routing

Input

- More registers (K×)
  - 7-10Kλ²/register+mux
  - 4-LUT => 30-40Kλ²/depth
- No more interconnect than unretimed
  - open: compare savings to additional reg. cost
- Area: 1 LUT (1M+d*40Kλ²) get Kd regs
  - d=4, 1.2Mλ²
- Bandwidth: K/cycle
  - 1/d th capacity

Preclass 3

<table>
<thead>
<tr>
<th>Diagram</th>
<th>Description</th>
<th>Area per Block</th>
<th>Blocks Needed</th>
<th>Total Area</th>
<th>Best?</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Diagram" /></td>
<td>4-LUT with single optional flip-flop on output</td>
<td>1.03MA²</td>
<td>1.03MA²</td>
<td>0.03MA²</td>
<td></td>
</tr>
<tr>
<td><img src="image2.png" alt="Diagram" /></td>
<td>4-LUT with configurable depth output register with maximum depth 4</td>
<td>1.04MA²</td>
<td>1.04MA²</td>
<td>0.04MA²</td>
<td></td>
</tr>
<tr>
<td><img src="image3.png" alt="Diagram" /></td>
<td>4-LUT with configurable depth input registers with maximum depth 4</td>
<td>1.10MA²</td>
<td>1.10MA²</td>
<td>0.10MA²</td>
<td></td>
</tr>
</tbody>
</table>

Day 5

Some Numbers (memory)

- Unit of area = λ² (F=2λ)
  - [more next week]
- Register as stand-alone element ≈ 4Kλ²
  - e.g. as needed/used last lecture
- Static RAM cell = 1Kλ²
  - SRAM Memory (single ported)
- Dynamic RAM cell (DRAM process) ≈ 100λ²
- Dynamic RAM cell (SRAM process) ≈ 300λ²
Retiming Density

- LUT+interconnect \( \approx 1M\lambda^2 \)
- Register as stand-alone element \( \approx 4K\lambda^2 \)
- Static RAM cell \( \approx 1K\lambda^2 \)
  - SRAM Memory (single ported)
- Dynamic RAM cell (DRAM process) \( \approx 100\lambda^2 \)
- Dynamic RAM cell (SRAM process) \( \approx 300\lambda^2 \)
- Can have much more retiming memory per chip if put it in large arrays
  - ...but then cannot get to it as frequently

Retiming Structure Concerns

- Area: \( \lambda^2/\text{bit} \)
- Throughput: bandwidth (bits/time)
- Energy

Just Logic Blocks

- Most primitive
  - build flip-flop out of logic blocks
    - \( I \leftarrow D^*/\text{Clk} + I^*\text{Clk} \)
    - \( Q \leftarrow Q^*/\text{Clk} + I^*\text{Clk} \)
  - Area: 2 LUTs (800K \( \rightarrow 1M\lambda^2 \)/LUT each)
  - Bandwidth: 1b/cycle

Separate Flip-Flops

- Network flip flop w/ own interconnect
  + can deploy where needed
  - requires more interconnect
  + Vary LUT/FF ratio
    - Arch. Parameter
  - Assume routing \( \propto \) inputs
    - 1/4 size of LUT
  - Area: 200K\( \lambda^2 \) each
  - Bandwidth: 1b/cycle

Virtex SRL16

- Xilinx Virtex 4-LUT
  - Use as 16b shiftreg
- Area: \( \approx 1M\lambda^2/16 = 60K\lambda^2/\text{bit} \)
  - Does not need CLBs to control
- Bandwidth: 1b/2 cycle (1/2 CLB)
  - 1/16 th capacity

Register File Memory Bank

- From MIPS-X
  - \( 1K\lambda^2/\text{bit} + 500\lambda^2/\text{port} \)
  - Area(RF) = \( (d+6)(W+6)(1K\lambda^2 + \text{ports} \times 500\lambda^2) \)
  - \( w >> 6, d >> 6, I+o=2 \Rightarrow 2K\lambda^2/\text{bit} \)
  - \( w=1, d >> 6, I=o=4 \Rightarrow 35K\lambda^2/\text{bit} \)
  - comparable to input chain
  - More efficient for wide-word cases
**Xilinx CLB**

- Xilinx 4K CLB
  - as memory
  - works like RF
- Area: 1/2 CLB \( \frac{640K \lambda^2}{16} \approx 40K \lambda^2 \) /bit
  - but need 4 CLBs to control
- Bandwidth: 1b/2 cycle (1/2 CLB)
  - 1/16 th capacity

**Memory Blocks**

- SRAM bit \( \approx 1200 \lambda^2 \) (large arrays)
- DRAM bit \( \approx 100 \lambda^2 \) (large arrays)

- Bandwidth: W bits / 2 cycles
  - usually single read/write
  - 1/2^th capacity

**Dual-Ported Block RAMs**

- Virtex-6 Series 36Kb memories
- Stratix-4
  - 640b, 9Kb, 144Kb
  - Can put 1M/1K=1K bits in space of 4-LUT
  - Trade few 4-LUTs for considerable memory

**Hierarchy/Structure Summary**

- “Memory Hierarchy” arises from area/bandwidth tradeoffs
  - Smaller/cheaper to store words/blocks
    - (saves routing and control)
  - Smaller/cheaper to handle long retiming in larger arrays (reduce interconnect)
  - High bandwidth out of shallow memories
  - Applications have mix of retiming needs

**Modern FPGAs**

- Output Flop (depth 1)
- Use LUT as Shift Register (16)
- Embedded RAMs (9Kb,36Kb)
- Interface off-chip DRAM (~0.1—1Gb)
- No retiming in interconnect
  - …yet

**Modern Processors**

- DSPs have accumulator (depth 1)
- Inter-stage pipelines (depth 1)
  - Lots of pipelining in memory path...
- Reorder Buffer (4—32)
- Architected RF (16, 32, 128)
- Actual RF (256, 512…)
- L1 Cache (~64Kb)
- L2 Cache (~1Mb)
- L3 Cache (10-100Mb)
- Main Memory in DRAM (~10-100Gb)
Admin

- Final Exercise
  - Basic statement out
  - Will continue to refine and detail
  - One arch. Desc. to add by next Monday
- No new homework grading
- Reading for Wednesday on web

Big Ideas

[MSB Ideas]

- Tasks have a wide variety of retiming distances (depths)
  - Within design, among tasks
- Retiming requirements vary independently of compute, interconnect requirements (balance)
- Wide variety of retiming costs
  - $100\lambda^2 \rightarrow 1M\lambda^2$
- Routing and I/O bandwidth
  - big factors in costs
- Gives rise to memory (retiming) hierarchy