It took a lot of work, but this latest Linux patch enables support for machines with 4,096 CPUs, up from the old limit of 1,024.

Do you have support for smooth full-screen Flash video yet?

No, but who uses that?
Pitfall: Amdahl’s Law

Execution time after improvement =

\[
\frac{\text{affected execution time}}{\text{amount of improvement}} + \text{execution time unaffected}
\]

\[
T_{\text{improved}} = \frac{T_{\text{affected}}}{\text{improvement factor}} + T_{\text{unaffected}}
\]
Pitfall: Amdahl’s Law

Improving an aspect of a computer and expecting a proportional improvement in overall performance

\[ T_{\text{improved}} = \frac{T_{\text{affected}}}{\text{improvement factor}} + T_{\text{unaffected}} \]

Example: multiply accounts for 80s out of 100s

- How much improvement do we need in the multiply performance to get 5× overall improvement?

\[ 20 = \frac{80}{n} + 20 \]

– Can’t be done!
Scaling Example

Workload: sum of 10 scalars, and $10 \times 10$ matrix sum

• Speed up from 10 to 100 processors?

Single processor: Time = $(10 + 100) \times t_{\text{add}}$

10 processors

• Time = $100/10 \times t_{\text{add}} + 10 \times t_{\text{add}} = 20 \times t_{\text{add}}$
• Speedup = $110/20 = 5.5$ (55% of potential)

100 processors

• Time = $100/100 \times t_{\text{add}} + 10 \times t_{\text{add}} = 11 \times t_{\text{add}}$
• Speedup = $110/11 = 10$ (10% of potential)

Assumes load can be balanced across processors
Scaling Example

What if matrix size is $100 \times 100$?

Single processor: \( \text{Time} = (10 + 10000) \times t_{\text{add}} \)

10 processors
- \( \text{Time} = 10 \times t_{\text{add}} + 10000/10 \times t_{\text{add}} = 1010 \times t_{\text{add}} \)
- Speedup = \(10010/1010 = 9.9\) (99% of potential)

100 processors
- \( \text{Time} = 10 \times t_{\text{add}} + 10000/100 \times t_{\text{add}} = 110 \times t_{\text{add}} \)
- Speedup = \(10010/110 = 91\) (91% of potential)

Assuming load balanced
Goals for Today

How to improve System Performance?

• Instruction Level Parallelism (ILP)
• Multicore
  – Increase clock frequency vs multicore
• Beware of Amdahls Law

Next time:

• Concurrency, programming, and synchronization
Problem Statement

Q: How to improve system performance?

→ Increase CPU clock rate?
  → But I/O speeds are limited
    Disk, Memory, Networks, etc.

Recall: Amdahl’s Law

Solution: Parallelism
Instruction-Level Parallelism (ILP)

Pipelining: execute multiple instructions in parallel

Q: How to get more instruction level parallelism?

A: Deeper pipeline

- E.g. 250MHz 1-stage; 500Mhz 2-stage; 1GHz 4-stage; 4GHz 16-stage

Pipeline depth limited by...

- max clock speed (less work per stage ⇒ shorter clock cycle)
- min unit of work
- dependencies, hazards / forwarding logic
Instruction-Level Parallelism (ILP)

Pipelining: execute multiple instructions in parallel

Q: How to get more instruction level parallelism?

A: Multiple issue pipeline

- Start multiple instructions per clock cycle in duplicate stages

\[ \text{ALU/Br} \rightarrow \square \quad \square \quad \square \quad \ldots \]

\[ \text{LW/SW} \rightarrow \square \quad \square \quad \square \quad \ldots \]

Hazards 3
Static Multiple Issue

a.k.a. Very Long Instruction Word (VLIW)

Compiler groups instructions to be issued together
  • Packages them into “issue slots”

Q: How does HW detect and resolve hazards?
A: It doesn’t.
  → Simple HW, assumes compiler avoids hazards

Example: Static Dual-Issue 32-bit MIPS
  • Instructions come in pairs (64-bit aligned)
    – One ALU/branch instruction (or nop)
    – One load/store instruction (or nop)
MIPS with Static Dual Issue

Two-issue packets

- One ALU/branch instruction
- One load/store instruction
- 64-bit aligned
  - ALU/branch, then load/store
  - Pad an unused instruction with nop

<table>
<thead>
<tr>
<th>Address</th>
<th>Instruction type</th>
<th>Pipeline Stages</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>ALU/branch</td>
<td>IF  ID  EX  MEM  WB</td>
</tr>
<tr>
<td>n + 4</td>
<td>Load/store</td>
<td>IF  ID  EX  MEM  WB</td>
</tr>
<tr>
<td>n + 8</td>
<td>ALU/branch</td>
<td>IF  ID  EX  MEM  WB</td>
</tr>
<tr>
<td>n + 12</td>
<td>Load/store</td>
<td>IF  ID  EX  MEM  WB</td>
</tr>
<tr>
<td>n + 16</td>
<td>ALU/branch</td>
<td>IF  ID  EX  MEM  WB</td>
</tr>
<tr>
<td>n + 20</td>
<td>Load/store</td>
<td>IF  ID  EX  MEM  WB</td>
</tr>
</tbody>
</table>
Scheduling Example

Schedule this for dual-issue MIPS

**Loop:**

```
lw   $t0, 0($s1)  # $t0=array element
addu $t0, $t0, $s2  # add scalar in $s2
sw   $t0, 0($s1)  # store result
addi $s1, $s1, -4  # decrement pointer
bne  $s1, $zero, Loop  # branch $s1!=0
```

<table>
<thead>
<tr>
<th>ALU/branch</th>
<th>Load/store</th>
<th>cycle</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Loop:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nop</td>
<td>lw $t0, 0($s1)</td>
<td>1</td>
</tr>
<tr>
<td>addi $s1, $s1, -4</td>
<td>nop</td>
<td>2</td>
</tr>
<tr>
<td>addu $t0, $t0, $s2</td>
<td>nop</td>
<td>3</td>
</tr>
<tr>
<td>bne $s1, $zero, Loop</td>
<td>sw $t0, 4($s1)</td>
<td>4</td>
</tr>
</tbody>
</table>

\[
\frac{5\text{ inst}}{4\text{ cycles}} = 1.25 \text{ IPC} = 1.25 \times 0.8 = 1.00
\]
Scheduling Example

Compiler scheduling for dual-issue MIPS...

Loop:  

lw  $t0, 0($s1)  # $t0 = A[i]
lw  $t1, 4($s1)  # $t1 = A[i+1]
addu $t0, $t0, $s2  # add $s2
addu $t1, $t1, $s2  # add $s2
sw  $t0, 0($s1)  # store A[i]
sw  $t1, −4($s1)  # store A[i+1]
addi $s1, $s1, +8  # increment pointer
bne $s1, $s3, TOP  # continue if $s1!=end

ALU/branch slot

Loop:  
nop
nop
addu $t0, $t0, $s2
addu $t1, $t1, $s2
addi $s1, $s1, +8
bne $s1, $s3, TOP

Load/store slot

lw  $t0, 0($s1)  1
lw  $t1, 4($s1)  2
nop  3
sw  $t0, 0($s1)  4
sw  $t1, 4($s1)  5
nop  6

\[ \frac{6}{6} = 0.75 \text{ CPI} \]
Scheduling Example

Compiler scheduling for dual-issue MIPS...

Loop:  lw  $t0, 0($s1)  # $t0 = A[i]
lw  $t1, 4($s1)  # $t1 = A[i+1]
addu $t0, $t0, $s2  # add $s2
addu $t1, $t1, $s2  # add $s2
sw  $t0,-8($s1)  # store A[i]
sw  $t1,-4($s1)  # store A[i+1]
addi $s1, $s1, +8  # increment pointer
bne $s1, $s3, TOP  # continue if $s1!=end

ALU/branch slot
Loop:  nop
addi $s1, $s1, +8
addu $t0, $t0, $s2
addu $t1, $t1, $s2
bne $s1, $s3, Loop

Load/store slot  cycle
lw  $t0, 0($s1)  1
lw  $t1, 4($s1)  2
nop  3
sw  $t0, -8($s1)  4
sw  $t1, -4($s1)  5

$\frac{5}{8} = 0.625$
Limits of Static Scheduling

Compiler scheduling for dual-issue MIPS...

lw $t0, 0($s1)  # load A
addi $t0, $t0, +1  # increment A
sw $t0, 0($s1)  # store A
lw $t0, 0($s2)  # load B
addi $t0, $t0, +1  # increment B
sw $t0, 0($s2)  # store B

ALU/branch slot
nop
nop
addi $t0, $t0, +1
nop
nop
addi $t0, $t0, +1
nop

Load/store slot  cycle
lw $t0, 0($s1)  1
lw $t0, 0($s2)  5
sw $t0, 0($s1)  4
sw $t0, 0($s2)  8

nop
nop
nop
nop
nop
nop
nop
nop
Limits of Static Scheduling

Compiler scheduling for dual-issue MIPS...

lw  $t0, 0($s1)  # load A
addi $t0, $t0, +1  # increment A
sw  $t0, 0($s1)  # store A
lw  $t1, 0($s2)  # load B
addi $t1, $t1, +1  # increment B
sw  $t0, 0($s2)  # store B

ALU/branch slot    Load/store slot    cycle
nop                  lw  $t0, 0($s1)    1
nop                  lw  $t1, 0($s2)    2
addi $t0, $t0, +1   nop                  3
addi $t1, $t1, +1   sw  $t0, 0($s1)    4
nop                  sw  $t1, 0($s2)    5

Problem: What if $s1 and $s2 are equal (aliasing)? Won’t work
Dynamic Multiple Issue

a.k.a. SuperScalar Processor (c.f. Intel)

- CPU examines instruction stream and chooses multiple instructions to issue each cycle
- Compiler can help by reordering instructions....
- ... but CPU is responsible for resolving hazards

Even better: Speculation/Out-of-order Execution

- Execute instructions as early as possible
- Aggressive register renaming
- Guess results of branches, loads, etc.
- Roll back if guesses were wrong
- Don’t commit results until all previous insts. are retired
Dynamic Multiple Issue

Instruction fetch and decode unit

Reservation station
Reservation station
Reservation station
Reservation station

Functional units
Integer
Integer
Floating point
Load-store

Commit unit

In-order issue
Out-of-order execute
In-order commit
Q: Does multiple issue / ILP work?

A: Kind of... but not as much as we’d like

Limiting factors?

• Programs dependencies
• Hard to detect dependencies → be conservative
  – e.g. Pointer Aliasing: A[0] += 1; B[0] *= 2;
• Hard to expose parallelism
  – Can only issue a few instructions ahead of PC
• Structural limits
  – Memory delays and limited bandwidth
• Hard to keep pipelines full
Question: Does multiple issue / ILP cost much?

Answer: Yes.

Dynamic issue and speculation requires power

<table>
<thead>
<tr>
<th>CPU</th>
<th>Year</th>
<th>Clock Rate</th>
<th>Pipeline Stages</th>
<th>Issue width</th>
<th>Out-of-order/Speculation</th>
<th>Cores</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>i486</td>
<td>1989</td>
<td>25MHz</td>
<td>5</td>
<td>1</td>
<td>No</td>
<td>1</td>
<td>5W</td>
</tr>
<tr>
<td>Pentium</td>
<td>1993</td>
<td>66MHz</td>
<td>5</td>
<td>2</td>
<td>No</td>
<td>1</td>
<td>10W</td>
</tr>
<tr>
<td>Pentium Pro</td>
<td>1997</td>
<td>200MHz</td>
<td>10</td>
<td>3</td>
<td>Yes</td>
<td>1</td>
<td>29W</td>
</tr>
<tr>
<td>P4 Willamette</td>
<td>2001</td>
<td>2000MHz</td>
<td>22</td>
<td>3</td>
<td>Yes</td>
<td>1</td>
<td>75W</td>
</tr>
<tr>
<td>UltraSparc III</td>
<td>2003</td>
<td>1950MHz</td>
<td>14</td>
<td>4</td>
<td>No</td>
<td>1</td>
<td>90W</td>
</tr>
<tr>
<td>P4 Prescott</td>
<td>2004</td>
<td>3600MHz</td>
<td>31</td>
<td>3</td>
<td>Yes</td>
<td>1</td>
<td>103W</td>
</tr>
<tr>
<td>Core</td>
<td>2006</td>
<td>2930MHz</td>
<td>14</td>
<td>4</td>
<td>Yes</td>
<td>2</td>
<td>75W</td>
</tr>
<tr>
<td>UltraSparc T1</td>
<td>2005</td>
<td>1200MHz</td>
<td>6</td>
<td>1</td>
<td>No</td>
<td>8</td>
<td>70W</td>
</tr>
</tbody>
</table>

Multiple simpler cores may be better?
Moore's Law

286
8088
8080
4004
386
Pentium
P4
Atom
Itanium 2
K8
K10
Dual-core Itanium 2

Curve shows 'Moore's Law': transistor count doubling every two years.
Why Multicore?

Moore’s law

- A law about transistors
- Smaller means more transistors per die
- And smaller means faster too

But: Power consumption growing too...
Surface of Sun
Rocket Nozzle
Nuclear Reactor
Hot Plate

Xeon

180nm
32nm
Power Wall

Power = capacitance * voltage\(^2\) * frequency

In practice: Power \(\sim\) voltage\(^3\)

Reducing voltage helps (a lot)

... so does reducing clock speed

Better cooling helps

The power wall

- We can’t reduce voltage further
- We can’t remove more heat
Why Multicore?

- **Single-Core**
  - Performance: 1.2x
  - Power: 1.7x
  - Overclocked +20%

- **Single-Core**
  - Performance: 1.0x
  - Power: 1.0x

- **Dual-Core**
  - Performance: 1.0x
  - Power: 1.02x
  - Underclocked -20%
Inside the Processor

AMD Barcelona Quad-Core: 4 processor cores
Inside the Processor

Intel Nehalem Hex-Core

4 wide pipeline
## Hyperthreading

### Multi-Core vs. Multi-Issue vs. HT

<table>
<thead>
<tr>
<th>Programs:</th>
<th>Num. Pipelines:</th>
<th>Pipeline Width:</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td>$\checkmark$</td>
<td>$1$</td>
</tr>
<tr>
<td>$N$</td>
<td>$\checkmark$</td>
<td>$1$</td>
</tr>
<tr>
<td>$1$</td>
<td></td>
<td>$N$</td>
</tr>
</tbody>
</table>

29
Hyperthreading

Multi-Core vs. Multi-Issue vs. HT

<table>
<thead>
<tr>
<th>Programs:</th>
<th>Num. Pipelines:</th>
<th>Pipeline Width:</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N )</td>
<td>( N )</td>
<td>( 1 )</td>
</tr>
<tr>
<td>( N )</td>
<td>( 1 )</td>
<td>( 1 )</td>
</tr>
<tr>
<td>( 1 )</td>
<td>( N )</td>
<td>( N )</td>
</tr>
</tbody>
</table>

Hyperthreads

- HT = Multilssue + extra PCs and registers – dependency logic
- HT = MultiCore – redundant functional units + hazard avoidance

Hyperthreads (Intel)

- Illusion of multiple cores on a single core
- Easy to keep HT pipelines full + share functional units
Example: All of the above

8 die
4 core/die
2 HT
Parallel Programming

Q: So lets just all use multicore from now on!
A: Software must be written as parallel program

Multicore difficulties

• Partitioning work
• Coordination & synchronization
• Communications overhead
• Balancing load over cores
• How do you write parallel programs?
  – ... without knowing exact underlying architecture?
Work Partitioning

Partition work so all cores have something to do
Load Balancing

Need to partition so all cores are actually working
Amdahl’s Law

If tasks have a **serial part** and a **parallel part**...

Example:

  step 1: divide input data into $n$ pieces
  step 2: do work on each piece
  step 3: combine all results

Recall: **Amdahl’s Law**

As number of cores increases ...

  • time to execute parallel part? goes to zero
  • time to execute serial part? Remains the same
  • **Serial part eventually dominates**
Amdahl’s Law
Q: So lets just all use multicore from now on!
A: Software must be written as parallel program

Multicore difficulties

- Partitioning work
- Coordination & synchronization
- Communications overhead
- Balancing load over cores
- How do you write parallel programs?
  - ... without knowing exact underlying architecture?
FlameWar Games Night Next Friday, April 27th

- 5pm in Upson B17
- Please come, eat, drink and have fun

No Lab4 or Lab Section next week!
PA3: FlameWar is due next Monday, April 23rd
• The goal is to have fun with it
• Recitations today will talk about it

HW6 Due next Tuesday, April 24th

Prelim3 next Thursday, April 26th