Performance and Power Analysis of Globally Asynchronous Locally Synchronous Multiprocessor Systems

Zhiyi Yu, Bevan M. Baas

VLSI Computation Lab,
ECE department, UC Davis
Outline

• Motivation
• Experimental platform: a GALS array processor
• Performance analysis of GALS multiprocessors
• Power analysis of GALS multiprocessors
Why GALS Chip Multiprocessor

- Why GALS clocking style
  - The challenge of globally synchronous systems
  - The challenge of totally asynchronous systems
  - GALS is a good compromise

- Why chip multiprocessor
  - The challenge of increasing clock frequency
  - High performance and high energy efficiency of multiprocessor system
GALS Effects

- Performance penalty due to additional synchronization delay

A simple GALS system

\[
\text{Sync. delay} = \text{clk edge alignment (} t_e \text{)} + \text{sync. circuit (} t_s \text{)}
\]

- High energy efficiency due to independent clock/voltage scaling
Synchronous vs. GALS comparisons

<table>
<thead>
<tr>
<th>Uniprocessor</th>
<th>Synchronous</th>
<th>GALS (multiple clock domains)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Previous work</td>
<td>This work</td>
</tr>
</tbody>
</table>

- **Synchronous**: Uniprocessor
- **GALS**: Array processor

ALU, MULT, MEM
Outline

• Motivation
• **Experimental platform: a GALS array processor**
  • Performance analysis of GALS multiprocessor
  • Power analysis of GALS multiprocessor
Our Implementation of a GALS Array and Synchronous Array

• Contains multiple identical simple processors
• Local oscillator and dual-clock FIFOs are key components for GALS style
Micrograph of the 6 x 6 GALS Array

Technology: 0.18 µm CMOS
Max speed: 475 MHz
Area (1 Proc): 0.66 mm²
Fully functional

[Yu, ISSCC06]
Outline

• Motivation
• Experimental platform: a GALS array processor
• **Performance analysis of GALS multiprocessor**
• Power analysis of GALS multiprocessor
Performance Penalty of GALS Uni-processor

- Extra pipeline hazards result in ~10% throughput penalty compared with synchronous uni-processor.

Branch penalty of 3 cycles in a synchronous uni-processor:

Branch penalty of 3 cycles and 4 SYNC delays in a GALS uni-processor.

[Lyer, ISCA02; Semeraro, HPCA02; Talpes, ISLPED03]
Application Performance of GALS and Synchronous Array

- GALS array has only ~1% performance penalty
- Simulation conditions: 32-word FIFO, same clock frequency, 2 synchronization registers for GALS

<table>
<thead>
<tr>
<th></th>
<th>8-pt DCT</th>
<th>8x8 DCT</th>
<th>Zig-zag</th>
<th>merg sort</th>
<th>bub. sort</th>
<th>ma-trix</th>
<th>64 FFT</th>
<th>JPEG</th>
<th>802.11a/g</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sync. array</td>
<td>41</td>
<td>498</td>
<td>168</td>
<td>254</td>
<td>444</td>
<td>817.5</td>
<td>11439</td>
<td>1439</td>
<td>87857</td>
</tr>
<tr>
<td>GALS array</td>
<td>41</td>
<td>505</td>
<td>168</td>
<td>254</td>
<td>444</td>
<td>819</td>
<td>11710</td>
<td>1443</td>
<td>88989</td>
</tr>
<tr>
<td>GALS perf. reduction(%)</td>
<td>0</td>
<td>1.4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.1</td>
<td>2.3</td>
<td>0.3</td>
<td>1.3</td>
</tr>
</tbody>
</table>

Clock cycles for several applications
The Source of Performance Penalty of GALS Array Processor

• For all systems, communication delay affects system throughput only when it generates a loop.

• For GALS array processor, communication loop is the FIFO stall loop.
  – Performance simulation results show that the chance of FIFO stall loop is low for many DSP applications.

• FIFO depth affects FIFO stalls and hence a reasonable FIFO size is required.
Importance of the Communication Loop Delay

- One way communication does not affect system throughput
- Communication loop degrades throughput
  - In uni-processor, it is caused by pipeline hazards
  - GALS system has longer communication loop delay

One way communication:
The throughput is $1/\max(T1, T2, T3)$

Communication loop:
The throughput is $1/(T1 + T2 + T3 + T4)$
Examples of Stall and Stall Loops in a GALS Array Processor

- Data producer proc. 1 too slow causes frequent FIFO empty stalls.
- Data consumer proc. 2 too slow causes frequent FIFO full stalls.
- Data producer proc. 1 and data consumer proc. 2 both too slow at different times cause FIFO empty and full stalls.
- Example of multiple-link loop between two processors.
Performance of Synchronous and GALS Array with Different FIFO Sizes

Synchronous Array

GALS Array

GALS vs. Synchronous Array

Relative Performance

Relative Performance

Performance Ratio

8 DCT  zig-zag  bsort  64 FFT  802.11
8x8 DCT  msort  matrix  JPEG
Outline

- Motivation
- Experimental platform: a GALS array processor
- Performance analysis of GALS multiprocessor
- **Power analysis of GALS multiprocessor**
Clock/Voltage Scaling in GALS Uni-processor

• All GALS designs
  – Independently control clock frequencies to save power
  – Reduced clock frequency allows voltage reduction
• Around 25% energy savings with more than 10% performance penalty

[Lyer, ISCA02; Semeraro, HPCA02; Talpes, ISLPED03]
Clock/Voltage Scaling in GALS Array Processor

• Similar basic idea as uni-processor
  – Use low clock frequency for processors with light computation load
  – Benefit from unbalanced processor computation loads
• Both static and dynamic clock scaling methods
  – We study only static scaling here
• The optimal processor clock frequency is determined by its
  – Computational load
  – Position
• Can achieve power savings without performance reduction!
Unbalanced Processor Computation Loads in Nine Applications

- 8-pt DCT
- 8x8 DCT
- zig-zag
- merge-sort
- bubble-sort
- matrix
- 64 FFT
- JPEG
- 802.11a/g
Throughput Changes with Statically Configured Clock for 8x8 DCT

Comput load (clk cycles)

408 204 408 204

Relative Throughput

Relative clock frequency

Optimal clock scaling points

- Scale 1st processor
- Scale 2nd processor
- Scale 3rd processor
- Scale 4th processor
Relationship of Processors in an 8x8 DCT

- Proc. 2 and 4 have identical computational load
- Different position results in a different FIFO stall style, which causes different clock scaling behavior
Power of GALS Array with Static Clock/Voltage Scaling

- 40% power savings without a performance penalty
- From simulated clock frequency and referenced clk/voltage/pow relationship
Summary

- Compared to a synchronous array processor, the proposed GALS array processor has:
  - $< 1\%$ throughput reduction
  - $\sim 40\%$ energy savings
- These results compare well with reported GALS uni-processors:
  - $\sim 10\%$ throughput reduction
  - $\sim 25\%$ energy savings
- Source of throughput reduction in GALS system
  - Extra cost for communication loops
  - Extra cost for FIFO stall loops in GALS array processors
- Energy benefit of GALS clock/voltage scaling
  - Unbalanced processor computation loads
Acknowledgments

• Funding
  – Intel Corporation
  – UC Micro
  – NSF Grant No. 0430090
  – UCD Faculty Research Grant

• Special Thanks
  – E. Work, T. Mohsenin, other VCL processor co-designers, R. Krishnamurthy, M. Anders, S. Mathew