Scalable Streaming-Array of Simple Soft-Processors for Stencil Computations with Constant Memory Bandwidth

Kentaro Sano  Yoshiaki Hatsuda  Satoru Yamamoto

Graduate School of Information Sciences
Tohoku University, Japan

Outline

- Introduction
- Stencil computation and its streaming
- Architecture and design
- Performance model
- Evaluation
- Conclusions

9-FPGA prototype system
Introduction

General-purpose µ-processors used for high-performance computing systems

Increasing cores for higher performance, but low growing rate of off-chip I/O bandwidth

✓ Insufficient memory/network bandwidth for the entire arithmetic performance
✓ Only a fraction of the peak performance usually utilized in large-scale systems

Stencil computation

✓ Typical scientific computing kernel
✓ Small operational intensity: 0.5 Flop/Byte
✓ µ-processors assume higher op. intensity

AMD Opteron X4: 4.2 Flop/Byte

3.7 Flop/Byte is lost. (88% of the peak)
Poor performance!

Stencil Computation on CPUs/GPUs

 Highly-optimized implementation techniques achieve:

<table>
<thead>
<tr>
<th>Processor</th>
<th>Performance [GFlop/s]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 core of Xeon E5220 (quad-core)</td>
<td>Peak: 9.0, Sustained: 2.8 (31%)</td>
</tr>
<tr>
<td>8 cores of Xeon E5220 SMP node</td>
<td>Peak: 72.3, Sustained: 15.9 (22%)</td>
</tr>
<tr>
<td>1 x NVidia Tesla C1016</td>
<td>Peak: 78, Sustained: 51 (65%)</td>
</tr>
<tr>
<td>16 x NVidia Tesla C1016</td>
<td>Peak: 1248, Sustained: 530 (42%)</td>
</tr>
</tbody>
</table>

Inefficiency is caused by

✓ General-purpose structure, unsuitable for stencil computation
✓ Imbalanced bandwidth for the entire performance

Another promising way: scalable custom computing on FPGAs
This Research

Linear scalability for high-performance stencil computation by custom computing on multiple FPGAs

Scalable Streaming-Array (SSA)
- Domain-specific architecture
- Policy to overcome high design-cost (Custom & programmable soft-cores)

Major contributions
- Extensible architecture for scalable multi-FPGA system with constant memory-bandwidth
- Programmable design of soft-cores tailored for the app domain
- Performance model to predict the scalability
- 9-FPGA system demonstrating 260 GFlop/s with only 2 GB/s

Iterative Stencil Computation

2D iterative stencil computation
- Update all the grid-points with local stencil computation
- Repeat for successive time-steps
  - Ex: Linear eq. solvers (Jacobi method, etc.)
  - Difference-scheme computations (CFD, etc.)
Streaming with a Cyclic Buffer

Cyclic buffer

Pipelining Multiple Iterations

Pipelining execution with constant bandwidth!

### Scalable Streaming-Array Architecture

![Diagram of Scalable Streaming-Array Architecture](image)

- **Multiple time-step execution per memory-read.**

### SSA Design for Jacobi Computation

#### Requirements
- ✓ High density & utilization of FP units: simple processing elements (PEs)
- ✓ Flexibility in application domain: programmable PEs & sequencers

![Overview of SSA on a single FPGA](image)
Processing Element (PE)

8-stage pipeline
- Buffer Memory
- Constant Memory
- Multiplexor
- Floating-point multiply and accumulate unit (FMAC)

Sequencer of each PSM
- Control PEs in PSM
- Micro operations
- PSM = SIMD core

SSA Assembly Language

Two opecodes (control, comput.)

lset  loop-set with # of iterations
bnz   branch if loop-counter is not zero
mulp  multiply
accp  multiply and accumulate-positive
Performance Model of SSA

- Cycles of pipelined exec. on multiple FPGAs
- Speedup to single FPGA
  - Grid: $128^2$ to $2048^2$
  - Iteration: 5000 to 80000
- Linear speedup up to 100s FPGAs for medium size

Prototyping SSA for Evaluation

ALTERA Stratix III FPGA x 9

- IEEE754 single-precision FP
- Sustained 2GB/s for reading and writing DDR2 Memory
- PSM has 8 PEs.
- Freq = 133MHz
**Photograph of 9-FPGA System**

Terasic DE3-150 boards

**Benchmark Problems**

- **2D heat-propagation**
  - Laplace’s equation
  - Jacobi method
  - 78000 iterations
  - Grid sizes = 128 x 128, 512 x 512, 2048 x 2048

- **Same hardware, different programs**

**Boundary condition and results**

- Simple condition
- Complex condition
Results

- Linear scalability already for small problem
- Sustained 260GFlop/s for 2GB/s bandwidth
- Utilization: 87.5% (ideal for 4 mul. & 3 add)
- Power-efficiency of 1.36GFlop/sW competing with the top 3 of Green500 list

Conclusions

- Scalable Streaming-Array
  - High-performance stencil-computation
  - Extensible multi-FPGA system
  - Programmable simple soft-cores
- Performance model
- 9-FPGA prototype system
  - Linear scalability & 260GFlop/s for only 2GB/s
  - 1.36GFlops/W competing with Green500 list
- Future work
  - Larger and complex 3D problems (fluid dynamics)
  - SSA compiler
  - SSA synthesis tool