TOHOKU UNIVERSITY

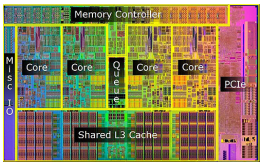
Scalable Streaming-Array of Simple Soft-Processors for Stencil Computations with Constant Memory-Bandwidth

Background

Problem: Low utilization of multi-core/many-core processors

- Many cores for high peak-performance ... but
- Biased mix of operations
- Insufficient bandwidth
- Overhead in parallel computing

low performance
low scalability



Intel Corei7 processor

nVidia Tesla c1060

It's difficult to fully exploit the fixed hardware for various algorithms...
How good with reconfigurable hardware'

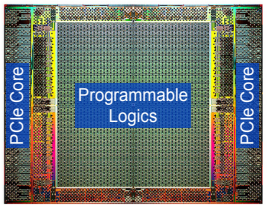
This research : low-power & high-performance accelerator

- Architecture & design of custom computing machines for stencil computation
- Reconfigurable accelerators with programmable-logic devices (FPGAs)
- Prototyping accelerators extensible with muliple FPGAs
- Demonstrating high scalability and high performance per power

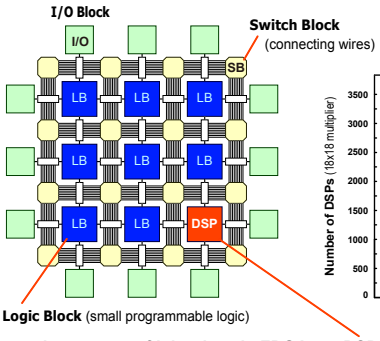
FPGA's Potential

Programmable-logic device

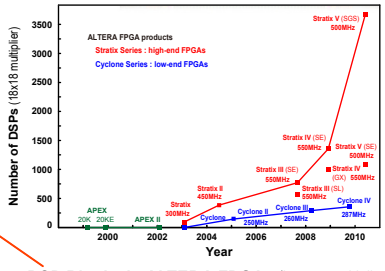
- Typical usage: VLSI emulation, Network processor for mobile-phone station
- Capable of high-performance computing



FPGA Die Photograph
(ALTERA Stratix IV GX)



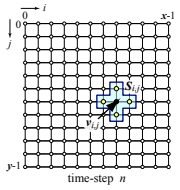
Internal structure of island-style FPGAs



DSP Blocks in ALTERA FPGAs (integer multipliers)
(more and faster DSPs in recent years)

Streaming Iterative Stencil-Computation

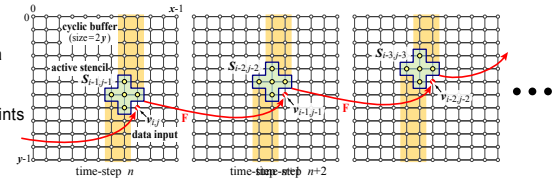
```
1: for(n=0; n<g ; n++) { // for g iterations
2:
4:   for(i=0; i<x; i++)
3:     for(j=0; j<y ; j++) { // for data-space
5:       // Update datum v(i,j) from n to n+1
5:       v(n+1,i,j) := F( v(n,i,j) in S(i,j) )
6:     }
7: }
```



Pseudo code of 2D iterative stencil computation

Iterative stencil-computation

- Update computing only with data inside "stencil" $S_{i,j}$
- Repeat updating all the grid-pooints
- More data accesses than computations

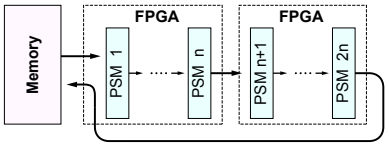


Streamed stencil-computation (multiple iterations for single sweep)

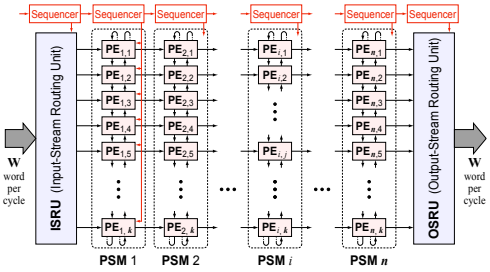
Scalable Streaming Array (SSA)

Design requirements

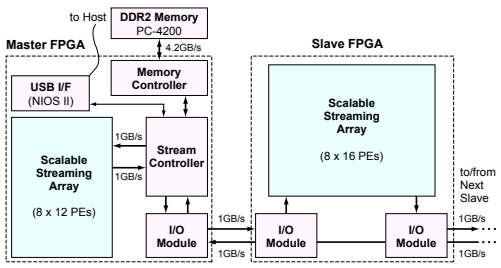
- Balancing memory bandwidth and arithmetic performance
- High scalability
- Computational flexibility
- High density & high utilization



Architecture of Scalable Streaming Array
(PSM: Pipeline-Stage Module)



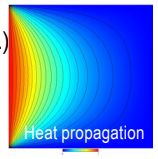
Structure of SSA on a single FPGA (k PEs x n PSMs)



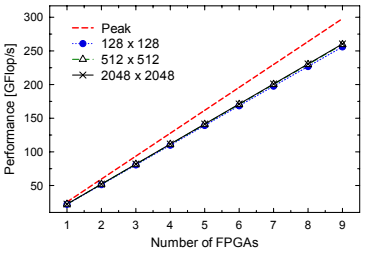
Prototyped system with Terasic DE3 boards (ALTERA Stratix III FPGAs)

Prototyping and Evaluation

2D Heat-propagation problem (Laplace's eq.)

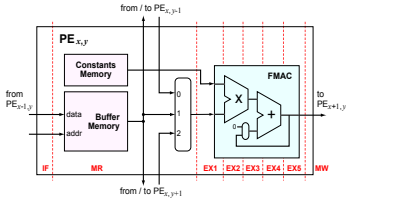


- Jacobi method (2048 x 2048 grid, 80000 iterations)
- Memory bandwidth of 2GB/s for R&W (DDR2 DRAM)
- Peak 300GFlop/s for 9 FPGAs
- Sustained 260GFlop/s
- FMAC utilization of 87.5%
- Higher than 1300MFlop/sW
- Almost linear speedup

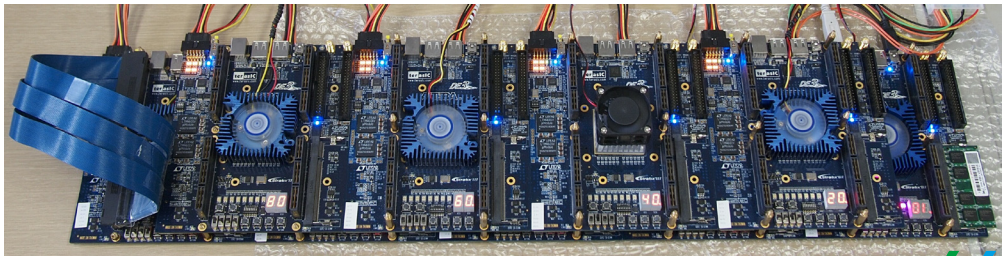


Future work

- Evaluation with practical applications
- High-speed host connection (PCI-Express)
- Optimal array generation and compiler



Structure of programmable processing elements (PEs)
(FMAC: Floating-Point Multiplier and Accumulator)



SSA prototyped on nine ALTERA StratixIII FPGAs

佐野研究室
Sano Laboratory
Graduate School of Information Sciences, Tohoku University, Japan
GSIS