Surviving the End of Scaling of Traditional Micro Processors in HPC

Olav Lindtjørn (Schlumberger, Stanford), Robert G. Clapp (Stanford), Oliver Pell, Oskar Mencer, Michael J Flynn (Maxeler)
The Memory Wall and the Power Wall

• Moore’s Law continues to deliver double the transistors on a chip every 18-24 months
  – But we are having trouble making those extra transistors deliver performance.

• Memory Wall
  – Parallel processing elements on-chip must share the same off-chip bandwidth

• Power Wall
  – Chips still need to be cooled in the same physical space
CPUs vs. FPGA Processing

Streaming Data through a data flow machine
Outline

• Oil and Gas HPC applications
• Maxeler FPGA Compiler and Accelerators
• Key Computational Kernels in Oil&Gas
  – Sparse Matrix
  – Convolution based methods
• Applications scalability – Technology trends
• Conclusions
HPC – Its role in Oil & Gas exploration

- Identify resources
- Access resources
- Maximize recovery

Courtesy of Statoil
Where to Drill

Seismic – Acoustic measurement
Electromagnetic
Gravity
Seismic Detectors
Seismic Source
Recording Vessel
Sea Surface
Water Bottom
Oil
Gas
Hydrocarbon "trap"
Seismic Wave
Ray Paths
Seismic Data Acquisition
Data Intensity and Complex Physics

Isotropic

Anisotropic
Data Rates and Computational needs

- 20 – 25,000 sensors
- 500 MB – 2 GB
- 50 – 200,000 shots
- 50 – 200 TB Data
- 1000s node
- 5 – 7 days

Relative computational cost:
- Isotropic
- VTI
- TTI

30 Hz RTM

Relative Disk Space
Data Rates and Computational needs

<table>
<thead>
<tr>
<th>20 – 25,000 sensors</th>
<th>50 – 200,000 shots</th>
<th>1000s node</th>
</tr>
</thead>
<tbody>
<tr>
<td>500 MB – 2 GB</td>
<td>50 – 200 TB Data</td>
<td>5 – 7 days</td>
</tr>
</tbody>
</table>

Relative computational cost

- Isotropic
- VTI
- TTI

60 Hz RTM

30 Hz RTM

Relative Disk Space
Data Rates and Computational Needs

<table>
<thead>
<tr>
<th>Data Rates</th>
<th>Computational Needs</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 – 25,000 sensors</td>
<td>500 MB – 2 GB</td>
</tr>
<tr>
<td>50 – 200,000 shots</td>
<td>50 – 200 TB Data</td>
</tr>
<tr>
<td>1000s node 5 – 7 days</td>
<td></td>
</tr>
<tr>
<td>15 - 20,000 nodes Days - weeks</td>
<td></td>
</tr>
</tbody>
</table>

Relative computational cost

- Isotropic
- VTI
- TTI

Relative Disk Space

30 Hz RTM

60 Hz RTM
Cost of Imaging Algorithms

- FWI - Elastic
- FWI - Acoustic
- Reverse Time Migration (RTM)
- Shot WEM (VTI)

○: Memory and disk space cost

Relative computational cost vs Imaging Complexity

- ‘09
- ‘10
- ‘11
- ‘?

FWI-Elastic

Shot WEM (VTI)

FWI-Acoustic

Reverse Time Migration (RTM)
HPC – Its role in Hydrocarbon exploration

- Identify resources
- Access resources
HPC – Its role in Hydrocarbon exploration

- Identify resources
- Access resources
- Maximize recovery
- Geomechanics
- Reservoir Flow Simulation
- Geomechanics
Oil and Gas Computational Kernels

Physics
- Wave propagation
- Diffusion
- Fluid Flow

Kernels
- Finite Difference
- FFT
- Finite Element
- Sparse matrix
Oil and Gas Computational Kernels

Physics

- Wave propagation
- Diffusion
- Fluid Flow

Kernels

- Convolution
- FFT
- Sparse Matrix
- Sparse matrix
Oil and Gas Computational Kernels

Physics

- Wave propagation
- Diffusion
- Fluid Flow

Kernels

- Convolution
- FFT
- Sparse Matrix
Outline

• Oil and Gas HPC applications
• Maxeler FPGA Compiler and Accelerators
• Key Computational Kernels in Geophysics
  – Sparse Matrix
  – Convolution based methods
• Applications scalability – Technology trends
• Conclusions
Accelerating Convolution and Sparse Matrix in the Maxeler Environment
Maxeler Accelerators

- Commodity silicon chips configurable to implement any digital circuit
  - $\sim 10^6$ small processing elements that operate in parallel
  - Several megabytes of on-chip memory
  - Run at several hundred megahertz
  - Support large on-board memory (24GB+)
MaxNode with MAX3

Specifications:

<table>
<thead>
<tr>
<th></th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compute</td>
<td>8x 2.8GHz Nahelem Cores</td>
</tr>
<tr>
<td></td>
<td>4x Virtex 6-SX475T FPGAs</td>
</tr>
<tr>
<td>Interconnect</td>
<td>PCI-Express Gen. 2</td>
</tr>
<tr>
<td></td>
<td>MaxRing</td>
</tr>
<tr>
<td>Storage</td>
<td>3x 2TB Hard disks</td>
</tr>
<tr>
<td>Memory</td>
<td>96GB DRAM</td>
</tr>
<tr>
<td>Form Factor</td>
<td>1U</td>
</tr>
</tbody>
</table>

MAX3 Node Architecture

- FPGA
- PCIe
- MaxRing
- Nahelem Cores

Image of MAX3 Node with specifications and architecture diagram.
MAX3 System Bandwidths

- SX475T FPGA (4.68MB): 38.4 GBytes/s
- Mem 24GB: 8 GBytes/s
- PCIe x8 Gen 2: 8 GBytes/s
- 8x Nehalem Cores: 8 GBytes/s
- MaxRing: 8 GBytes/s

Memory 24GB connections:
- SX475T FPGA (4.68MB) to Mem 24GB: 38.4 GBytes/s
- Mem 24GB to SX475T FPGA (4.68MB): 38.4 GBytes/s
- PCIe x8 Gen 2 to Mem 24GB: 8 GBytes/s
public class MovingAverageKernel extends Kernel {

    public MovingAverageKernel(KernelParameters parameters, int N) {
        super(parameters);

        // Input
        HWVar x = io.input("x", hwFloat(8, 24));

        // Data
        HWVar prev = stream.offset(x, -1);
        HWVar next = stream.offset(x, 1);
        HWVar sum = prev + x + next;
        HWVar result = sum / 3;

        // Output
        io.output("y", result, hwFloat(8, 24));
    }
}
Sparse Matrix Format

(DIA) Diagonal
(ELL) ELLPACK
(CSR) Compressed Row
(HYB) Hybrid
(COO) Coordinate

Structured

Unstructured
Typical scalability of SLB Sparse Matrix Applications

Eclipse Benchmark
(2 node Westmere 3.06 GHz)

Visage – Geomechanics
(2 node Nehalem 2.93 GHz)

E300 2 Mcell Benchmark

FEM Benchmark

Relative Speed

# cores

0 1 2 3 4 5 6 7 8

# cores

0 1 2 3 4 5 6 7 8

Relative Speed

0 1 2 3 4 5

0 2 4 6 8 10 12

0 2 4 6
Sparse Matrix on FPGAs

- 4 MB BLK RAM
- Pipelining
- Addressing scheme optimized for Matrix structure
- Domain Specific Data Encoding
Sparse Matrix on FPGAs

SPEEDUP is 20x-40x per 1U at 200MHz
Sparse Matrix on FPGAs

SPEEDUP is 20x-40x per 1U at 200MHz

Domain Specific Address and Data Encoding
3D Convolution

- Low Flop/Byte ratio
- Sparse structure requires large streaming memory buffers \((14 \times nx \times ny\) for 14\(^{th}\) order in space).
- Data Structure >> Data Caches

- CPUs:
  - Constrained by:
    - Small L1/L2 cache
    - Limited utilization of pipeline
    - Limited by Streaming BW
    - Limited data element reuse
    - Fraction of peak performance
FPGA Opportunities

- FPGA opportunities
- 4 MB on-chip Memory
- Hundreds of pipeline stages
- Optimal trade off between streams for BW utilization and Pipe line depth

CPU limits:
- Constrained by:
  - Small L1/L2 cache
  - Limited depth of pipeline
  - Limited by Streaming BW
  - Limited data element reuse
  - → Fraction of peak performance
### Performance

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Hardware</th>
<th>Design</th>
<th>Speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>Star stencil</td>
<td>VIRTEX 5</td>
<td>3 pipe</td>
<td>20x</td>
</tr>
<tr>
<td>Star stencil</td>
<td>VIRTEX 6</td>
<td>9 pipe</td>
<td>73x</td>
</tr>
</tbody>
</table>

8-core Nehalem 2.93 GHz 1U server vs 1U MaxNode
Outline

• Oil and Gas HPC applications
• Maxeler FPGA Compiler and Accelerators
• Key Computational Kernels in Geophysics
  – Sparse Matrix
  – Convolution based methods
• Applications scalability – Technology trends
• Conclusions
Application scalability and Technology trends

- Transistor count keeps increasing
- Memory BW continues to trail
- How will our algorithms scale?

- Convolution:
  - Deeper pipelines:
    - An example: Cascading multiple time steps
  - Specialized macros on FPGAs
FPGA: Time step Cascading

Stencil width
FPGA: Time step Cascading
FPGA: Time step Cascading

Advance t1

Stencil width
FPGA: Time step Cascading

All information needed to update t2 at X is now available
FPGA: Time step Cascading

All information needed to update $t_2$ at $X$ is now available.
FPGA: Time step Cascading

In one pass through the data to multiple steps in time
FPGA: Time step Cascading

- Requires more computational units per pass but
- reduce memory bandwidth requirements

Stencil width
Technology opportunities

- Added Resources (Transistor scaling) translates directly into performance using Multiple time step techniques
- Independent of Memory BW increase

Resource costs for a symmetric 15-point stencil:

<table>
<thead>
<tr>
<th></th>
<th>LUT/FFs</th>
<th>DSPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>MaxGenFD on Virtex-5</td>
<td>207</td>
<td>8</td>
</tr>
<tr>
<td>MaxGenFD on Virtex-6</td>
<td>33</td>
<td>8</td>
</tr>
<tr>
<td>Resulting perf. improvement</td>
<td>50 %</td>
<td></td>
</tr>
</tbody>
</table>

Virtex-6 DSP enhanced with Pre-Adder
Outline

• Oil and Gas HPC applications
• Maxeler FPGA Compiler and Accelerators
• Key Computational Kernels in Geophysics
  – Sparse Matrix
  – Convolution based methods
• Applications scalability – Technology trends
• Conclusions
Surviving the End of Scaling of Traditional Micro Processors in HPC

- Conclusions:
  - FPGA Streaming has come of age
  - Development Environment is here today
  - Application will scale with predicted technology evolution
  - Considerable upside for “smart macros”

Conventional Road Map
Surviving the End of Scaling of Traditional Micro Processors in HPC

Conclusions:
- FPGA Streaming has come of age
- Development Environment is here today
- Application will scale with predicted technology evolution
- Considerable upside for “smart macros”
Thank You
GPU Comments