Lecture 14 – Parallelism in Software V

Mattan Erez

The University of Texas at Austin
Quick recap

• Decomposition
  – High-level and fairly abstract
  – Consider machine scale for the most part
  – Task, Data, Pipeline
  – Find dependencies

• Algorithm structure
  – Still abstract, but a bit less so
  – Consider communication, sync, and bookkeeping
  – Task (collection/recursive)
  – Data (geometric/recursive)
  – Dataflow (pipeline/event-based-coordination)

• Supporting structures
  – Loop
  – Master/worker
  – Fork/join
  – SPMD
  – MapReduce
## Algorithm Structure and Organization (my view)

<table>
<thead>
<tr>
<th></th>
<th>Task parallelism</th>
<th>Divide and conquer</th>
<th>Geometric decomposition</th>
<th>Recursive data</th>
<th>Pipeline</th>
<th>Event-based coordination</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SPMD</strong></td>
<td>*****</td>
<td>**</td>
<td>*****</td>
<td>**</td>
<td>*****</td>
<td>*</td>
</tr>
<tr>
<td><strong>Loop Parallelism</strong></td>
<td>*****</td>
<td>*</td>
<td>****</td>
<td>*</td>
<td>*****</td>
<td>** SWP to hide comm.</td>
</tr>
<tr>
<td>when no dependencies</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Master/Worker</strong></td>
<td>*****</td>
<td>***</td>
<td>***</td>
<td>***</td>
<td>**</td>
<td>****</td>
</tr>
<tr>
<td><strong>Fork/Join</strong></td>
<td>*****</td>
<td>*****</td>
<td>**</td>
<td>****</td>
<td></td>
<td>*</td>
</tr>
</tbody>
</table>

- Patterns can be hierarchically composed so that a program uses more than one pattern
Patterns for Parallelizing Programs

4 Design Spaces

Algorithm Expression
- Finding Concurrency
  - Expose concurrent tasks

- Algorithm Structure
  - Map tasks to processes to exploit parallel architecture

Software Construction
- Supporting Structures
  - Code and data structuring patterns

- Implementation Mechanisms
  - Low level mechanisms used to write parallel programs

ILP, DLP, and TLP in SW and HW

- **ILP**
  - OOO
  - Dataflow
  - VLIW

- **DLP**
  - SIMD
  - Vector

- **TLP**
  - Essentially multiple cores with multiple sequencers

- **ILP**
  - Within straight-line code

- **DLP**
  - Parallel loops
  - Tasks operating on disjoint data
  - No dependencies within parallelism phase

- **TLP**
  - All of DLP +
  - Producer-consumer chains
ILP, DLP, and TLP and Supporting Patterns

<table>
<thead>
<tr>
<th>Task parallelism</th>
<th>Divide and conquer</th>
<th>Geometric decomposition</th>
<th>Recursive data</th>
<th>Pipeline</th>
<th>Event-based coordination</th>
</tr>
</thead>
<tbody>
<tr>
<td>ILP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DLP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TLP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# ILP, DLP, and TLP and Supporting Patterns

<table>
<thead>
<tr>
<th></th>
<th>Task parallelism</th>
<th>Divide and conquer</th>
<th>Geometric decomposition</th>
<th>Recursive data</th>
<th>Pipeline</th>
<th>Event-based coordination</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ILP</strong></td>
<td>inline / unroll</td>
<td>inline</td>
<td>unroll</td>
<td>inline</td>
<td>inline / unroll</td>
<td>inline</td>
</tr>
<tr>
<td><strong>DLP</strong></td>
<td>natural or local-conditions</td>
<td>after enough divisions</td>
<td>natural</td>
<td>after enough branches</td>
<td>difficult</td>
<td>local-conditions</td>
</tr>
<tr>
<td><strong>TLP</strong></td>
<td>natural</td>
<td>natural</td>
<td>natural</td>
<td>natural</td>
<td>natural</td>
<td>natural</td>
</tr>
</tbody>
</table>
## ILP, DLP, and TLP and Implementation Patterns

<table>
<thead>
<tr>
<th></th>
<th>SPMD</th>
<th>Loop Parallelism</th>
<th>Mater/Worker</th>
<th>Fork/Join</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ILP</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>DLP</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>TLP</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **SPMD**
  - SPMD
  - Loop Parallelism
  - Mater/Worker
  - Fork/Join

- **ILP**
  - ILP
  - Pipeline
  - Unroll
  - Inline

- **DLP**
  - DLP
  - Natural or local
  - Conditional

- **TLP**
  - TLP
  - Natural
  - Conditional

After enough divisions, local

---

8
### ILP, DLP, and TLP and Implementation Patterns

<table>
<thead>
<tr>
<th>SPMD</th>
<th>Loop Parallelism</th>
<th>Master/Worker</th>
<th>Fork/Join</th>
</tr>
</thead>
<tbody>
<tr>
<td>ILP</td>
<td>pipeline</td>
<td>unroll</td>
<td>inline</td>
</tr>
<tr>
<td>DLP</td>
<td>natural or local-conditional</td>
<td>natural</td>
<td>local-conditional</td>
</tr>
<tr>
<td>TLP</td>
<td>natural</td>
<td>natural</td>
<td>natural</td>
</tr>
</tbody>
</table>
Outline

• Molecular dynamics example
  – Problem description
  – Steps to solution
    • Build data structures; Compute forces; Integrate for new positions; Check global solution; Repeat
  – Finding concurrency
    • Scans; data decomposition; reductions
  – Algorithm structure
  – Supporting structures
Credits

• Parallel Scan slides courtesy David Kirk (NVIDIA) and Wen-Mei Hwu (UIUC)
  – Taken from EE493-AI taught at UIUC in Spring 2007

• Reduction slides courtesy Dr. Rodric Rabbah (IBM)
  – Taken from 6.189 IAP taught at MIT in 2007
GROMACS

• Highly optimized molecular-dynamics package
  – Popular code
  – Specifically tuned for protein folding
  – Hand optimized loops for SSE3 (and other extensions)
Gromacs Components

- Non-bonded forces
  - Water-water with cutoff
  - Protein-protein tabulated
  - Water-water tabulated
  - Protein-water tabulated

- Bonded forces
  - Angles
  - Dihedrals

- Boundary conditions

- Verlet integrator

- Constraints
  - SHAKE
  - SETTLE

- Other
  - Temperature-pressure coupling
  - Virial calculation
GROMACS Water-Water Force Calculation

- Non-bonded long-range interactions
  - Coulomb
  - Lennard-Jones
  - 234 operations per interaction

\[ V_{nb} = \sum_{i,j} \left[ \frac{1}{4\pi\varepsilon_0} \frac{q_i q_j}{r_{ij}} + \left( \frac{C_{12}}{r_{ij}^{12}} - \frac{C_6}{r_{ij}^6} \right) \right] \]

Water-water interaction \(~75\%\) of GROMACS run-time
GROMACS Uses Non-Trivial Neighbor-List Algorithm

- Full non-bonded force calculation is $o(n^2)$
- GROMACS approximates with a cutoff
  - Molecules located more than $r_c$ apart do not interact
  - $O(nr_c^3)$

Efficient algorithm leads to variable rate input streams
GROMACS Uses Non-Trivial Neighbor-List Algorithm

- Full non-bonded force calculation is \( o(n^2) \)
- GROMACS approximates with a cutoff
  - Molecules located more than \( r_c \) apart do not interact
  - \( O(nr_c^3) \)

Efficient algorithm leads to variable rate input streams
GROMACS Uses Non-Trivial Neighbor-List Algorithm

- Full non-bonded force calculation is $o(n^2)$
- GROMACS approximates with a cutoff
  - Molecules located more than $r_c$ apart do not interact
  - $O(nr_c^3)$

Efficient algorithm leads to variable rate input streams
GROMACS Uses Non-Trivial Neighbor-List Algorithm

- Full non-bonded force calculation is $o(n^2)$
- GROMACS approximates with a cutoff
  - Molecules located more than $r_c$ apart do not interact
  - $O(nr_c^3)$

Efficient algorithm leads to variable rate input streams
GROMACS Uses Non-Trivial Neighbor-List Algorithm

• Full non-bonded force calculation is $o(n^2)$

• GROMACS approximates with a cutoff
  
  – Molecules located more than $r_c$ apart do not interact
  
  – $O(nr_c^3)$

• Separate neighbor-list for each molecule
  
  – Neighbor-lists have variable number of elements

Efficient algorithm leads to variable rate input streams
Parallel Prefix Sum (Scan)

• **Definition:**
  The all-prefix-sums operation takes a binary associative operator $\oplus$ with identity $I$, and an array of $n$ elements

  $$[a_0, a_1, \ldots, a_{n-1}]$$

  and returns the ordered set

  $$[I, a_0, (a_0 \oplus a_1), \ldots, (a_0 \oplus a_1 \oplus \ldots \oplus a_{n-2})].$$

• **Example:**
  if $\oplus$ is addition, then scan on the set

  $$[3 \ 1 \ 7 \ 0 \ 4 \ 1 \ 6 \ 3]$$

  returns the set

  $$[0 \ 3 \ 4 \ 11 \ 11 \ 15 \ 16 \ 22]$$

  (From Blelloch, 1990, “Prefix Sums and Their Applications”)
Applications of Scan

• Scan is a simple and useful parallel building block
  - Convert recurrences from sequential:
    for (j=1; j<n; j++)
    out[j] = out[j-1] + f(j);
  - into parallel:
    forall (j) { temp[j] = f(j) };
    scan(out, temp);

• Useful for many parallel algorithms:
  - radix sort
  - quicksort
  - String comparison
  - Lexical analysis
  - Stream compaction
  - Polynomial evaluation
  - Solving recurrences
  - Tree operations
  - Building data structures
  - Etc.
Building Data Structures with Scans

• Fun on the board
Scan on a serial CPU

```c
void scan( float* scanned, float* input, int length)
{
    scanned[0] = 0;
    for(int i = 1; i < length; ++i)
    {
        scanned[i] = input[i-1] + scanned[i-1];
    }
}
```

- Just add each element to the sum of the elements before it
- Trivial, but sequential
- Exactly $n$ adds: optimal
A First-Attempt Parallel Scan Algorithm

1. Read input to shared memory. Set first element to zero and shift others right by one.

Each UE reads one value from the input array in device memory into shared memory array T0. UE 0 writes 0 into shared memory array.
A First-Attempt Parallel Scan Algorithm

1. (previous slide)

2. Iterate \( \log(n) \) times: UEs \textit{stride} to \( n \):
   Add pairs of elements \textit{stride} elements apart.
   Double \textit{stride} at each iteration. (note must double buffer shared mem arrays)

---

Iteration #1
Stride = 1

- Active UEs: \textit{stride} to \( n-1 \) (\( n \)-\textit{stride} UEs)
- UE \( j \) adds elements \( j \) and \( j \)-\textit{stride} from T0 and writes result into shared memory buffer T1 (ping-pong)
A First-Attempt Parallel Scan Algorithm

<table>
<thead>
<tr>
<th></th>
<th>In</th>
<th>3</th>
<th>1</th>
<th>7</th>
<th>0</th>
<th>4</th>
<th>1</th>
<th>6</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>T0</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>7</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>T1</td>
<td>0</td>
<td>3</td>
<td>4</td>
<td>8</td>
<td>7</td>
<td>4</td>
<td>5</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>T0</td>
<td>0</td>
<td>3</td>
<td>4</td>
<td>11</td>
<td>11</td>
<td>12</td>
<td>12</td>
<td>11</td>
<td></td>
</tr>
</tbody>
</table>

1. Read input from device memory to shared memory. Set first element to zero and shift others right by one.

2. Iterate \(\log(n)\) times: UEs **stride** to \(n\):
Add pairs of elements **stride** elements apart. Double **stride** at each iteration. (note must double buffer shared mem arrays)

© David Kirk/NVIDIA and Wen-mei W. Hwu, 2007
ECE 498AL, University of Illinois, Urbana-Champaign
A First-Attempt Parallel Scan Algorithm

1. Read input from device memory to shared memory. Set first element to zero and shift others right by one.

2. Iterate log(n) times: UEs stride to n: Add pairs of elements stride elements apart. Double stride at each iteration. (note must double buffer shared mem arrays)

© David Kirk/NVIDIA and Wen-mei W. Hwu, 2007
ECE 498AL, University of Illinois, Urbana-Champaign
A First-Attempt Parallel Scan Algorithm

1. Read input from device memory to shared memory. Set first element to zero and shift others right by one.

2. Iterate \( \log(n) \) times: UEs stride to \( n \): Add pairs of elements stride elements apart. Double stride at each iteration. (note must double buffer shared mem arrays)

3. Write output.
What is wrong with our first-attempt parallel scan?

• **Work Efficient:**
  - A parallel algorithm is work efficient if it does the same amount of work as an optimal sequential complexity

• **Scan executes log(n) parallel iterations**
  - The steps do \( n-1, n-2, n-4, \ldots, n/2 \) adds each
  - Total adds: \( n \times (\log(n) - 1) + 1 \rightarrow O(n \times \log(n)) \) work

• **This scan algorithm is NOT work efficient**
  - Sequential scan algorithm does \( n \) adds
  - A factor of \( \log(n) \) hurts: 20x for \( 10^6 \) elements!
Improving Efficiency

• A common parallel algorithm pattern: Balanced Trees
  – Build a balanced binary tree on the input data and sweep it to and from the root
  – Tree is not an actual data structure, but a concept to determine what each UE does at each step

• For scan:
  – Traverse down from leaves to root building partial sums at internal nodes in the tree
    • Root holds sum of all leaves
  – Traverse back up the tree building the scan from the partial sums
Build the Sum Tree

Assume array is already in shared memory
Build the Sum Tree

Stride 1

Iteration 1, \( n/2 \) UEs

Iterate \( \log(n) \) times. Each UE adds value \( stride \) elements away to its own value

Each \( \circ \) corresponds to a single UE.
Iterate \(\log(n)\) times. Each UE adds value \(\text{stride}\) elements away to its own value.
Iterate log(n) times. Each UE adds value *stride* elements away to its own value.

Note that this algorithm operates in-place: no need for double buffering.
Zero the Last Element

We now have an array of partial sums. Since this is an exclusive scan, set the last element to zero. It will propagate back to the first element.
## Build Scan From Partial Sums

| 1 | 3 | 4 | 7 | 11 | 4 | 5 | 6 | 0 |

---

© David Kirk/NVIDIA and Wen-mei W. Hwu, 2007

eCE 498AL, University of Illinois, Urbana-Champaign

and Locality, Fall 2011 -- Lecture 14 (c) Rodric Rabbah, Mattan Erez
Iterate \( \log(n) \) times. Each UE adds value \( \text{stride} \) elements away to its own value, and sets the value \( \text{stride} \) elements away to its own \text{previous} \ value.
Build Scan From Partial Sums

Iterate log(n) times. Each UE adds value \textit{stride} elements away to its own value, and sets the value \textit{stride} elements away to its own previous value.
## Build Scan From Partial Sums

<table>
<thead>
<tr>
<th>T</th>
<th>3</th>
<th>4</th>
<th>7</th>
<th>11</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>0</th>
</tr>
</thead>
</table>

### Stride 4
- Total steps: \(2 \times \log(n)\).
- Total work: \(2 \times (n-1)\) adds = \(O(n)\)  
  **Work Efficient!**

### Stride 2

### Stride 1

Done! We now have a completed scan that we can write out to device memory.

© David Kirk/NVIDIA and Wen-mei W. Hwu, 2007  
ECE 498AL, University of Illinois, Urbana-Champaign
Reductions

• Many to one
• Many to many
  – Simply multiple reductions
    • Also known as scatter-add and subset of parallel prefix sums
• Use
  – Histograms
  – Superposition
    • Physical properties
Serial Reduction

- When reduction operator is not associative
- Usually followed by a broadcast of result
Tree-based Reduction

- \( n \) steps for \( 2^n \) units of execution
- When reduction operator is associative
- Especially attractive when only one task needs result
Recursive-doubling Reduction

- \( n \) steps for \( 2^n \) units of execution
- If all units of execution need the result of the reduction
Recursive-doubling Reduction

• Better than tree-based approach with broadcast
  – Each units of execution has a copy of the reduced value at the end of n steps
  – In tree-based approach with broadcast
    • Reduction takes $n$ steps
    • Broadcast cannot begin until reduction is complete
    • Broadcast can take $n$ steps (architecture dependent)
Other Examples

• More patterns
  – Reductions
  – Scans
    • Building a data structure

• More examples
  – Search
  – Sort
  – FFT as divide and conquer
  – Structured meshes and grids
  – Sparse algebra
  – Unstructured meshes and graphs
  – Trees
  – Collections
    • Particles
    • Rays