Advanced Algorithmics (6EAP)
Parallel algorithms

Jaak Vilo
2014 Fall
• 67 + 56 = ?
\text{SUM}( 10 \ 7 \ 98 \ 82 \ 84 \ 66 \ 62 \ 94 \ 6 \ 25 \ 73 \ 62 \ 65 \ 5 \ 36 \ 983 \ 73 \ 93 \ 77 \ 10 \ 7 \ 98 \ 82 \ 84 \ 66 \ 62 \ 94 \ 6 \ 25 \ 73 \ 62 \ 65 \ 5 \ 36 \ 983 \ 73 \ 93 \ 77 \ 10 \ 7 \ 98 \ 82 \ 84 \ 66 \ 62 \ 94 \ 6 \ 25 \ 73 \ 62 \ 65 \ 5 \ 36 \ 983 \ 73 \ 93 \ 77 \ 10 \ 7 \ 98 \ 82 \ 84 \ 66 \ 62 \ 94 \ 6 \ 25 \ 73 \ 62 \ 65 \ 5 \ 36 \ 983 \ 73 \ 93 \ 77 \ 10 \ 7 \ 98 \ 82 \ 84 \ 66 \ 62 \ 94 \ 6 \ 25 \ 73 \ 62 \ 65 \ 5 \ 36 \ 983 \ 73 \ 93 \ 77 \ 10 \ 7 \ 98 \ 82 \ 84 \ 66 \ 62 \ 94 \ 6 \ 25 \ 73 \ 62 \ 65 \ 5 \ 36 \ 983 \ 73 \ 93 \ 77 \ ) = ?
Researchers Give Update on Road to Parallelism

EE Times (03/19/10) Merritt, Rick

University of Illinois researchers have taken several small steps toward developing new parallel programming models to tap the many-core processors of the future. The DeNovo project attempts to define a new and more rigorous way of utilizing shared memory. It is working concurrently with a separate effort to define a deterministic, parallel language primarily based on a parallel version of Java and eventually migrating to a parallel version of C++. The chip project that is nearest to testing is the 1,024-core Rigel processor architecture targeting high density, high throughput computing, which would be programmed through a task-level applications programming interface aimed at chores in imaging, computer vision, physics, and simulations. The Bulk Architecture chip design is testing the notion of atomic transactions.

View Full Article  |  Return to Headlines
NASA, Air Force define cutting-edge next-generation space computer

The processor will be applicable to a broad variety of military and civil space missions

By Michael Cooney | Network World US | Published: 21:19, 11 April 2013

It's not every day a customer wants a computer that can do everything from direct a landing on the surface of Mars, to control descent onto a speeding asteroid or coordinate a flight of satellites - but NASA and the US Air Force's space group aren't everyday customers.

NASA and the Air Force Research Laboratory's Space Vehicles Directorate put out a call today for research and development of what likely will become the next computer system - what the agencies call the Next Generation Space Processor (NGSP) - to fly onboard a variety of future spacecraft.

The processor is envisioned to be a radiation-hardened, general-purpose multicore chip and associated software, applicable to a broad variety of military and civil space missions, and a broad range of spacecraft sizes and power/mass/volume constraints, NASA stated.
Titan has AMD Opteron CPUs in conjunction with Nvidia Tesla GPUs to improve energy efficiency while providing an order of magnitude increase in computational power over Jaguar. It uses 18,688 CPUs paired with an equal number of GPUs to perform at a theoretical peak of 27 petaFLOPS; however, in the LINPACK benchmark used to rank supercomputers' speed, it performed at 17.59 petaFLOPS. This was enough to take first place in the November 2012 list by the TOP500 organisation.
Tianhe-2 (MilkyWay-2) - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P

| Site: | National Super Computer Center in Guangzhou |
| Manufacturer: | NUDT |
| Cores: | 3,120,000 |
| Linpack Performance (Rmax) | 33,862.7 TFlop/s |
| Theoretical Peak (Rpeak) | 54,902.4 TFlop/s |
| Power: | 17,808.00 kW |
| Memory: | 1,024,000 GB |
| Interconnect: | TH Express-2 |
| Operating System: | Kylin Linux |
| Compiler: | icc |
| Math Library: | Intel MKL-11.0.0 |
| MKL: | MKL with a customized GSY channel |

TOP10 November 2013

1. Tianhe-2 (MilkyWay-2) - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P
   NUDT

2. Titan - Cray XK7, Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x
   Cray Inc.
Computing

One can calculate FLOPS using this equation:[1]

\[ \text{FLOPS} = \text{chassis} \times \frac{\text{nodes}}{\text{chassis}} \times \frac{\text{sockets}}{\text{node}} \times \frac{\text{cores}}{\text{socket}} \times \text{clock} \times \frac{\text{FLOPs}}{\text{cycle}} \]

or

\[ \text{FLOPS} = \text{cores} \times \text{clock} \times \frac{\text{FLOPs}}{\text{cycle}} \]

Most microprocessors today can do 4 FLOPs per clock cycle. Therefore, a single-core 2.5 GHz processor has a theoretical performance of 10 billion FLOPS = 10 GFLOPS.

Note: In this context, sockets is referring to chip sockets on a motherboard, in other words, how many computer chips are in use, with each chip having one or more cores on it. This equation only applies to one very specific (but common) hardware architecture and it ignores limits imposed by memory bandwidth and other constraints. In general, GigaFLOPS are not determined by theoretical calculations such as this one; instead, they are measured by actual benchmarks of actual performance/throughput. Because this equation ignores all sources of overhead, in the real world, you will never get actual performance that is anywhere near to what this equation predicts.
Instruction-level parallelism

A canonical five-stage pipeline in a RISC machine (IF = Instruction Fetch, ID = Instruction Decode, EX = Execute, MEM = Memory access, WB = Register write back)
A five-stage pipelined **superscalar processor**, capable of issuing two instructions per cycle. It can have two instructions in each stage of the pipeline, for a total of up to 10 instructions (shown in green) being simultaneously executed.

Instructions can be grouped together only if there is no **data dependency** between them.
TÜ HPC – v1, Feb 2, 2009

- 42 nodes, 2x4-core = 336 core
- 32GB RAM / node
- Infiniband fast interconnect

- [http://www.hpc.ut.ee/](http://www.hpc.ut.ee/)

- Job scheduling (Torque)
- MPI
Aurumasina InfiniBand võrgu skeem
Latency
The single data rate switch chips have a latency of 200 nanoseconds, and DDR switch chips have a latency of 140 nanoseconds. The end-to-end latency range is from 1.07 microseconds MPI latency to 1.29 microseconds MPI latency to 2.6 microseconds.

Compare:
The speed of a normal ethernet ping/pong (request and response) is roughly 350us (microseconds) or about .35 milliseconds or .00035 seconds.
Grace Hopper - Nanosecods

http://www.youtube.com/watch?v=JEpsKnWZrJ8
Single large-memory machines
@TartuUniversity

- 80-core, 256GB RAM
- HPC: 512GB,
- EXCS: 160-core cpu, 1TB RAM fall 2011
- >3000 cores now
- 2TB RAM machine

- [http://hpc.ut.ee/](http://hpc.ut.ee/) for more info!
Drivers for parallel computing

• Multi-core processors (2, 4, ... , 64 ... )
• Computer clusters (and NUMA)
• Specialised computers (e.g. vector processors, massively parallel, ...)
• Distributed computing: GRID, cloud, ...
• Need to create computer systems from cheaper (and weaker) components, but many ...

• One of the major challenges of modern IT
Slides based on materials from

- Joe Davey
- Matt Maul
- Ashok Srinivasan
- Edward Chrzanowski
- Ananth Grama, Anshul Gupta, George Karypis, and Vipin Kumar
- Dr. Amitava Datta and Prof. Dr. Thomas Ottmann
- ... and many others
books


• CLRS: (3rd edition) pp 772-812 (Chapter 27)
What is a Parallel Algorithm?

• Imagine you needed to find a lost child in the woods.

• Even in a small area, searching by yourself would be very time consuming.

• Now if you gathered some friends and family to help you, you could cover the woods in much faster manner...
A simple parallel algorithm

Adding n numbers in parallel
Terms (CLRS)

- work, span
Parallel complexity

• Schedule on p processors
• schedule depth – max /longest path in schedule/
• N – set of nodes on DAG
• time ti – assigned for each DAG node

• \( T_p(n) = \min \{ \max_{i \in N} t_i \} \)

• Length of the longest (critical) path!
Boolean circuits, bit-level parallelism

• (Micro)processors

• Basic building blocks,

• bit-level operations

• combine into word-level operations
### NOT

<table>
<thead>
<tr>
<th>A</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

$Q = \neg A$

also written

$Q = \overline{A}$

### XOR

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

$Q = A \oplus B$

### AND

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

$Q = A \land B$

also written

$Q = A \cdot B$

### NAND

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

$Q = \neg(A \land B)$

### OR

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

$Q = A \lor B$

also written

$Q = A + B$
This is the NOR gate as seen from above.

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
8-bit adder
8-bit adder
Circuit complexity

- Longest path (= time)
- Nr or elements (= size)
• Most of the available algorithms to compute \( \pi \), on the other hand, cannot be easily split up into parallel portions. They require the results from a preceding step to effectively carry on with the next step. Such problems are called inherently serial problems.
Background

• Terminology
  – Time complexity
  – Speedup
  – Efficiency
  – Scalability

• Communication cost model
Time complexity

• Parallel computation
  – A group of processors work together to solve a problem
  – Time required for the computation is the period from when the first processor starts working until when the last processor stops
Other terminology

- **Speedup**: \( S = T_1 / T_P \)
- **Efficiency**: \( E = S / P \)
- **Work**: \( W = P \, T_P \)

**Scalability**
- How does \( T_P \) decrease as we increase \( P \) to solve the same problem?
- How should the problem size increase with \( P \), to keep \( E \) constant?

**Notation**
- \( P = \) Number of processors
- \( T_1 = \) Time on one processor
- \( T_P = \) Time on \( P \) processors
Sometimes a speedup of more than $N$ when using $N$ processors is observed in parallel computing, which is called super linear speedup. Super linear speedup rarely happens and often confuses beginners, who believe the theoretical maximum speedup should be $N$ when $N$ processors are used.
• One possible reason for a super linear speedup is the **cache effect**

• Super linear speedups can also occur when performing **backtracking** in parallel: One thread can prune a branch of the exhaustive search that another thread would have taken otherwise.
Amdahl law

is used to find the maximum expected improvement to an overall system when only part of the system is improved.
• e.g. 5% non parallelizable => speedup can not be more than 20x!

• [http://en.wikipedia.org/wiki/Amdahl%27s_law](http://en.wikipedia.org/wiki/Amdahl%27s_law)
Sequential programme

Two independent parts

Original process

Make B 5x faster

Make A 2x faster
Communication cost model

• Processes spend some time doing useful work, and some time communicating

• Model communication cost as
  – $T_C = t_s + L t_b$
  – $L =$ message size
  – Independent of location of processes
  – Any process can communicate with any other process
  – A process can simultaneously send and receive one message
Communications

• Simple model for analyzing algorithm communications performance:

\[ t_{\text{comm}} = \text{communication time} \]
\[ = t_{\text{startup}} + (#\text{words} \times t_{\text{data}}) \]
\[ = \text{latency} + ( #\text{words} \times \frac{1}{\text{words/second}} ) \]
\[ = \alpha + w \times \beta \]

*latency* and *bandwidth*
I/O model

• We will ignore I/O issues, for the most part
• We will assume that input and output are distributed across the processors in a manner of our choosing
• Example: Sorting
  – Input: \( x_1, x_2, \ldots, x_n \)
    • Initially, \( x_i \) is on processor \( i \)
  – Output \( x_{p_1}, x_{p_2}, \ldots, x_{p_n} \)
    • \( x_{p_i} \) on processor \( i \)
    • \( x_{p_i} \leq x_{p_{i+1}} \)
Important points

• Efficiency
  – Increases with increase in problem size
  – Decreases with increase in number of processors

• Aggregation of tasks to increase granularity
  – Reduces communication overhead

• Data distribution
  – 2-dimensional may be more scalable than 1-dimensional
  – Has an effect on load balance too

• General techniques
  – Divide and conquer
  – Pipelining
Parallel Architectures

• Single Instruction Stream, Multiple Data Stream (SIMD)
  – One global control unit connected to each processor

• Multiple Instruction Stream, Multiple Data Stream (MIMD)
  – Each processor has a local control unit
## Parallel Computing Architectures
### Flynn’s Taxonomy

<table>
<thead>
<tr>
<th></th>
<th>Single Data Stream</th>
<th>Multiple Data Stream</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Single</strong></td>
<td>SISD</td>
<td>SIMD</td>
</tr>
<tr>
<td><strong>Instruction</strong></td>
<td>uniprocessors</td>
<td>Processor arrays</td>
</tr>
<tr>
<td><strong>Stream</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Multiple</strong></td>
<td>MISD</td>
<td>MIMD</td>
</tr>
<tr>
<td><strong>Instruction</strong></td>
<td>Systolic arrays</td>
<td>Multiprocessors</td>
</tr>
<tr>
<td><strong>Stream</strong></td>
<td></td>
<td>multicomputers</td>
</tr>
</tbody>
</table>
Architecture (continued)

• Shared-Address-Space
  – Each processor has access to main memory
  – Processors may be given a small private memory for local variables

• Message-Passing
  – Each processor is given its own block of memory
  – Processors communicate by passing messages directly instead of modifying memory locations
<table>
<thead>
<tr>
<th>Centralized memory</th>
<th>SMP (Symmetric Multiprocessor)</th>
<th>N/A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distributed memory</td>
<td>NUMA (Non-Uniform Memory Access)</td>
<td>MPP (Massively Parallel Processors)</td>
</tr>
</tbody>
</table>
NUMA architecture

Distributed shared memory network with directory
AMD reveals potent parallel processing breakthrough

Upcoming Kaveri processor will drink from shared-memory Holy Grail

By Rik Myslewski in San Francisco, 1st May 2013

AMD has released details on its implementation of The Next Big Thing in processor evolution, and in the process has unleashed the TNBT of acronyms: the AMD APU (CPU+GPU) HSA hUMA.

Before your eyes glaze over and you click away from this page, know that if this scheme is widely adopted, it could be of great benefit to both processor performance and developer convenience — and to you.

Simply put, what AMD’s heterogeneous Uniform Memory Access (hUMA) does is allow central processing units (CPUs) and graphics processing units (GPUs) — which AMD places on a single die in their accelerated processing units (APUs) — to seamlessly share the same memory in a heterogeneous system architecture (HSA). And that’s a very big deal, indeed.
INTRODUCING hUMA

CPU

UMA

Memory

CPU  CPU  CPU  CPU

APU

NUMA

CPU Memory

CPU  CPU  CPU  CPU  GPU

APU with HSA

hUMA

Memory

CPU  CPU  CPU  CPU  GPU
SMP - Symmetric multiprocessing

• Shared Memory
  – All processes share the same address space
  – Easy to program; also easy to program poorly
  – Performance is hardware dependent; limited memory bandwidth can create contention for memory

• MIMD (multiple instruction multiple data)
  – Each parallel computing unit has an instruction thread
  – Each processor has local memory
  – Processors share data by message passing
  – Synchronization must be explicitly programmed into a code

• NUMA (non-uniform memory access)
  – Distributed memory in hardware, shared memory in software, with hardware assistance (for performance)
  – Better scalability than SMP, but not as good as a full distributed memory architecture
MPP – Massively Parallel

• Each node is an independent system having its own:
  – Physical memory
  – Address space
  – Local disk and network connections
  – Operating system
MPP

- Short for *massively parallel processing*, a type of computing that uses many separate CPUs running in parallel to execute a single program. MPP is similar to symmetric processing (SMP), with the main difference being that in SMP systems all the CPUs share the same memory, whereas in MPP systems, each CPU has its own memory. MPP systems are therefore more difficult to program because the application must be divided in such a way that all the executing segments can communicate with each other. On the other hand, MPP don't suffer from the bottleneck problems inherent in SMP systems when all the CPUs attempt to access the same memory at once.
Interconnection Networks

• Static
  – Each processor is hard-wired to every other processor

  ![Completely Connected](image1)
  ![Star-Connected](image2)
  ![Bounded-Degree (Degree 4)](image3)

• Dynamic
  – Processors are connected to a series of switches
Adding $n$ numbers on the mesh

Adding $n$ numbers in $\sqrt{n}$ steps
Adding $n$ numbers on the hypercube

Adding $n$ numbers in $\log n$ steps
128-way fat tree

http://en.wikipedia.org/wiki/Hypercube
Why Do Parallel Computing

• Time: Reduce the turnaround time of applications
• Performance: Parallel computing is the only way to extend performance toward the TFLOP realm
• Cost/Performance: Traditional vector computers become too expensive as one pushes the performance barrier
• Memory: Applications often require memory that goes beyond that addressable by a single processor
• Whole classes of important algorithms are ideal for parallel execution. Most algorithms can benefit from parallel processing such as Laplace equation, Monte Carlo, FFT (signal processing), image processing

• Life itself is a set of concurrent processes
  – Scientists use modelling so why not model systems in a way closer to nature
Some Misconceptions

• Requires new parallel languages?
  – No. Uses standard languages and compilers (Fortran, C, C++, Java, Occam)
  • However, there are some specific parallel languages such as Qlisp, Mul-T and others – check out:
    http://ceu.fi.udc.es/SAL/C/1/index.shtml

• Requires new code?
  – No. Most existing code can be used. Many production installations use the same code base for serial and parallel code.
Cont...

• Requires confusing parallel extensions?
  – No. They are not that bad. Depends on how complex you want to make it. From nothing at all (letting the compiler do the parallelism) to installing semaphores yourself

• Parallel computing is difficult:
  – No. Just different and subtle. Can be akin to assembler language programming 😊
The PRAM Model

• **Parallel Random Access Machine**
  – Theoretical model for parallel machines
  – p processors with uniform access to a large memory bank
  – MIMD
  – UMA (uniform memory access) – Equal memory access time for any processor to any address
Memory Protocols

• Exclusive-Read Exclusive-Write
• Exclusive-Read Concurrent-Write
• Concurrent-Read Exclusive-Write
• Concurrent-Read Concurrent-Write

• If concurrent write is allowed we must decide which value to accept
An Example: Read-Write Register

Suppose we have a read-write register initialized to value 0.

Both READS can output 1 or 2. For linearizability, they should both output the same value.
An Example: Read-Write Register

Suppose we have a read-write register initialized to value $0$.

<table>
<thead>
<tr>
<th>Process</th>
<th></th>
<th></th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P$</td>
<td>WRITE(1)</td>
<td>WRITE(3)</td>
<td></td>
</tr>
<tr>
<td>$Q$</td>
<td>WRITE(2)</td>
<td>READ→?</td>
<td></td>
</tr>
<tr>
<td>$R$</td>
<td></td>
<td>READ→?</td>
<td></td>
</tr>
</tbody>
</table>

Both READS can output 1, 2 or 3. For linearizability, they should not output 1 and 2.
An Example: Read-Write Register

Suppose we have a read-write register initialized to value 0.

Process

\[
\begin{align*}
P & \quad \text{WRITE}(1) \\
Q & \quad \text{WRITE}(2) \quad \text{READ} \rightarrow 1 \\
R & \quad \text{READ} \rightarrow 2
\end{align*}
\]
An Example: Read-Write Register

Suppose we have a read-write register initialized to value 0.

<table>
<thead>
<tr>
<th>Process</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$P$</td>
<td>WRITE(1)</td>
<td>Not permitted ✗</td>
</tr>
<tr>
<td>$Q$</td>
<td>WRITE(2)</td>
<td>READ→1</td>
</tr>
<tr>
<td>$R$</td>
<td></td>
<td>READ→2</td>
</tr>
</tbody>
</table>

Eric Ruppert  Lock-free Shared Data Structures
Linearizability

An object (built in hardware or software) is linearizable if its operations seem to occur instantaneously.

More formally:

For every execution that uses the objects, we can choose a ☆ inside each operation’s time interval so that all operations would return the same results if they were performed sequentially in the order of their ☆s.

The ☆ is called the linearization point of the operation.
Linearizability

In **concurrent programming**, an operation (or set of operations) is **atomic, linearizable, indivisible** or **uninterruptible** if it appears to the rest of the system to occur instantaneously. Atomicity is a guarantee of **isolation** from **concurrent processes**. Additionally, atomic operations commonly have a **succeed-or-fail** definition—they either successfully change the state of the system, or have no apparent effect.

Primaive level operations

- Processors have instructions that can be used to implement locking and lock-free and wait-free algorithms. The ability to temporarily inhibit interrupts, ensuring that the currently running process cannot be context switched, also suffices on a uniprocessor. These instructions are used directly by compiler and operating system writers but are also abstracted and exposed as bytecodes and library functions in higher-level languages:
Primitive operations:

- atomic read-write;
- test-and-set;
- fetch-and-add;
- compare-and-swap;
- load-link/store-conditional.

Read Value, Modify, Write Back – in an “instant”
- Locking – coarse or fine-grained
  - Non-blocking (some can always advance)
  - Wait-free (each thread will eventually finish)
- Update value (if the same)
- ABA problem
- Update counter
- CAS (compare and swap)
- Lock-free data structures...
- Slides from Eric Ruppert
• Queues

Non-blocking algorithm

In computer science, an algorithm is called non-blocking if failure or suspension of any thread cannot cause failure or suspension of another thread;¹ for some operations, these algorithms provide a useful alternative to traditional blocking implementations. A non-blocking algorithm is lock-free if there is guaranteed system-wide progress, and wait-free if there is also guaranteed per-thread progress.
Example: Merge Sort

P1
1 8
4 5
2 7
3 6

P2
1 4 5 8
2 3 6 7

P1
1 2 3 4 5 6 7 8
Merge Sort Analysis

• Number of compares
  - \(1 + 3 + \ldots + (2^i - 1) + \ldots + (n-1)\)
  - \(\sum_{i=1..\lg(n)} 2^i - 1 = 2n - 2\lg n = \Theta(n)\)

• We have improved from \(n\lg(n)\) to \(n\) simply by applying the old algorithm to parallel computing, by altering the algorithm we can further improve merge sort to \((\lg n)^2\)
CLRS

- **parallel** (to execute for-loop iterations in parallel)
- **spawn** (create a new task/thread), and
- **sync** (wait for all spawned children)

\[
\begin{align*}
\text{Fib}(n) &: \text{if } n \leq 1 \text{ return } n \\
& \text{else} \\
& x = \text{Fib}(n-1) \\
& y = \text{Fib}(n-2) \\
& \text{return } x + y
\end{align*}
\]
Parallel loops

Mat-Vec( A, x )
1 n = A.rows
2 let y be a new vector of length n
3 parallel for i = 1 to n
4 yi = 0
5 parallel for i = 1 to n
6 for j = 1 to n
7 yi = yi + aij * xj
8 return y
Parallel loop (5-7) divide & conquer

Mat-Vec-Main-Loop( A, x,y,n,i,i’ )
1   if i== i’
2       for j=1 to n
3           yi = yi + aij * xj
4    else   mid = \left\lfloor \frac{(i+i’)}{2} \right\rfloor
5    spawn Mat-Vec-Main-Loop( A, x,y,n,i,mid )
6    Mat-Vec-Main-Loop( A, x,y,n,mid+1, i’ )
7    sync
• deterministic – always the same result
• non-deterministic – may depend on order, runto run
• race

Race-Example()
  x=0
  parallel for i=1 to 2
    x = x+1
  print x
Parallel Design and Dynamic Programming

• Often in a dynamic programming algorithm a given row or diagonal can be computed simultaneously

• This makes many dynamic programming algorithms amenable for parallel architectures
Current Applications that Utilize Parallel Architectures

• Computer Graphics Processing
• Video Encoding
• Accurate weather forecasting
• Scientific computing, modelling
• ...
Parallel addition features

• If n >> P
  – * Each processor adds $n/P$ distinct numbers
  – Perform parallel reduction on P numbers
  – $T_P \sim n/P + (1 + t_s + t_b) \log P$
  – Optimal P obtained by differentiating wrt P
    • $P_{opt} \sim n/(1 + t_s + t_b)$
    • If communication cost is high, then fewer processors ought to be used
  – $E = [1 + (1 + t_s + t_b) P \log P/n]^{-1}$
    • * As problem size increases, efficiency increases
    • * As number of processors increases, efficiency decreases
Some common collective operations

- **Broadcast**: A sends A to A, B, C, D.
- **Gather**: A, B, C, D receive A, A, A, A.
- **Scatter**: A, B, C, D send A, B, C, D to A, B, C, D.
- **All Gather**: A, B, C, D gather A, B, C, D.
Broadcast

- $T \sim (t_s + L t_b) \log P$
  - $L$: Length of data
Gather/Scatter

- Gather: Data move towards the root
- Scatter: Review question
- $T \sim t_s \log P + PL t_b$

Note: $\sum_{i=0}^{\log P-1} 2^i = \frac{2^{\log P} - 1}{2-1} = P-1 \sim P$
All gather

- Equivalent to each processor broadcasting to all the processors
All gather
All gather
All gather

- $T_n \sim t_s \log P + PLt_b$
Matrix-vector multiplication

• $c = A \, b$
  – Often performed repeatedly
    • $b_i = A \, b_{i-1}$
  – We need same data distribution for $c$ and $b$

• One dimensional decomposition
  – Example: row-wise block striped for $A$
    • $b$ and $c$ replicated
  – Each process computes its components of $c$ independently
  – Then all-gather the components of $c$
1-D matrix-vector multiplication

- Each process computes its components of $c$ independently
  - Time = $\Theta(n^2/P)$
- Then all-gather the components of $c$
  - Time = $t_s \log P + t_b n$
- Note: $P \leq n$
2-D matrix-vector multiplication

- Processes $P_{i0}$ sends $B_i$ to $P_{0i}$
  - Time: $t_s + t_b n/P^{0.5}$
- Processes $P_{0j}$ broadcast $B_j$ to all $P_{ij}$
  - Time = $t_s \log P^{0.5} + t_b n \log P^{0.5} / P^{0.5}$
- Processes $P_{ij}$ compute $C_{ij} = A_{ij}B_j$
  - Time = $\Theta(n^2 / P)$
- Processes $P_{ij}$ reduce $C_{ij}$ on to $P_{i0}$, $0 \leq i \leq P^{0.5}$
  - Time = $t_s \log P^{0.5} + t_b n \log P^{0.5} / P^{0.5}$
- Total time = $\Theta(n^2 / P + t_s \log P + t_b n \log P / P^{0.5})$
  - $P \leq n^2$
  - * More scalable than 1-dimensional decomposition
Matrix multiplication

In mathematics, matrix multiplication is the operation of multiplying a matrix with either a scalar or another matrix. This article gives an overview of the various ways to perform matrix multiplication.

Ordinary matrix product

This is the most often used and most important way to multiply matrices. It is defined between two matrices only if the number of columns of the first matrix is the same as the number of rows of the second matrix. If $A$ is an $m$-by-$n$ matrix and $B$ is an $n$-by-$p$ matrix, then their product is an $m$-by-$p$ matrix denoted by $AB$ (or sometimes $A \cdot B$). If $C = AB$, and $c_{ij}$ denotes the entry in $C$ at position $(i,j)$, then

$$c_{ij} = \sum_{r=1}^{n} a_{ir} b_{rj} = a_{i1} b_{1j} + a_{i2} b_{2j} + \cdots + a_{in} b_{nj},$$

for each pair $i$ and $j$ with $1 \leq i \leq m$ and $1 \leq j \leq p$. The algebraic system of "matrix units" summarises the abstract properties of this kind of multiplication.

Calculating directly from the definition

The picture to the left shows how to calculate the $(1,2)$ element and the $(3,3)$ element of $AB$ if $A$ is a $4 \times 2$ matrix, and $B$ is a $2 \times 3$ matrix. Elements from each matrix are paired off in the direction of the arrows, each pair is multiplied and the products are added. The location of the resulting number in $AB$ corresponds to the row and column that were considered.

$$(AB)_{1,2} = \sum_{r=1}^{2} a_{1r} b_{r2} = a_{11} b_{12} + a_{12} b_{22}$$

$$(AB)_{3,3} = \sum_{r=1}^{2} a_{3r} b_{r3} = a_{31} b_{13} + a_{32} b_{23}$$

For example:

$$\begin{bmatrix} 1 & 0 & 2 \\ -1 & 3 & 1 \end{bmatrix} \cdot \begin{bmatrix} 3 & 1 \\ 2 & 1 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 1 \times 3 + 0 \times 2 + 2 \times 1 & 1 \times 1 + 0 \times 1 + 2 \times 0 \\ -1 \times 3 + 3 \times 2 + 1 \times 1 & -1 \times 1 + 3 \times 1 + 1 \times 0 \end{bmatrix} = \begin{bmatrix} 5 \\ 4 \end{bmatrix}$$

The coefficients-vectors method

This matrix multiplication can also be considered from a slightly different viewpoint: it adds vectors together after being multiplied by different coefficients. If $A$ and $B$ are matrices given by:

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \quad B = \begin{bmatrix} b_{1} \\ b_{2} \end{bmatrix}$$

then

$$AB = \begin{bmatrix} a_{11}b_{1} + a_{12}b_{2} \\ a_{21}b_{1} + a_{22}b_{2} \end{bmatrix}$$

The above example shows that this method is equivalent to the ordinary matrix product.
• 31.2 Strassen's algorithm for matrix multiplication
• This section presents Strassen's remarkable recursive algorithm for multiplying $n \times n$ matrices that runs in $(n^{\lg 7}) = O(n^{2.81})$ time. For sufficiently large $n$, therefore, it outperforms the naive $(n^3)$ matrix-multiplication algorithm MATRIX-MULTIPLY from Section 26.1.
\[ C = AB \]

\[
\begin{pmatrix}
  r & s \\
  t & u
\end{pmatrix} =
\begin{pmatrix}
  a & b \\
  c & d
\end{pmatrix}
\begin{pmatrix}
  e & g \\
  f & h
\end{pmatrix}
\]

- \( r = ae + bf \)
- \( s = ag + bh \)
- \( t = ce + df \)
- \( u = cg + dh \)
- \( T(n) = 8T(n/2) + \Theta(n^2). = O(n^3) \)
Strassen discovered a different recursive approach that requires only 7 recursive multiplications of \( n/2 \times n/2 \) matrices and \( \Theta(n^2) \) scalar additions and subtractions, yielding the recurrence

\[
T(n) = 7T(n/2) + \Theta(n^2)
\]

\[
= \Theta(n^{\log_2 7})
\]

\[
= \mathcal{O}(n^{2.81})
\]

(31.15)

Strassen’s method has four steps:

1. Divide the input matrices \( A \) and \( B \) into \( n/2 \times n/2 \) submatrices, as in equation (31.9).

2. Using \( \Theta(n^2) \) scalar additions and subtractions, compute 14 \( n/2 \times n/2 \) matrices \( A_1, B_1, A_2, B_2, \ldots, A_7, B_7 \).

3. Recursively compute the seven matrix products \( P_i = A_iB_i \) for \( i = 1, 2, \ldots, 7 \).

4. Compute the desired submatrices \( r, s, t, u \) of the result matrix \( C \) by adding and/or subtracting various combinations of the \( P_i \) matrices, using only \( \Theta(n^2) \) scalar additions and subtractions.
Matrix operations are parallelizable

• Problems that are expressed in forms of matrix operations are often easy to automatically parallelise

• Fortran, etc – programming languages can achieve that at no extra effort
The Process

- A running executable of a (compiled and linked) program written in a standard sequential language (i.e. F77 or C) with library calls to implement the message passing.
- A process executes on a processor.
  - All processes are assigned to processors in a one-to-one mapping (simplest model of parallel programming).
  - Other processes may execute on other processors.
- A process communicates and synchronizes with other processes via messages.
- A process is uniquely identified by:
  - The node on which it is running.
  - Its process id (PID).
- A process does not migrate from node to node (though it is possible for it to migrate from one processor to another within an SMP node).
Solving Problems in Parallel

It is true that the hardware defines the parallel computer. However, it is the software that makes it usable.

Parallel programmers have the same concern as any other programmer:
- Algorithm design,
- Efficiency
- Debugging ease
- Code reuse, and
- Lifecycle.
However, they are also concerned with:
- Concurrency and communication
- Need for speed (nee high performance), and
- Plethora and diversity of architecture
Fosters Four step Process for Designing Parallel Algorithms

1. Partitioning – process of dividing the computation and the data into many small pieces – decomposition

2. Communication – local and global (called overhead) minimizing parallel overhead is an important goal and the following check list should help the communication structure of the algorithm
   1. The communication operations are balanced among tasks
   2. Each task communicates with only a small number of neighbours
   3. Tasks can perform their communications concurrently
   4. Tasks can perform their computations concurrently
3. Agglomeration is the process of grouping tasks into larger tasks in order to improve the performance or simplify programming. Often in using MPI this is one task per processor.

4. Mapping is the process of assigning tasks to processors with the goal to maximize processor utilization.
Solving Problems in Parallel

- Decomposition determines:
  - Data structures
  - Communication topology
  - Communication protocols
- Must be looked at early in the process of application development
- Standard approaches
Decomposition methods

- Perfectly parallel
- Domain
- Control
- Object-oriented
- Hybrid/layered (multiple uses of the above)
For the program

- Choose a decomposition
  - Perfectly parallel, domain, control etc.
- Map the decomposition to the processors
  - Ignore topology of the system interconnect
  - Use natural topology of the problem
- Define the inter-process communication protocol
  - Specify the different types of messages which need to be sent
  - See if standard libraries efficiently support the proposed message patterns
Perfectly parallel

- Applications that require little or no inter-processor communication when running in parallel
- Easiest type of problem to decompose
- Results in nearly perfect speed-up
In simulation and modelling this is the most common solution

- The solution space (which often corresponds to the real space) is divided up among the processors. Each processor solves its own little piece
- Finite-difference methods and finite-element methods lend themselves well to this approach
- The method of solution often leads naturally to a set of simultaneous equations that can be solved by parallel matrix solvers
- Sometimes the solution involves some kind of transformation of variables (i.e. Fourier Transform). Here the domain is some kind of phase space. The solution and the various transformations involved can be parallelized
Output Data Decomposition: Example

Consider the problem of multiplying two $n \times n$ matrices $A$ and $B$ to yield matrix $C$. The output matrix $C$ can be partitioned into four tasks as follows:

\[
\begin{pmatrix}
A_{1,1} & A_{1,2} \\
A_{2,1} & A_{2,2}
\end{pmatrix}
\cdot
\begin{pmatrix}
B_{1,1} & B_{1,2} \\
B_{2,1} & B_{2,2}
\end{pmatrix}
\rightarrow
\begin{pmatrix}
C_{1,1} & C_{1,2} \\
C_{2,1} & C_{2,2}
\end{pmatrix}
\]

Task 1: $C_{1,1} = A_{1,1}B_{1,1} + A_{1,2}B_{2,1}$

Task 2: $C_{1,2} = A_{1,1}B_{1,2} + A_{1,2}B_{2,2}$

Task 3: $C_{2,1} = A_{2,1}B_{1,1} + A_{2,2}B_{2,1}$

Task 4: $C_{2,2} = A_{2,1}B_{1,2} + A_{2,2}B_{2,2}$
Output Data Decomposition: Example

A partitioning of output data does not result in a unique decomposition into tasks. For example, for the same problem as in previous foil, with identical output data distribution, we can derive the following two (other) decompositions:

<table>
<thead>
<tr>
<th>Decomposition I</th>
<th>Decomposition II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1: $C_{1,1} = A_{1,1} B_{1,1}$</td>
<td>Task 1: $C_{1,1} = A_{1,1} B_{1,1}$</td>
</tr>
<tr>
<td>Task 2: $C_{1,1} = C_{1,1} + A_{1,2} B_{2,1}$</td>
<td>Task 2: $C_{1,1} = C_{1,1} + A_{1,2} B_{2,1}$</td>
</tr>
<tr>
<td>Task 3: $C_{1,2} = A_{1,1} B_{1,2}$</td>
<td>Task 3: $C_{1,2} = A_{1,2} B_{2,2}$</td>
</tr>
<tr>
<td>Task 4: $C_{1,2} = C_{1,2} + A_{1,2} B_{2,2}$</td>
<td>Task 4: $C_{1,2} = C_{1,2} + A_{1,1} B_{1,2}$</td>
</tr>
<tr>
<td>Task 5: $C_{2,1} = A_{2,1} B_{1,1}$</td>
<td>Task 5: $C_{2,1} = A_{2,2} B_{2,1}$</td>
</tr>
<tr>
<td>Task 6: $C_{2,1} = C_{2,1} + A_{2,2} B_{2,1}$</td>
<td>Task 6: $C_{2,1} = C_{2,1} + A_{2,1} B_{1,1}$</td>
</tr>
<tr>
<td>Task 7: $C_{2,2} = A_{2,1} B_{1,2}$</td>
<td>Task 7: $C_{2,2} = A_{2,1} B_{1,2}$</td>
</tr>
<tr>
<td>Task 8: $C_{2,2} = C_{2,2} + A_{2,2} B_{2,2}$</td>
<td>Task 8: $C_{2,2} = C_{2,2} + A_{2,2} B_{2,2}$</td>
</tr>
</tbody>
</table>
Control decomposition

- If you cannot find a good domain to decompose, your problem might lend itself to control decomposition
  - Good for:
    - Unpredictable workloads
    - Problems with no convenient static structures
  - One set of control decomposition is functional decomposition
    - Problem is viewed as a set of operations. It is among operations where parallelization is done
    - Many examples in industrial engineering (i.e. modelling an assembly line, a chemical plant, etc.)
    - Many examples in data processing where a series of operations is performed on a continuous stream of data
Control is distributed, usually with some distribution of data structures

Some processes may be dedicated to achieve better load balance

Examples

- Image processing: given a series of raw images, perform a series of transformation that yield a final enhanced image. Solve this in a functional decomposition (each process represents a different function in the problem) using data pipelining

- Game playing: games feature an irregular search space. One possible move may lead to a rich set of possible subsequent moves to search.
  - Need an approach where work can be dynamically assigned to improve load balancing
  - May need to assign multiple processes to work on a particularly promising lead
Any problem that involve search (or computations) whose scope cannot be determined a priori, are candidates for control decomposition

- Calculations involving multiple levels of recursion (i.e. genetic algorithms, simulated annealing, artificial intelligence)
- Discrete phenomena in an otherwise regular medium (i.e. modelling localized storms within a weather model)
- Design-rule checking in micro-electronic circuits
- Simulation of complex systems
- Game playing, music composing, etc..
Object-oriented decomposition

- Object-oriented decomposition is really a combination of functional and domain decomposition
  - Rather than thinking about a dividing data or functionality, we look at the objects in the problem
  - The object can be decomposed as a set of data structures plus the procedures that act on those data structures
  - The goal of object-oriented parallel programming is distributed objects
- Although conceptually clear, in practice it can be difficult to achieve good load balancing among the objects without a great deal of fine tuning
  - Works best for fine-grained problems and in environments where having functionally ready at-the-call is more important than worrying about under-worked processors (i.e. battlefield simulation)
  - Message passing is still explicit (no standard C++ compiler automatically parallelizes over objects).
Partitioning the Graph of Lake Superior

Random Partitioning

Partitioning for minimum edge-cut.
Decomposition summary

- A good decomposition strategy is
  - Key to potential application performance
  - Key to programmability of the solution
- There are many different ways of thinking about decomposition
  - Decomposition models (domain, control, object-oriented, etc.) provide standard templates for thinking about the decomposition of a problem
  - Decomposition should be natural to the problem rather than natural to the computer architecture
  - Communication does no useful work; keep it to a minimum
  - Always wise to see if a library solution already exists for your problem
  - Don’t be afraid to use multiple decompositions in a problem if it seems to fit
Summary of Software

- **Compilers**
  - Moderate $O(4-10)$ parallelism
  - Not concerned with portability
  - Platform has a parallelizing compiler

- **OpenMP**
  - Moderate $O(10)$ parallelism
  - Good quality implementation exists on the platform
  - Not scalable

- **MPI**
  - Scalability and portability are important
  - Needs some type of message passing platform
  - A substantive coding effort is required

- **PVM**
  - All MPI conditions plus fault tolerance
  - Still provides better functionality in some settings

- **High Performance Fortran (HPF)**
  - Like OpenMP, but new language constructs provide a data-parallel implicit programming model

- **P-Threads**
  - Not recommended
  - Difficult to correct and maintain programs
  - Not scalable to large number of processors

- **High level libraries**
  - POOMA and HPC++
  - Library is available and it addresses a specific problem
Apache Hadoop NextGen MapReduce (YARN)

MapReduce has undergone a complete overhaul in hadoop-0.23 and we now have, what we call, MapReduce 2.0 (MRv2) or YARN.

The fundamental idea of MRv2 is to split up the two major functionalities of the JobTracker, resource management and job scheduling/monitoring, into separate daemons. The idea is to have a global ResourceManager (RM) and per-application ApplicationMaster (AM). An application is either a single job in the classical sense of Map-Reduce jobs or a DAG of jobs. The ResourceManager and per-node slave, the NodeManager (NM), form the data-computation framework. The ResourceManager is the ultimate authority that arbitrates resources among all the applications in the system.

The per-application ApplicationMaster is, in effect, a framework specific library and is tasked with negotiating resources from the ResourceManager and working with the NodeManager(s) to execute and monitor the tasks.

The ResourceManager has two main components: Scheduler and ApplicationsManager.

The Scheduler is responsible for allocating resources to the various running applications subject to familiar constraints of capacities, queues etc. The Scheduler is pure scheduler in