Serial vs. Parallel Algorithms

Serial algorithms are suitable for running on sequential uniprocessor computers. These sequential computers are not built anymore. We live in the age of parallel computers, where multiple instructions are executed at the same time. These parallel computers come in different forms:

- **Single multi core chips.** A core is a full-fledged processor. Each core can access a single shared memory.
- **Single multi core chips with accelerators.** An accelerator is a special co-processor that can execute certain (simple) codes in parallel, e.g. a vector processor that executes the same instruction on an array of values.
- **Distributed memory multi-computers,** where each processor’s memory is private, and processors communicate via an interconnection network.

We will concentrate on the first class: multi-core shared memory computers.
Dynamic Multithreading

Programs can specify parallelism through:

1. **Nested Parallelism**, where a function call is “spawned”, allowing the caller and spawned function to run in parallel. We also call this Task Parallelism.

2. Loop **Parallelism**, where the iterations of the loop can execute in parallel.

These parallel loop iterations and tasks are executed by “virtual processors” or threads. Exactly when a thread executes and on which core it executes is not decided by the programmer, but by the run time system, which coordinates, schedules and manages the parallel computing resources. This lightens the task of writing parallel programs, as we don’t have to worry about data partitioning (shared memory) and task scheduling.
Parallel constructs

Parallel tasks are created by a *spawn* and at the end of the task’s execution synchronized with the parent by a *sync*. Parallel tasks naturally follow the divide-and-conquer paradigm.

Parallel loops are created using *parallel* and *new* constructs (later).

Removing spawn, sync, parallel and new from the program brings back the original sequential code.

There is a growing number of dynamic multi threading platforms. E.g., in cs475 we study OpenMP (open multi processing), built on top of C, C++, or Fortran.
The basics of dynamic multithreading

Fibonacci sequence:
\[ F_0 = 0 \]
\[ F_1 = 1 \]
\[ F_n = F_{n-1} + F_{n-2} \quad n>1 \]

Simple recursive solution:

```python
Fib(n) :
    if n<=1
        return n
    else
        x = Fib(n-1)
        y = Fib(n-2)
        return x+y
```

Why do you not want to compute Fibonacci for large \( n \) this way?
How many nodes in this tree? (order of magnitude)
How would you write an efficient Fibonacci?
Run time of Fib(n)

T(n) denotes the run time of Fib(n):

\[ T(n) = T(n-1) + T(n-2) + \Theta(1) \]

the two recursive calls and

some constant time split and combine extra work

Claim: \( T(n) = \Theta(F_n) \)

Proof: strong induction.

Base: all constants, OK.

Step: assume

\[ T(m) = \Theta(F_m) \leq aF_m - b \quad a, b \text{ non negative constants, } 0 \leq m < n \]

Then: \[ T(n) \leq aF_{n-1} - b + aF_{n-2} - b + \Theta(1) = a(F_{n-1} + F_{n-2}) - b - (b - \Theta(1)) \]

\[ = aF_n - b - (b - \Theta(1)) \leq aF_n - b \]

In fact, we can show that \( T(n) = \Theta(\phi^n) \quad \phi = (1+\sqrt{5})/2 \quad (CS420) \)
Parallel Fibonacci

P-Fib(n):
    if n<=1
        return n
    else
        x = spawn P-Fib(n-1) // spawn
        y = P-Fib(n-2) // call
        sync
        return x+y

**spawn**: the caller (parent) can compute on **in parallel with** the called (child); it does not have to, but it may (up to the RTS when and where to schedule tasks)

**sync**: the parent must wait for all its spawned children to have completed, before proceeding. The sync is required to avoid summing x+y, before x is computed

**return**: in addition to explicit sync-s, return statement execute a sync implicitly, so that the parent waits for its children to complete
A Multithreaded execution model

The multithreaded computation can be viewed as executing a Directed Acyclic Graph $G=(V,E)$, called computation DAG

Example: computation DAG for P-Fib(4)

Assuming each strand takes one time unit 
(totals) work equals 17 time units 
span equals 8 (#critical path strands)
A Multithreaded execution model

The multithreaded computation can be viewed as executing a Directed Acyclic Graph $G=(V,E)$, called computation DAG, which is embedded in the call tree.

Edge $(u,v)$: $u$ executes before $v$ 
(u,v) indicates a dependency:
- if a node (strand) has two successors, one of them is spawned
- if a strand has multiple predecessors, they sync before execution continues
If there is a path from $u$ to $v$, they execute in series, otherwise they execute in parallel

Spawn and call edges point downward. Horizontal (continuation) edges indicate that the parent may keep computing while spawn executes. Return edges point up.

Execution starts in a single initial strand (which one?) and ends in a single final strand (which one?)
Impact of schedule

Unfolded DAG for PF-3

2 Processors

Schedule 1
P2  3  5  7
P1  1  2  4  6  8  9
-------------------------
time  1  2  3  4  5  6

Schedule 2
P2  3  6
P1  1  2  4  5  7  8  9
-------------------------
time  1  2  3  4  5  6  7

Idle time: number of empty slots (processor not busy) in schedule
schedule 1: 3, schedule 2: 5
Performance Measures

**Work:** the total time to execute the program sequentially. Assuming 1 time unit per strand, this is the number of nodes in the DAG.

**Span:** longest time to execute the strands along any path in the tree, i.e., the critical path length of the DAG.

The run time of the program depends also on schedule and number of processors.
Performance Measures: time

Work: the total time to execute the program sequentially. Assuming 1 time unit per strand, this is the number of nodes in the DAG.

Span: longest path length of the DAG.

$T_1$: the time to execute the program with 1 processor ($T_1=$work)

$T_P$: the time to execute the program with $P$ processors

As we have seen, different schedules can sometimes take different time, but we always assume greedy scheduling: if there are ($\geq 1$) strands ready and a processor is available, a strand will be executed. (Which strand depends on the scheduler.)

Simplifying assumption: We are assuming no time cost for communication between the strands or memory accesses.

We call this model of computation ideal.
Work Law and Span Law

Work Law:
in one step $P$ processors can do at most $P$ units of work:

$$T_P \geq T_1/P$$

Span Law:
$T_\infty$: the time to execute the program with unlimited # processors ($T_\infty = \text{span}$) is less or equal the time to execute the program with a fixed #processors $P$

$$T_\infty \leq T_P \text{ or } T_P \geq T_\infty$$
Performance Measures: parallelism and speedup

$S_P$: speedup with $P$ processors: $T_1 / T_P$.

(Average) Parallelism: $T_1 / T_\infty$ (sometimes called $\pi$ (pi)):
- average amount of work that can be done per time step
- or, maximum possible speedup (when no idle time exists)

With $P$ processors you can only go $P$ times faster than with 1 processor: $S_P \leq P$

linear speedup: $S_P = fP$ ($0 < f \leq 1$)

ideal speedup: $f=1$ or $S_P = P$
- (no idle time, all processors busy all the time)

When $P > \pi$ there will be idle time and thus no perfect speedup.
Exercise

Fill in

$T_1$: 
$T_\infty$: 
$\Pi$: 

Is there idle time for: 
$P=1$  $P=2$  $P=3$  ?

Create a schedule for $P=3$

$P_3$

$P_2$

$P_1$

Time

$T_3$:  
$S_3$: 

Is $T_4 < T_3$,  ?  explain

Unfolded DAG for PF-3
We consider greedy schedulers only.

If there are at least $P$ strands available in a time step, all processors execute, and we call this a complete step.

If there are fewer than $P$ strands available in a time step, some processors will be idle, and we call that an incomplete step.

From the work law we know that at best $T_P = T_1 / P$.

From the span law we know that at best $T_P = T_\infty$. 

Bound on $T_P$
**Theorem: bound on** $T_P$

**Theorem:** $T_P \leq T_1/P + T_\infty$

**Proof:**
- There can be at most $\lfloor T_1/P \rfloor$ complete steps, otherwise there would be more than $T_1$ work.
- There can be at most $T_\infty$ (critical path length) incomplete steps. This happens when all steps are incomplete in which case in every step the remaining critical path length is decreased.

Steps are either complete or incomplete, therefore: $T_P \leq T_1/P + T_\infty$

**QED**

We define **slackness** as $(T_1/T_\infty)/P$. Slackness is the factor by which program parallelism $\pi$ exceeds machine parallelism $P$. More slackness, more idle time. If all steps are incomplete, the number of processors is larger than the maximum parallelism in the program.
Exercises

1. Draw a task graph $G$ with 8 nodes, where every step is incomplete for $P=2$. Determine $T_2$, $T_\infty$

2. Use the schedule for $P=3$ from the previous exercise
   - Determine #incomplete steps, #complete steps
   - Determine $T_1/P$, $T_\infty$, $T_P$
   - Verify the theorem for this case
3 - How many critical paths?

- Critical path length?

- For $P = 1, 2$ and $3$ draw a schedule and check that for each incomplete step the critical path length decreases.

- Does the critical path length ONLY decrease on incomplete steps?
Corollary:

\[ T_P \] of any computation scheduled by a greedy scheduler is within a factor of 2 of optimal

Proof: Let \( T^*_p \) be the run time of an optimal schedule

Work law: \( T^*_p \geq T_1/P \) Span law: \( T^*_p \geq T_\infty \)

therefore \( T^*_p \geq \max(T_1/P, T_\infty) \)

For any \( P \) processor computation we have the theorem:

\[ T_P \leq T_1/P + T_\infty \]
\[ \leq 2 \max(T_1/P, T_\infty) \]
\[ \leq 2 T^*_p \]

QED

Also, if \( P \ll T_1/T_\infty \) i.e., \( P \ll \pi \) (\ll: much less than) then most steps are complete and thus \( T_p \sim T_1/T_\infty \) (\sim: is about), i.e. \( S_p \sim P \) i.e. the speedup is near perfect. The scheduling algorithm has thus a low impact on the performance.
We can compose two computations \( A \) and \( B \) in series or in parallel.

**In series:** \( A \) is followed by \( B \)

- **Work:** \( T_1(A \cup B) = T_1(A) + T_1(B) \)
- **Span:** \( T_\infty(A \cup B) = T_\infty(A) + T_\infty(B) \)

**In parallel:** \( A \) and \( B \) execute in parallel

- **Work:** \( T_1(A \cup B) = T_1(A) + T_1(B) \)
- **Span:** \( T_\infty(A \cup B) = \max(T_\infty(A), T_\infty(B)) \)
Critique of the ideal execution model

Why are the previous observations highly (unrealistically) optimistic?

1. Communication between strands is NOT free of time cost. Determining that a strand is ready for execution, and starting it on an available processor, takes time.

2. Accessing memory is not free; it takes (A LOT OF) time. In modern computers memory accesses take many processor clock cycles.

This is why modern computers have a complex cache architecture.
Analyzing multithreaded computation: Fibonacci

\[
P_{\text{Fib}}(n): \quad \begin{align*}
\text{if } n & \leq 1 \\
& \quad \text{return } n \\
\text{else} \\
x &= \text{spawn } P_{\text{Fib}}(n-1) \quad \text{// spawn} \\
y &= P_{\text{Fib}}(n-2) \quad \text{// call} \\
\text{sync} \\
& \quad \text{return } x+y
\end{align*}
\]

- Work: we already did this: execution time of Fib:

\[
T_1(n) = \Theta(\varphi^n) \quad \varphi = (1+\sqrt{5})/2
\]

- Span: spawn \( P_{\text{Fib}}(n-1) \) and call \( P_{\text{Fib}}(n-2) \) can run in parallel:

\[
T_\infty(n) = \max(T_\infty(n-1), T_\infty(n-2)) + \Theta(1) = T_\infty(n-1) + \Theta(1) \\
T_\infty(n) = \Theta(n)
\]

- Parallelism: \( \Pi = T_1(n)/T_\infty(n) = \Theta(\varphi^n / n) \) which grows fast, and so near perfect speedup can be achieved. BUT!!: WHAT IS WRONG HERE? 
  (hint: It is often easy to write inefficient parallel algorithms 😊 )
How fast can we add up an array of $n$ numbers?

- Hint divide and conquer
  - divide into halves;
  - (recursively) add them up
  - add the result

```python
Reduction(lo, hi):
    if lo = hi
        return A[lo]
    else
        mid = (hi-lo)/2
        x = Reduction(lo, mid):
        y = Reduction(mid+1, hi):
        return x+y
```

- Reduction has (work) complexity $\Theta(n)$
Next step: parallel Reductions - `spawn-sync` one child

```python
P_Reduction(lo, hi):
    if lo = hi
        return A[lo]
    else
        mid = (hi-lo)/2
        x = spawn P_Reduction(lo, mid):
        y = P_Reduction(mid+1, hi):
        sync
        return x+y
```

- Work remains $\Theta(n)$
- Span is $\Theta(lg n)$
- Can be used for any associative operator, not just addition.
Scan: The third step

Instead of a single result, compute an array (prefix sum):

- $i$-th element of the output is the sum of the first $i$ elements of the input array
- divide into halves; (recursively) scan them.
- And what else?

```python
def Scan(lo, hi, X):
    if lo == hi:
        return X[lo]
    else:
        mid = (hi - lo) / 2
        Y[1:mid] = Scan(lo, mid)
        Y[mid+1:hi] = Scan(mid+1, hi)
        // just a for loop
    return Y
```

- Work complexity? $\Theta(n)$
- What is the constant factor?
- Now parallelize (spawn-sync alone is not enough)
Digression: Parallel Loops

The `parallel` keyword before a for statement indicates that all the iterations of the for loop can execute in parallel.

Example: matrix vector product  \( Y_i = \sum_{j=1}^{n} a_{ij}x_j \)  for \( i = 1..n \)

Each \( Y_i \) can be computed by spawning a loop iteration \( i \). However, each spawned iteration needs a private version of \( j \) (\( j \) cannot be shared, WHY?). This is expressed using the `new` keyword:

\[
\text{Mat-Vec}(A,x) :
\]

\[
\begin{align*}
n & = A.\text{rows} \\
y & = \text{float}[n] \\
\text{parallel for } i = 1 \text{ to } n & \\
y_i & = 0 \\
\text{parallel for } i = 1 \text{ to } n & \quad \text{// parallel loop} \\
\text{for new } j = 1 \text{ to } n & \quad \text{// sequential loop} \\
y_i & = y_i + a_{ij}x_j \\
\text{return } y
\end{align*}
\]
The parallel for can be compiled into a divide and conquer tree of spawned processes much like merge sort. (There are many other ways to compile this type of loop.)

**Work:** each internal node \([lo,up]\) does constant spawn, compute, call work. There are \(n-1\) of these nodes, so set up work is \(\Theta(n)\). Each of the \(n\) leaves does \(\Theta(n)\) work. Hence the work is \(\Theta(n^2)\)

**Span:** To set up the tree takes span \(\Theta(\log(n))\). The bulk of the computation is in the (internally sequential) leaves. All leaves can run in parallel and take span \(\Theta(n)\), which dominates building the tree. Hence the span is \(\Theta(n)\).
Parallelizing the scan

\[
\text{P\_Scan}(lo, hi, X): \\
\text{if } lo = hi \text{ return } X[lo] \\
\text{else} \\
\quad \text{mid} = (h-\text{lo})/2 \\
\quad Y[1:\text{mid}] = \text{spawn P\_scan(lo, mid)} \\
\quad Y[\text{mid}+1:hi] = \text{P\_scan(mid}+1, hi) \\
\quad \text{sync} \\
\quad Y[\text{mid}+1:hi] = Y[\text{mid}] + Y[\text{mid}+1:hi] \\
\]

//just a for loop
\[
\text{return } Y
\]

- Use parallel \text{spawn-sync} across recursion, and parallel for for the result-update
- Work: Remains \(\Theta(n)\)
- Span: Now \(\Theta(\lg n)\)

\text{We can scan } n \text{ numbers in } \lg n \text{ time with } \Theta(n) \text{ processors}
Problem:
- Compute an array of the first \( n \) Fibonacci numbers
- Computing only the \( n^{th} \) one is a special case of this problem

Lower bound:
- Is it \( \Theta(n) \), the size of the output?
- No, processors can write outputs in parallel

Recall the memo-Fib: in each iteration, update one value using the previous and pre-previous

\[
F[0] = F[1] = 0;
for \ i \ in \ range(2:n)
F[i] = F[i-1] + F[i-2]
\]
How to **reduce** the Fib to a scan?

- Draw from the memory efficient memo-Fib: copy
  - the new to previous
  - Previous to pre-previous
- **Maintain a 2-element vector**
- Use a matrix notation, to express $F[i]$ and $F[i-1]$ in terms of previous vector $F[i-1]$ and $F[i-2]$:
  
  $F[i] = 1 \times F[i-1] + 1 \times F[i-2]$
  
  $F[i-1] = 1 \times F[i-1] + 0 \times F[i-2]$

\[
\begin{pmatrix}
F_i \\
F_{i-1}
\end{pmatrix}
= 
\begin{bmatrix}
1 & 1 \\
1 & 0
\end{bmatrix}
\begin{pmatrix}
F_{i-1} \\
F_{i-2}
\end{pmatrix}
= 
\begin{bmatrix}
1 & 1 \\
1 & 0
\end{bmatrix}
\begin{bmatrix}
1 & 1 \\
1 & 0
\end{bmatrix}
\begin{pmatrix}
F_{i-2} \\
F_{i-3}
\end{pmatrix}
= 
\begin{bmatrix}
1 & 1 \\
1 & 0
\end{bmatrix}
\cdots
\begin{bmatrix}
1 & 1 \\
1 & 0
\end{bmatrix}
\begin{pmatrix}
F_{i-1} \\
F_{i-0}
\end{pmatrix}
\]