Intro to Parallelism

CSE 332 Summer 2021

Instructor: Kristofer Wong

Teaching Assistants:

Alena Dickmann	Arya GJ	Finn Johnson
Joon Chong	Kimi Locke	Peyton Rapo
Rahul Misal	Winston Jodjana	

Announcements

- Ex7 & 8 due tonight!
- Ex9 (Communication Exercise) out tomorrow.
- Ex10, 11, 12, and the EC exercise also out tomorrow. You'll receive a repo for these, as they practice parallel programing
  - 9 and 10 due 7/30
  - 11 and 12 due 8/6
- P2 due Tuesday
- No lecture activity today.

Lecture Outline

- **Changing Another Major Assumption**
  - Definitions: Parallelism vs Concurrency
- Shared Memory with Threads
- Concurrency Frameworks in Java
  - Introducing java.lang.Thread
  - Writing *good* parallel code

Sequential Programming: A Major Assumption

- So far, most / all of your study has assumed:

One thing happened at a time

- This is **sequential programming**: everything in one sequence

- Removing this assumption creates major challenges & opportunities
  - *Programming*: How to divide work among **threads of execution** and coordinate (**synchronize**) among them
  - *Algorithms*: How to utilize parallel activity to gain speed
    - More **throughput**: work done per unit time
  - *Data structures*: May need to support **concurrent access**
    - ie, multiple threads operating on data at the same time

A Simplified View of Computing History

- Writing *correct* and *efficient* multithreaded code is much more difficult than for sequential code
  - Especially in common languages like Java and C
- Roughly 1980-2005, computers got exponentially faster
  - Sequentially-written programs doubled in speed every couple years
  - So there was little motivation to write non-sequential code
- But nobody knows how to continue making computers faster
  - Increasing clock rate generates too much heat
  - Relative cost of memory access is too high
- But we *can* continue "making wires exponentially smaller" (**"Moore's 'Law'"**)
  - Result: multiple processors on the same chip ("**multicore**")

What to do with Multiple Processors/Cores?

- Next computer you buy will likely have 4 cores
  - Wait a few years and it will be 8, 16, 32, …
  - The chip companies have decided to do this (not a "law")

- What can you do with these processors?
  - Run multiple, totally different, programs at the same time
    - Already do that? It certainly *appears* that way, thanks to **time-slicing**
  - Run multiple, possibly different, tasks at the same time in one single program
    - Our focus for the next few lectures; it's more difficult!
    - Requires rethinking everything from asymptotic complexity to how to implement data-structure operations

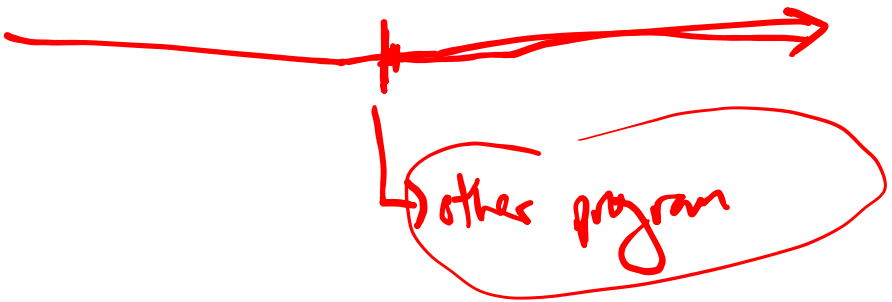
Lecture Outline

- Changing Another Major Assumption
  - **Definitions: Parallelism vs Concurrency**
- Shared Memory with Threads
- Concurrency Frameworks in Java
  - Introducing java.lang.Thread
  - Writing *good* parallel code

High Level Definitions:

- **Parallelism**: Using extra executors to solve a problem faster

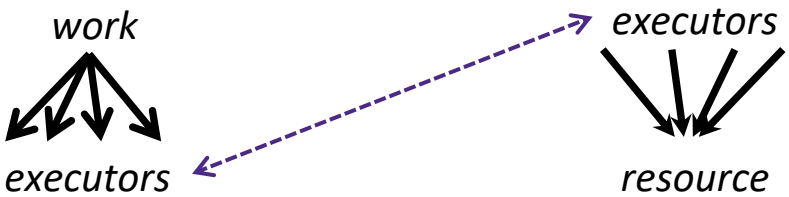
- **Concurrency**: *Correctly* and *efficiently* manage access to shared resources, from multiple possibly-simultaneous clients



Parallelism vs. Concurrency

Parallelism: Use extra executors to solve a problem faster

Concurrency: Manage access to shared resources



- There is some connection (confusion!) between them:
  - We commonly use threads for both parallelism and concurrency
  - If parallel computations access shared resources, the concurrency needs to be managed

Parallelism vs Concurrency: An Analogy

- **Sequential:** A program is like a cook making dinner
  - *One cook*: Makes gravy and stuffing one at a time!

- **Parallelism**: *"Extra executors gets the job done faster!"*
  - *Multiple cooks*: One cook in charge of the gravy (and its onions), another in charge of the stuffing (and its onions)
    - Increase throughput via simultaneous execution!
    - Too many cooks means you spend all your time coordinating

- **Concurrency**: *"We need to manage a shared resource"*
  - *Multiple cooks*: One cook per dish, but only one cutting board
    - Correctness: Don't want spills or ingredient mixing
    - Efficiency: Who should use the boards and in what order?

Lecture Outline

- Changing Another Major Assumption
  - Definitions: Parallelism vs Concurrency
- **Shared Memory with Threads**
- Concurrency Frameworks in Java
  - Introducing java.lang.Thread
  - Writing *good* parallel code

Our Model: Shared Memory with Threads

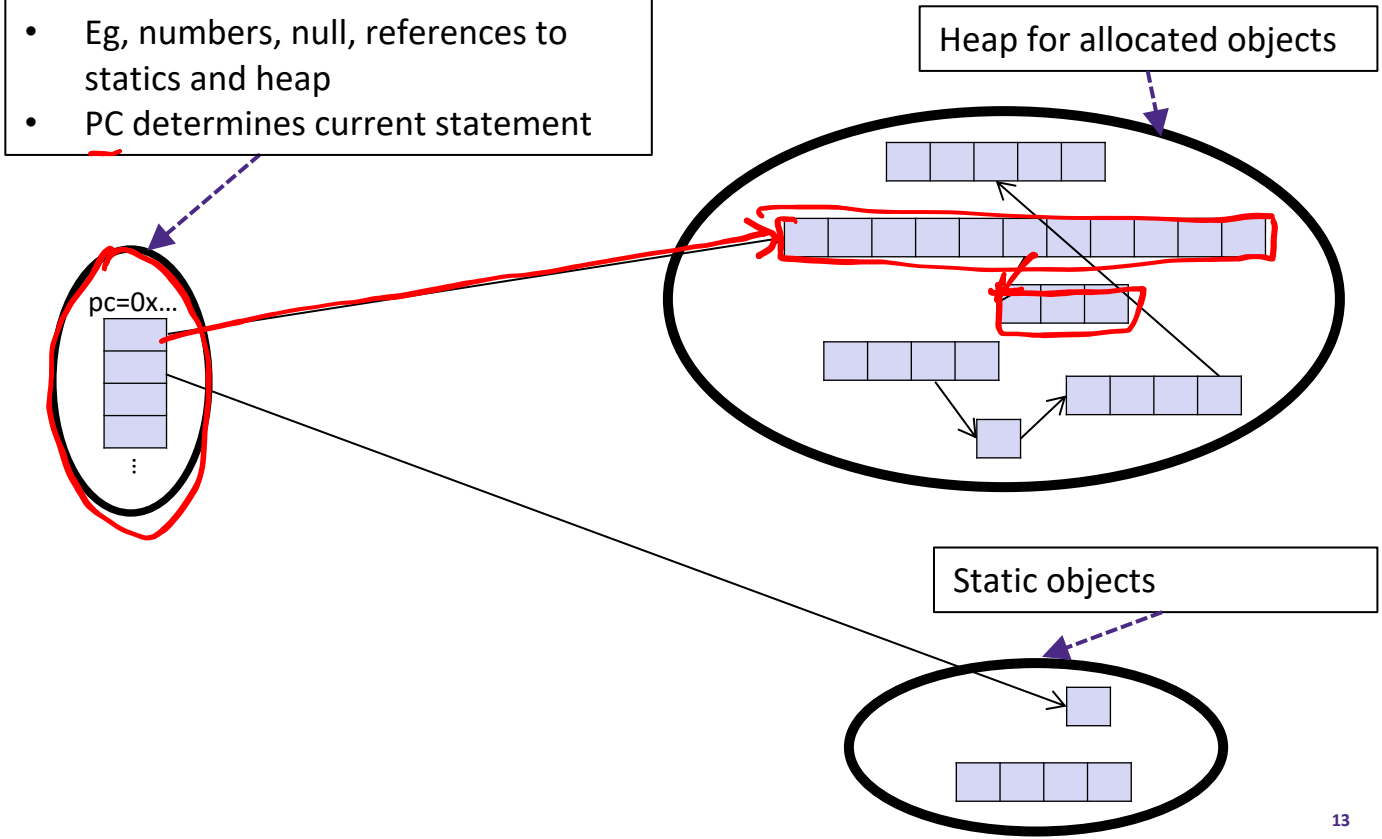
- We will assume **shared memory** with **explicit threads**

- *Sequential*: A running program has
  - One ***program counter*** ("PC"): currently executing statement
  - One ***call stack,*** with each stack frame holding its local variables
  - ***Objects in the heap*** created by memory allocation (i.e., new)
  - ***Static fields*** that are "global" to the entire program

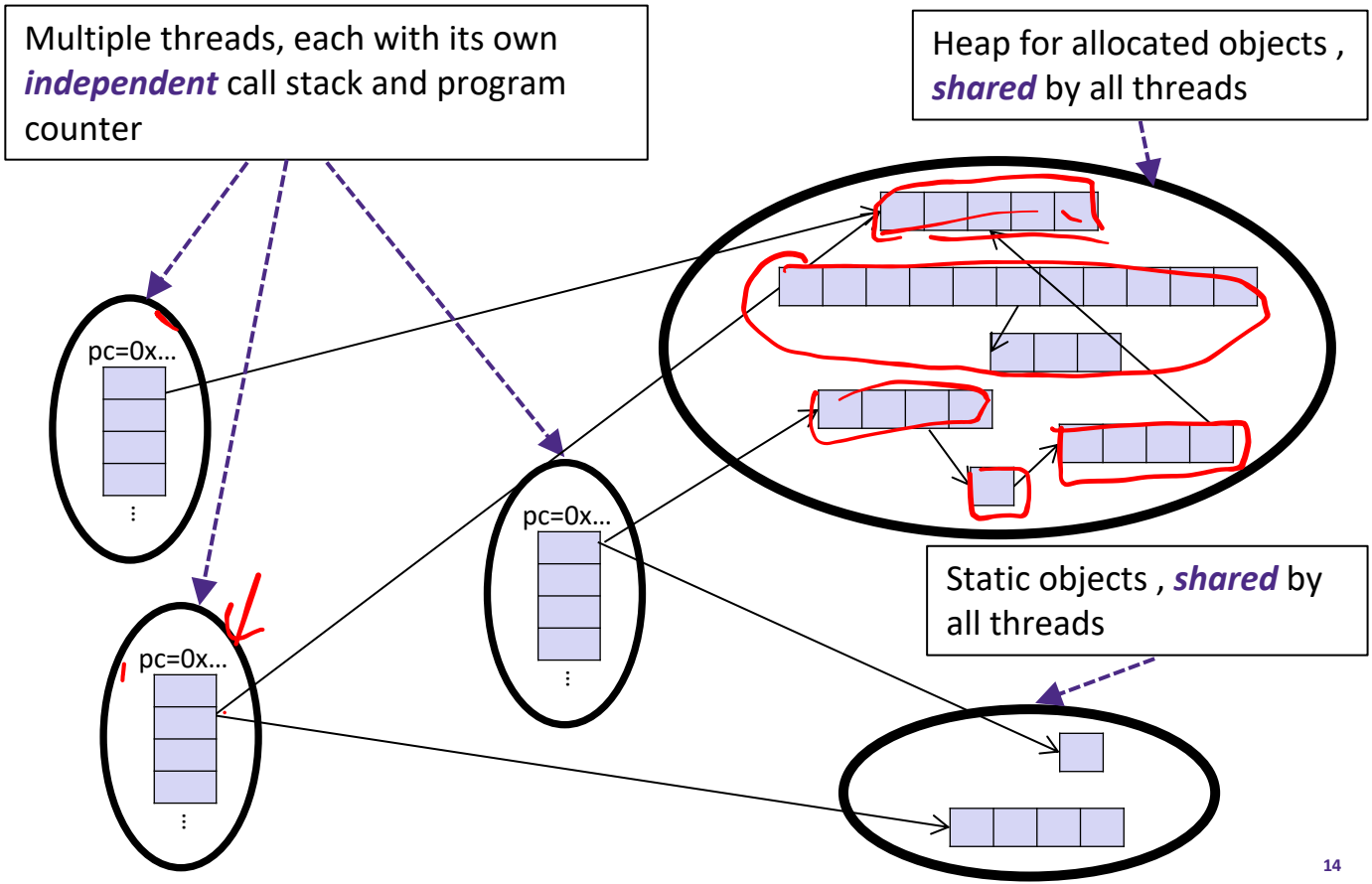
- *Shared Memory with Threads*: A running program has
  - A set of ***threads***, each with its own *program counter* and *call stack*
    - But each thread cannot access to another thread's local variables
  - Threads implicitly share *static fields* and the *heap* (ie, objects)
    - Communication via writing values to some shared location

Sequential: One Call Stack and One PC

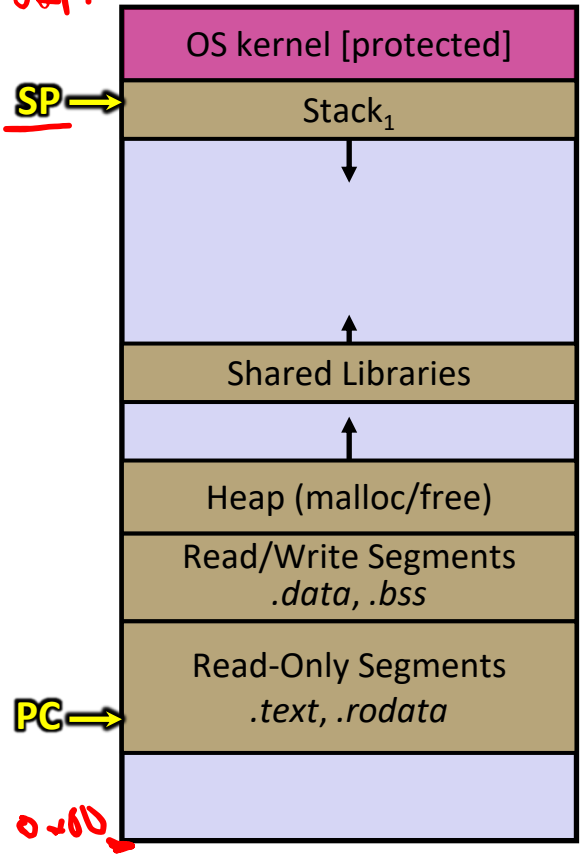
- Call stack with local variables
- Eg, numbers, null, references to statics and heap
- PC determines current statement

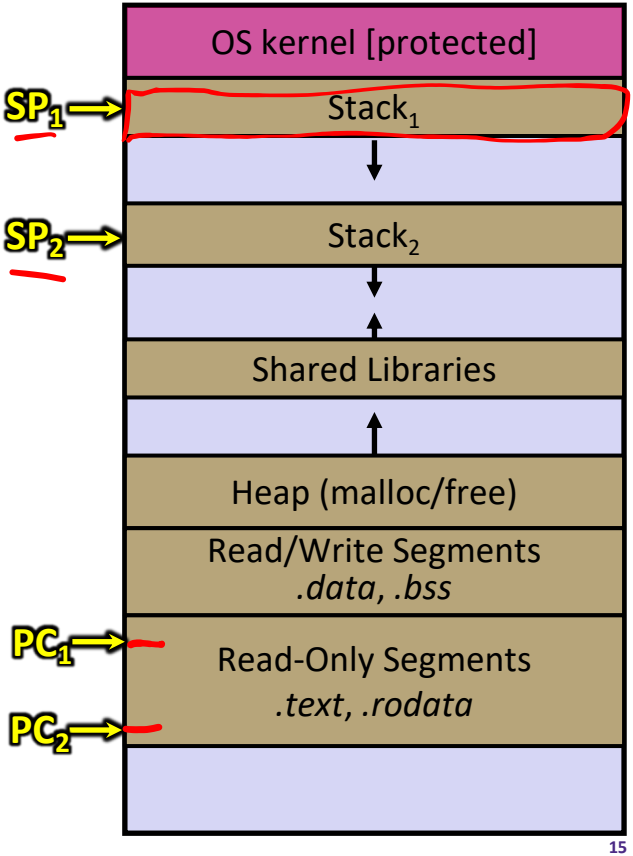


Shared Memory with Threads



Shared Memory with Threads *(if you've taken 351)*





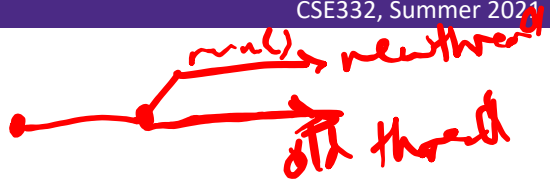
Lecture Outline

- Changing Another Major Assumption
  - Definitions: Parallelism vs Concurrency
- Shared Memory with Threads
- Concurrency Frameworks in Java
  - **Introducing java.lang.Thread**
  - Writing *good* parallel code

Our Requirements

- To write a shared-memory parallel program, we need new primitives from our *programming language* or a *library*
  - Ways to create and ***execute multiple things at once***
    - i.e. the parallel threads themselves!
  - Ways for threads to ***share memory*** or retain sole ownership
    - Often: just have threads contain references to the same objects
    - How will we pass thread-specific arguments to it? Does the thread have its own "private" (i.e., local) memory?
  - Ways for threads to ***coordinate*** (a.k.a. synchronize)
    - For now, all we need is a way for one thread to wait for another to finish
    - (we'll study other primitives when we get to concurrency)

Introducing java.lang.Thread



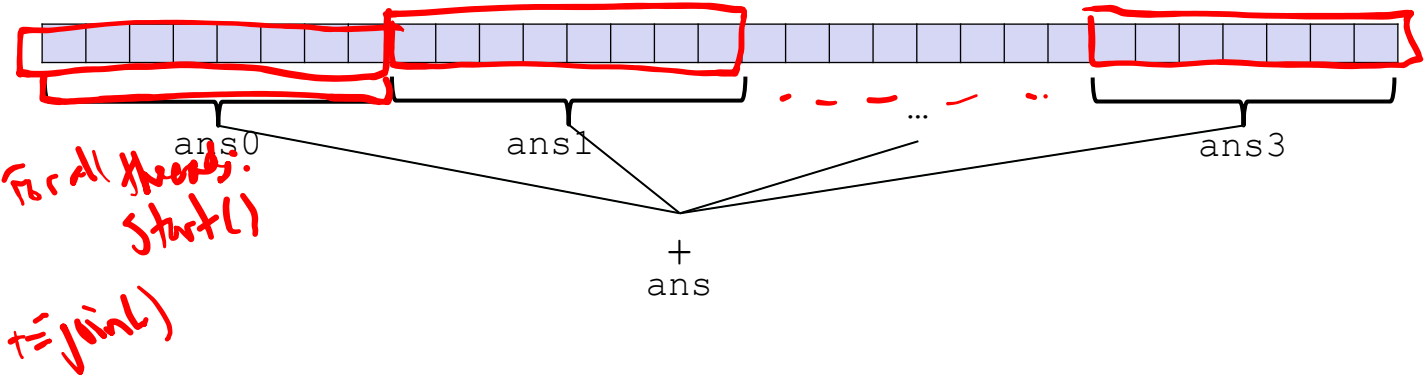
- First, we'll learn basic multithreading with java.lang.Thread
  - Then we'll discuss a different library (used in p3): ForkJoin

- To get a new thread to start executing something:
  1. Define a subclass C of `java.lang.Thread`, and override its `run()` method
  2. Create an instance of class `C`
  3. Call that object's `start()` method
     - `start()` creates a new thread and executes `run()` as its "main"

- What if we called C's `run()` method instead?
  - Normal method call executed in the current thread

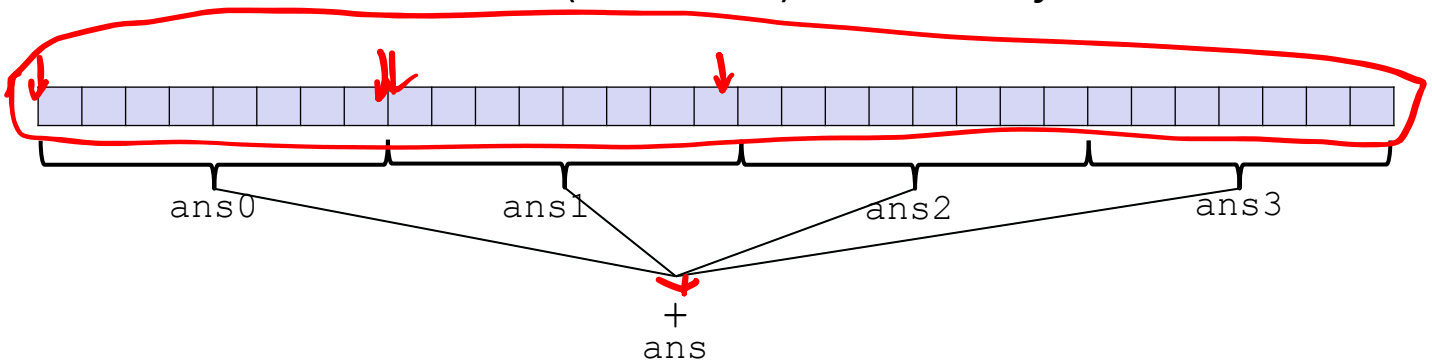
Example: Summing a Large Array

- *Problem*: Sum all the elements of a very large array
- *Idea*: Have n threads simultaneously sum a portion of the array
  - Create n ***thread objects***, each given a portion of the work
  - Call `start()` on each object to actually ***execute*** it in parallel
  - ***Wait*** for each thread to finish using its `join()` method
  - Combine their answers (via addition) to obtain the ***final result***



Attempt #1: Summing a Large Array

- (Warning: this is an inferior first approach)
- Have 4 threads simultaneously sum a portion of the array
  - Create 4 ***thread objects***, each given a 1/4 of the work
  - Call **`start()`** on each object to actually ***execute*** it in parallel
  - ***Wait*** for each thread to finish using its **`join()`** method
  - Combine their answers (via addition) to obtain the ***final result***



Attempt #1: Code (1 of 2)

```
class SumThread extends java.lang.Thread {
  // We pass arguments to the SumThread instance via
  // member fields that are initialized in the constructor
  int lo;          // input; start index
  int hi;          // input; end index, exclusive
  int[] arr;       // input; the (shared) array

  int ans = 0;     // output; the final sum

  SumThread(int[] a, int l, int h) { lo=l; hi=h; arr=a; }

  @Override
  public void run() { // must have this exact signature
    for (int i=lo; i < hi; i++)
      ans += arr[i];
  }
}
```

- Because we override a no-arguments/no-result `run`, we use member fields to communicate across threads

Attempt #1: Code (2 of 2)

```
class SumThread extends java.lang.Thread {
  int lo, int hi, int[] arr; // input: arguments
  int ans = 0;               // output: result
  SumThread(int[] a, int l, int h) { … }
  public void run(){ … }     // override: implement "main"
}
```

```
int sum(int[] arr){           // can be a static method
  int len = arr.length;
  int ans = 0;
  SumThread[] ts = new SumThread[4];
  for (int i=0; i < 4; i++) { // do parallel computations
    ts[i] = new SumThread(arr,i*len/4,(i+1)*len/4);
}
  for (int i=0; i < 4; i++) { // combine partial results
    ans += ts[i].ans;
  }
  return ans;
}
```

Attempt #2: Code

```
class SumThread extends java.lang.Thread {
  int lo, int hi, int[] arr; // input: arguments
  int ans = 0;               // output: result
  SumThread(int[] a, int l, int h) { … }
  public void run(){ … }     // override: implement "main"
}
```

```
int sum(int[] arr){           // can be a static method
  int len = arr.length;
  int ans = 0;
  SumThread[] ts = new SumThread[4];
  for (int i=0; i < 4; i++) { // do parallel computations
    ts[i] = new SumThread(arr,i*len/4,(i+1)*len/4);
    ts[i].start();            // call start(), not run!!!
  }
  for (int i=0; i < 4; i++) { // combine partial results
    ans += ts[i].ans;
  }
  return ans;
}
```

Attempt #3: Code

```
class SumThread extends java.lang.Thread {
  int lo, int hi, int[] arr; // input: arguments
  int ans = 0;               // output: result
  SumThread(int[] a, int l, int h) { … }
  public void run(){ … }     // override: implement "main"
}
```

```
int sum(int[] arr){           // can be a static method
  int len = arr.length;
  int ans = 0;
  SumThread[] ts = new SumThread[4];
  for (int i=0; i < 4; i++) { // do parallel computations
    ts[i] = new SumThread(arr,i*len/4,(i+1)*len/4);
    ts[i].start();            // call start(), not run!!!
  }
  for (int i=0; i < 4; i++) { // combine partial results
    ts[i].join();             // wait for thread to finish
    ans += ts[i].ans;
  }
  return ans;
}
```

join(): Our "wait" method for Threads

- Framework implements functionality you couldn't on your own
  - E.g.: **`start`**, which creates a new thread
- You "fill in the blanks" for the framework
  - E.g.: we implement **`run()`**, telling Java what to do in the thread
- Something else you can't implement: thread coordination
  - So it also provides the **`join()`** method!
  - **`join()`** blocks the caller until/unless the thread instance is done executing (i.e.: the call to **`run()`** finishes)
  - If it didn't, we would have a ***race condition*** on **`ts[i].ans`**

Incidentally …

- This code has a compile error because **`join`** may throw **`java.lang.InterruptedException`**
  - In basic parallel code, should be fine to catch-and-exit

```
int sum(int[] arr){              // can be a static method
  int len = arr.length;
  int ans = 0;
  SumThread[] ts = new SumThread[4];
  for (int i=0; i < 4; i++) { // do parallel computations
    ts[i] = new SumThread(arr,i*len/4,(i+1)*len/4);
    ts[i].start();             // call start(), not run!!!
  }
  for (int i=0; i < 4; i++) { // combine partial results
    ts[i].join();              // wait for thread to finish
    ans += ts[i].ans;
  return ans;
}
```

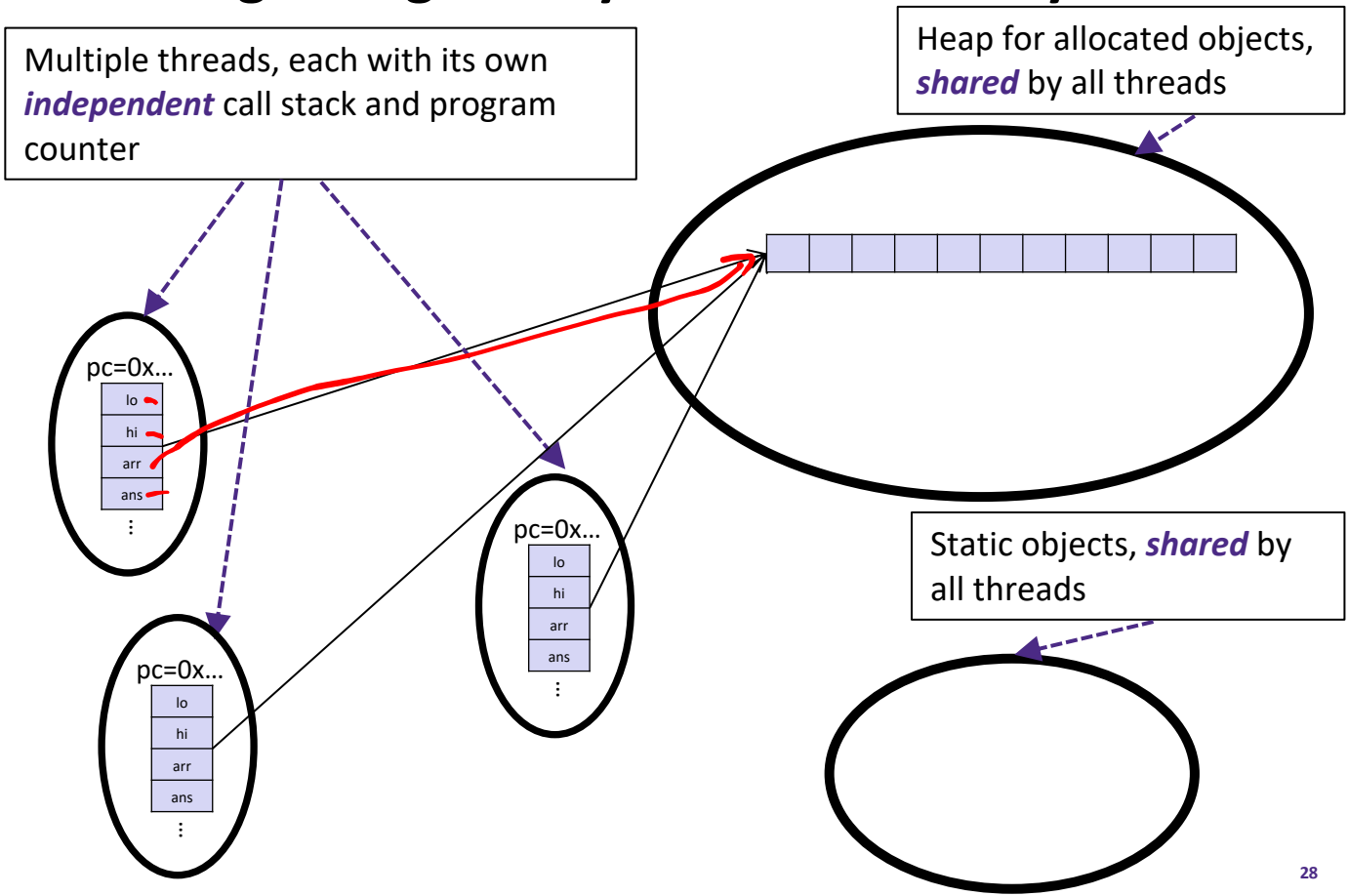
Where is the Shared Memory? Local Memory?

- Our program (implicitly!) shares memory
  - **`lo`** & **`hi`** are inputs: written by "main" thread, read by helper thread
  - **`arr`** reference also an input, but its referred array was shared
  - **`ans`** is an output: written by helper thread, read by "main" thread

- Our program didn't have thread-local memory
  - But you could imagine `SumThread` containing a local variable (maybe a loop counter?) which it doesn't share with other threads

- When using shared memory, you must avoid race conditions
  - While studying parallelism (now), we'll stick with **`join`**
  - With concurrency (later), we will learn other ways to synchronize

Summing a Large Array: Shared Memory



Lecture Outline

- Changing Another Major Assumption
  - Definitions: Parallelism vs Concurrency
- Shared Memory with Threads
- Concurrency Frameworks in Java
  - Introducing java.lang.Thread
  - **Writing *good* parallel code**

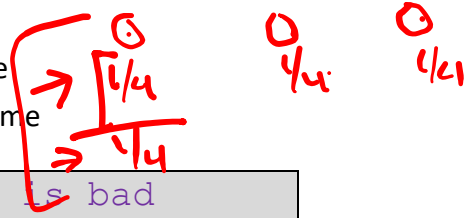
Issues with Our Earlier Approach (1 of 3)

1. Want code to be portable and efficient across platforms
   - So at the *very very* least, parameterize by the number of threads

```
int sum(int[] arr, int numTs){
  int len = arr.length;
  int chunkLen = arr.length/numTs;
  int ans = 0;
  SumThread[] ts = new SumThread[numTs];
  for(int i=0; i < numTs; i++) {
    ts[i] = new SumThread(arr, i*chunkLen,(i+1)*chunkLen);
    ts[i].start();
  }
  for(int i=0; i < numTs; i++) {
    ts[i].join();
    ans += ts[i].ans;
  }
  return ans;
}
```

Issues with Our Earlier Approach (2 of 3)

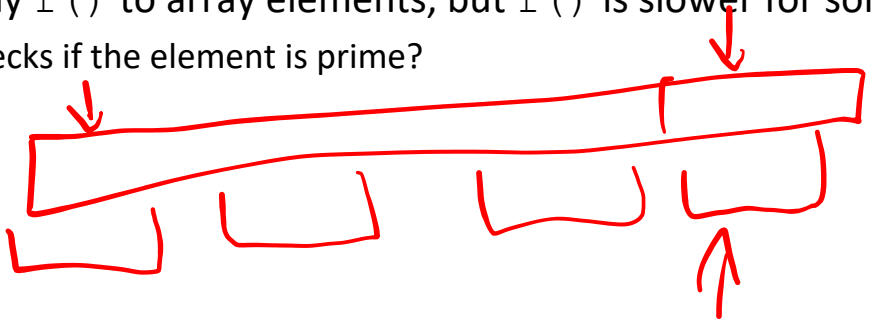
2. Want to use only processors "available to you now"
   - Processors used by other programs or threads aren't available!
     - Maybe caller is also using parallelism?
     - Number of available cores changes even while your threads run
   - E.g.: if you have 3 available processors and using 3 threads would take time **x**, then creating 4 threads would take time **1.5x**
     - Example: 12 units of work, 3 processors
       - Dividing work into 3 chunks will take 4 units of time
       - Dividing work into 4 chunks will take 3*2 units of time



```
// numThreads == numProcessors is bad
// if some are needed for other things
int sum(int[] arr, int numTs){
  …
}
```

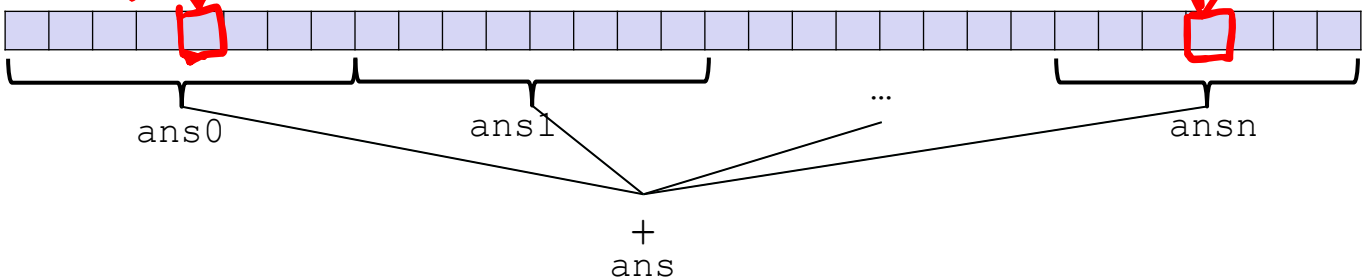
Issues with Our Earlier Approach (3 of 3)

3. In general, subproblems take different amounts of time
   - Sometimes drastically different!
   - If we create 100 threads but one chunk takes much much longer, we won't get a ~100x speedup
     - This is called a ***load imbalance***
   - E.g.: apply `f()` to array elements, but `f()` is slower for some elts
     - `f()` checks if the element is prime?



A Better Approach: Smaller Chunks

- The solution: *cut up our problem into many small chunks*
  - We want far more chunks than the number of processors!
  - … but this will require changing our algorithm



1. *Portable?* Yes! (Substantially) more chunks than processors
2. *Adapts to Available Processors?* Yes! Hand out chunks as you go
3. *Load Balanced?* Yes(ish)! Variation is smaller if chunks are small

A Better Approach: Abandoning java.lang.Thread

- For this specific problem (and for p3), the constants for Java's built-in thread framework are not great
- Plus, there's complexity in Java's Thread framework that confuse rather than illuminate

Lecture Outline

- Concurrency Frameworks in Java
  - Improving java.lang.Thread
    - **Asymptotically**
    - Constants
  - ForkJoin Library

Naïve Thread Creation/Joining Algorithm

- Suppose we create 1 thread to process every 1000 elements

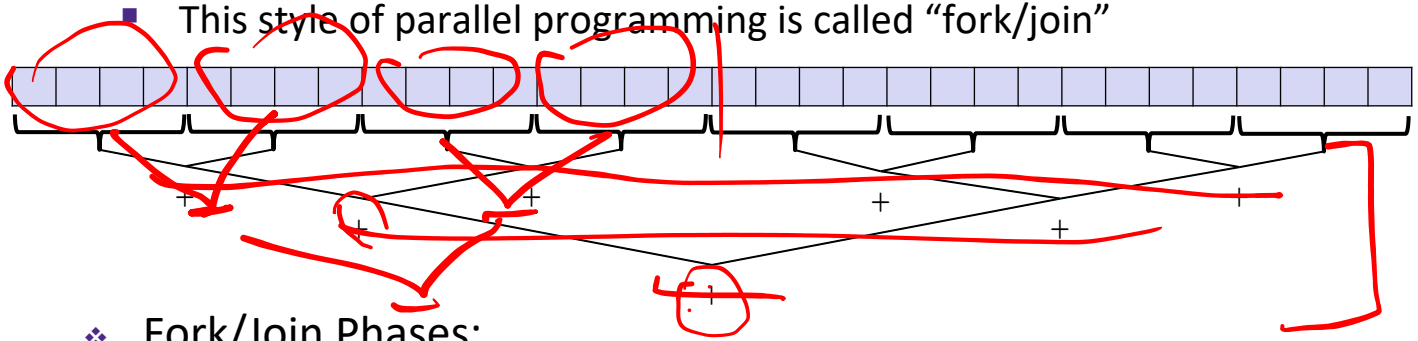
```
int sum(int[] arr){
  …
  int numThreads = arr.length / 1000;
  SumThread[] ts = new SumThread[numThreads];
  …
}
```

- "Combine results" part has `arr.length/1000` additions
  - Θ(N) to combine!
  - Previously, we had only 4 pieces (Θ(1) to combine)

→ Will a Θ(N) algorithm to create threads/combine results be a bottleneck?

Smarter Thread Creation/Joining: Divide and Conquer!

- Divide and Conquer:
  - "Grows" the number of threads to fit the problem
  - Uses parallelism for the recursive calls
  - This style of parallel programming is called "fork/join"



- Fork/Join Phases:

1. Divide the problem
   - Start with full problem at root
   - Make two new threads, halving the problem, until size is at cutoff
2. Combine answers as we return from recursion

Fork/Join-style Parallelism (1 of 2)

```
class SumThread extends java.lang.Thread {
  // … member fields and constructors elided …
  public void run() {  // override: implement "main"
    if(hi – lo < SEQUENTIAL_CUTOFF) {
      // Just do the calculation in this thread
      for (int i=lo; i < hi; i++)
        ans += arr[i];
    }
    else {
      // Create two new threads to calculate the left and right sums
      SumThread left = new SumThread(arr, lo, (hi+lo)/2);
      SumThread right= new SumThread(arr, (hi+lo)/2, hi);
      left.start();
      right.start();

      // Combine their results
      left.join();  // don't move this up a line (why?)
      right.join();
      ans = left.ans + right.ans;
    }
  }
}
```

Fork/Join-style Parallelism (2 of 2)

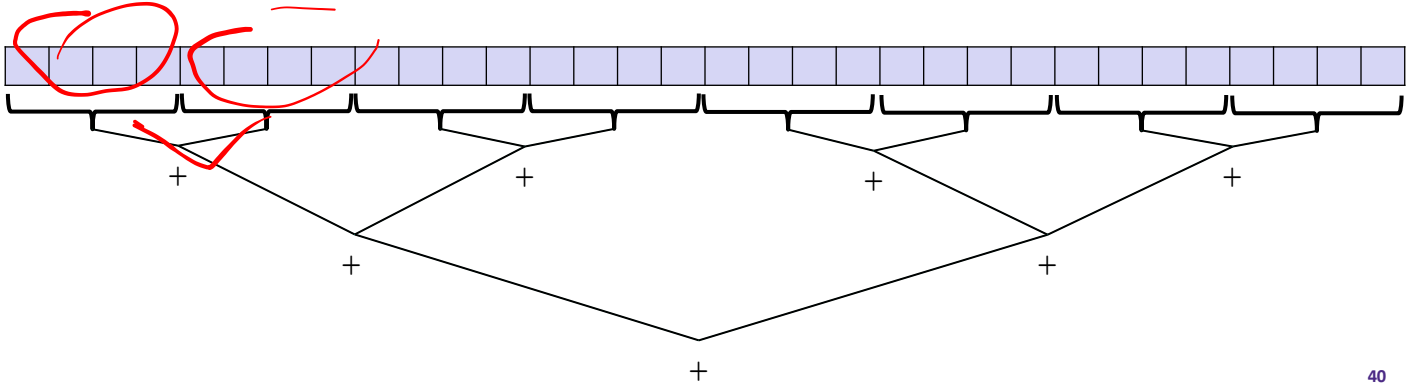
```
class SumThread extends java.lang.Thread {
  int lo, int hi, int[] arr; // input: arguments
  int ans = 0;               // output: result
  SumThread(int[] a, int l, int h) { … }
  public void run(){ … }     // override: implement "main"
}
```

```
int sum(int[] arr) {
  SumThread t = new SumThread(arr, 0, arr.length);// just 1 obj since
  t.run();                                        // we don't need
  return t.ans;                                   // parallelism to
}                                                 // start recursion
```

- The computation and the result-combining are both in parallel
  - Using recursive divide-and-conquer makes this natural
  - Easier to write *and* more efficient asymptotically!

Fork/Join-style Parallelism Really Works!

- The key is in parallelizing both the executor-creation and the result-combining phases
  - If enough processors, runtime is **height of the tree:** $O(\log n)$
    - Optimal and exponentially faster than sequential $O(n)$
  - Relies on operations being associative (like +)
- We'll write all our parallel algorithms in this style
  - But using a special library engineered for this style



Being Pragmatic #1: Performance Tuning

- Wait until computer has more processors ;)
  - Communication overhead may still dominate at 4 processors, but this configuration is rare for servers (circa 2020)
    - attu6 has 4 CPUs with 14 cores each = 56 "processors"
- Beware memory-hierarchy issues!
  - Won't focus on this, but crucial for parallel performance

Lecture Outline

- Concurrency Frameworks in Java
  - Improving java.lang.Thread
    - Asymptotically
    - **Constants**
  - ForkJoin Library

gradescope

gradescope.com/courses/275833

- Assume that thread creation and joining are expensive. Which of the following optimizations might improve our constants?
  1. Use a cutoff, after which computation proceeds sequentially
  2. Somehow create fewer threads
  3. Somehow reuse threads when they're done

A. Cutoff only
B. Cutoff + Fewer Threads
C. Cutoff + Thread Reuse
D. Cutoff + Fewer Threads + Thread Reuse
E. I'm not sure …

Being Pragmatic #2: Constants Matter

- *In theory*, can divide down to single elements, do all the result-combining in parallel, and get optimal speedup
  - Total time: O(n / numProcessors + log n)

- *In practice*, thread creation/joins eat into the savings, so:

1. Use a cutoff, after which computation proceeds sequentially
   - Cutoff value depends on type of computation; 500-1000 is a good start
   - Eliminates *almost all* the recursive thread creation (bottom levels of tree)
   - *Exactly* like QuickSort switching to InsertionSort, but more important here
2. Do not create *two* recursive threads; create one thread and do the other piece of work "yourself"
   - Halves the number of threads created (?!?!)

Halving the Created Threads: Code

- If the *language* had built-in support for fork/join-style parallelism, this hand-optimization would be unnecessary
- But the *library* we're using expects you to do it yourself
  - … and the difference is surprisingly substantial
- Again: no difference in theory, "only" the constants

run() is a nomal function call! Execution won't proceed until it completes

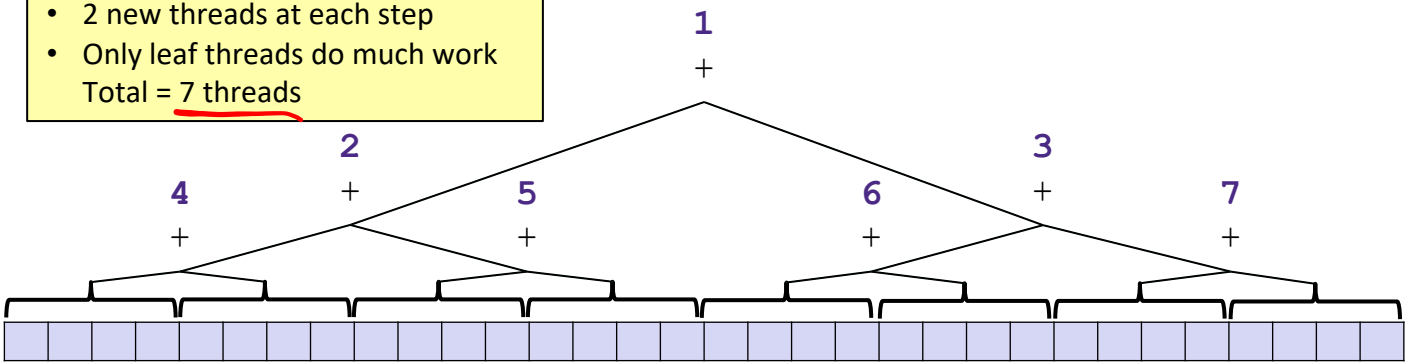
```
// Don't do this:
SumThread left  = …
SumThread right = …

left.start();
right.start();

left.join();
right.join();
ans = left.ans + right.ans;
```

```
// Do this instead:
SumThread left  = …
SumThread right = …

left.start();
right.run();

left.join();
// no right.join() needed
ans = left.ans + right.ans;
```

Halving the Created Threads: Pictorially

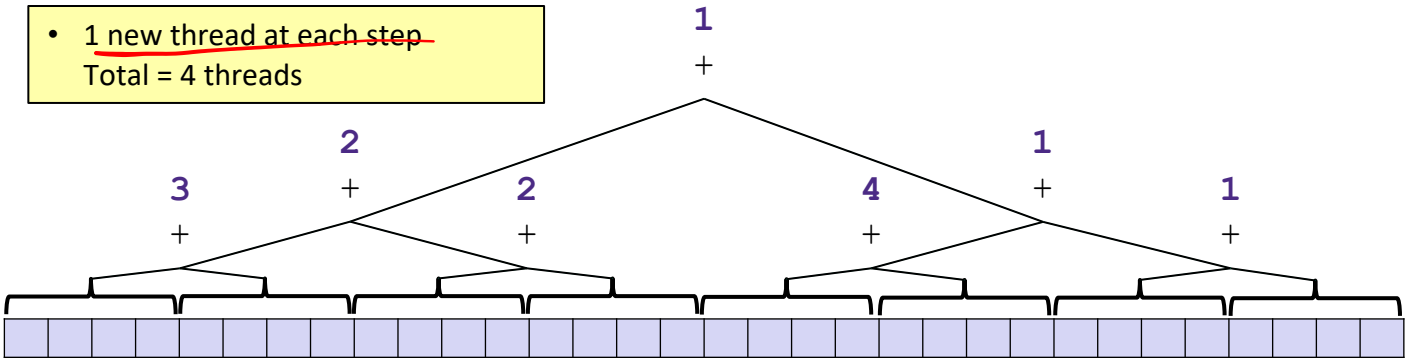
- 2 new threads at each step
- Only leaf threads do much work

Total = 7 threads



- 1 new thread at each step

Total = 4 threads



Lecture Outline

- Concurrency Frameworks in Java
  - Improving java.lang.Thread
    - Asymptotically
    - Constants
  - **ForkJoin Library**

Finally! The ForkJoin Library

- Even using fork/join-style code, java.lang.Thread is still too "heavyweight"
  - Constant factors, especially space overhead
  - Creating 20,000 Java threads just a bad idea ☹
- So use the **ForkJoin Library** instead
  - Introduced in Java 8 (2014)
  - Similar libraries available for other languages
    - C/C++: Cilk (inventors), Intel's Thread Building Blocks
    - C#: Task Parallel Library
    - …
  - Its implementation is a fascinating but advanced topic

Thread -> ForkJoin: Terminology

Java Built-in Threads	ForkJoin Library
Subclass `Thread`	Subclass `RecursiveTask<V>`
Override `run()`	Override `compute()`
Call `start()` to begin parallel computation	Call `fork()` to begin parallel computation
Return results via member fields (eg, `ans`)	Return results via return value (ie, an instance of `V`)
Call `join()`, then check its "returned" member field	Call `join()`, then check its return value
Halve created threads by calling `run()` directly	Halve created threads by calling `compute()` directly
Begin recursion with top-level call to `run()` (instead of `start()`)	Begin recursion by creating a `ForkJoinPool` and calling its `invoke()`

Fork/Join-style Parallelism with ForkJoin (1 of 2)

```
class SumTask extends RecursiveTask<Integer> {
  int lo; int hi; int[] arr;     // just the "input" arguments!

  protected Integer compute() {  // override: implement "main"
    if(hi – lo < SEQUENTIAL_CUTOFF) {
      // Just do the calculation in this thread
      int ans = 0;   // local variable instead of a member field
      for (int i=lo; i < hi; i++)
        ans += arr[i];
      return ans;    // direct return of answer
    } else {
      // Create ONE new thread to calculate the left sum
      SumTask left  = new SumTask(arr, lo, (hi+lo)/2);
      SumTask right = new SumTask(arr, (hi+lo)/2, hi);
      left.fork();   // create a thread and call its compute()
      int rightAns = right.compute(); // call compute() directly

      // Combine results
      int leftAns  = left.join();
      return leftAns + rightAns;
    }
  }
}
```

Fork/Join-style Parallelism with ForkJoin (2 of 2)

```
class SumTask extends RecursiveTask<Integer> {
  int lo; int hi; int[] arr;         // input: arguments
  SumTask(int[] a, int l, int h) { … }
  protected Integer compute() { … } // override: implement "main"
}
```

```
static final ForkJoinPool POOL = new ForkJoinPool();

int sum(int[] arr) {
  SumTask task = new SumTask(arr, 0, arr.length);

  // invoke() returns the value which is returned by the
  // top-level compute()
  return POOL.invoke(task);
}
```

Being Pragmatic #2: Performance Tuning the Library

- Sequential threshold
  - Library documentation recommends doing approximately 100-5000 basic operations in each "piece" of your algorithm
- ForkJoin library needs to "warm up"
  - May see slow results before JVM re-optimizes the library internals
  - Put computations in a loop to see the "long-term benefit"

Summary: Parallelism

- **Parallelism**: increasing efficiency/decreasing total runtime
- **Concurrency**: correctly accessing shared resources
  - They intersect when *parallel computations access shared resources*

- Model: shared memory with explicit threads:
  - **Threads** are the minimum fields necessary to represent "computation": a program counter and a stack
  - Everything else is **shared** (eg, static variables, heap)

- Threading:
  - `run()`/`compute()` are "regular" function calls, but `start()`/`fork()` create a new thread and then call `run()`/`compute()`
  - Parallelizing many small chunks of work is portable, adaptable, and load-balancable

Summary: Fork/Join Parallelism

- **Fork/Join Parallelism** is a *model* that grows the parallelism to fit the problem size using recursion

- The **ForkJoin library** that cleanly enables this model
  - You still need to manually specify the sequential cutoff
  - Halving the created threads also requires manual intervention