**Question 1:** (16 points)
Program P runs to completion in 100 seconds on a single, in-order processor. When Program P is parallelized, 20 seconds have to be spent in serial execution that cannot be parallelized, and the rest of execution time can be perfectly parallelized on any number of processors.

(1A) (6 points) When Program P runs on 16 processors, how much time would it take to run to completion?

Time = Serial time + (Parallel time/N) = 20 + 80 / 16 = 25 seconds

Another solution: speedup = 1/(S + P/N) = 1/(0.2 + 0.8 / 16) = 4
Time = original time / speedup = 100/4 = 25 seconds

(1B) (4 points) Assume we have a system with an infinite number of processors. How much time would it take to run Program P to completion?

When N = infinity, time = Serial time + (Parallel time / infinity) = serial time = 20 seconds

(1C) (6 points) If we have a fixed area on which we can build one of the following two systems:
- System 1 contains 16 in-order single-issue processors similar to the ones in part (1A)
- System 2 contains 4 out-of-order 4-wide superscalar processors. Program P runs to completion on one of these processors in 50 seconds. When parallelized, 4 seconds have to be spent in serial execution, and the rest can be perfectly parallelized.

Which of these two systems would you choose to run Program P?

System 1’s runtime = 25 seconds
System 2’s runtime = serial time + parallel time / number of processors = 4 + 46/4 = 15.5
System 2’s runtime is lower, so we choose System 2.
Question 2: (16 points)

(2A) (4 points) Identify one advantage and one disadvantage of UMA (compared to NUMA)

Advantage: Simpler to design
Disadvantage: Less scalable

(2B) (4 points) What are the characteristics of programs that perform better on a CMP compared to a superscalar processor?

Programs with high parallelism (e.g., IPC > 40) like most floating point applications
Programs with high thread-level parallelism but low instruction level parallelism
Parallel programs

(2C) (4 points) Identify one advantage and one disadvantage of a hyper-threading processor compared to a multiprocessor.

Advantage: Area and power efficiency
Disadvantage: Resource contention due to sharing of resources between multiple threads

(2D) (4 points) Identify one advantage and one disadvantage of the distributed memory architecture (compared to the shared memory architecture)

Advantage: More scalable
Disadvantage: Harder to program
**Question 3:** (16 points)

Consider the following program fragment that runs in parallel on a multiprocessor system that implements the MSI cache coherence protocol:

\[\begin{align*}
I1: & \quad \text{LOAD} \quad \text{R1, [A]} \quad // \text{load from memory address A to register R1} \\
I2: & \quad \text{LOAD} \quad \text{R2, [B]} \quad // \text{load from memory address B to register R2} \\
I3: & \quad \text{ADD} \quad \text{R3, R1, R2} \quad // \text{add the contents of R1 and R2, save result in R3} \\
I4: & \quad \text{STORE} \quad [A], \text{R3} \quad // \text{store R3 to memory address A} \\
I5: & \quad \text{SUB} \quad \text{R4, R4, 1} \quad // \text{Subtract 1 from R4} \\
I6: & \quad \text{BNZ} \quad I1 \quad // \text{return to I1 if R4 is not zero}
\end{align*}\]

(3A) (6 points) If A and B reside in two different cache blocks, complete the following table to identify the cache state of the blocks containing A and B after each instruction executes. Possible states are M, S, and I

\[
\begin{array}{|c|c|c|}
\hline
\text{Instruction} & \text{State for block “A”} & \text{State for block “B”} \\
\hline
\text{I1} & S & I \\
\hline
\text{I2} & S & S \\
\hline
\text{I3} & S & S \\
\hline
\text{I4} & M & S \\
\hline
\end{array}
\]

(3B) (10 points) Consider the following order of instructions on two processors P1 and P2:

P1: I1, P2: I1, P1: I2, P1: I3, P1: I4, P2: I2, P2: I3, P2: I4

Complete the following table to identify the cache state for the two blocks containing A and B in both processors’ caches after each instruction executes.

\[
\begin{array}{|c|c|c|c|c|}
\hline
\text{Instruction} & \text{P1’s cache state for “A”} & \text{P1’s cache state for “B”} & \text{P2’s cache state for “A”} & \text{P2’s cache state for “B”} \\
\hline
\text{Initial State} & I & I & I & I \\
\hline
\text{P1: I1} & S & I & I & I \\
\hline
\text{P2: I1} & S & I & S & I \\
\hline
\text{P1: I2} & S & S & S & I \\
\hline
\text{P1: I3} & S & S & S & I \\
\hline
\text{P1: I4} & M & S & I & I \\
\hline
\text{P2: I2} & M & S & I & S \\
\hline
\text{P2: I3} & M & S & I & S \\
\hline
\text{P2: I4} & I & S & M & S \\
\hline
\end{array}
\]
Question 4: (18 points)

(4A) (6 points) Identify one advantage and one disadvantage of write-invalidate cache coherence protocols (compare to write-update protocols).

Advantage: Less bandwidth used when multiple writes are performed to the same block by the same processor
Disadvantage: More bandwidth used when multiple writes are performed to the same block by more than one processor due to the ping-pong effect of invalidations

(4B) (6 points) Identify one advantage and one disadvantage of building an inclusive cache system (where an L2 cache has to contain all blocks in L1 caches of processors that share it).

Advantage: L2 cache handles coherence with other processors, so the L1 doesn’t have to snoop the bus or maintain directory information
Disadvantage: Total cache size is reduced, and L1 blocks need to be replaced if corresponding L2 block is evicted

(4C) (6 points) Identify one advantage and one disadvantage of using a split-transaction bus in a snooping-based cache coherence protocol.

Advantage: Less utilization of bus, and increase in effective bus bandwidth
Disadvantage: Intermediate states are needed in coherence protocol to address cases when request is sent and the cache is waiting for response
Question 5: (16 points)

(5A) (6 points) Using the terminology in the DASH paper, identify the owner of a memory block whose directory state is:

(i) Uncached-remote
    Home cluster
(ii) Shared-remote
    Home cluster
(iii) Dirty-remote
    Remote cluster

(5B) (4 points) In the DASH protocol, why does the directory need to keep track of all sharers of a memory block?

To avoid broadcasts, the directory only sends invalidate requests to sharers of a block on a remote write

(5C) (6 points) The SGI Origin coherence protocol added the Exclusive (i.e., “E”) cache state to the MSI cache coherence protocol implemented in the DASH prototype. Identify two reasons why that change can improve performance.

1. Single fetch from memory is needed for read-modify-write requests (saves bandwidth and reduces latency)
2. Permits processor to replace a block in the E state without informing the directory to save bandwidth
3. A request from a processor that replaced an E block can be immediately satisfied from memory without communicating with other processors (saves bandwidth and latency)
Question 6: (18 points)

(6A) (6 points) Identify two reasons why using IPC may be misleading when comparing performance of multiprocessor systems.

An architectural enhancement can lead to a completely different execution path with a different number of instructions due to operating system scheduling or different orders of lock acquisition. Programs with idle instruction and spin locks may execute instructions that do not contribute to useful work.

(6B) (8 points) In the Data Diffusion Machine (DDM) implementation of COMA, list the sequence of cache coherence states and transactions when a processor needs to read and then write a block in the “I” state, and then another processor needs to write to that same block. Complete the following table with all the state information:

<table>
<thead>
<tr>
<th>Current State</th>
<th>Requests and/or Responses</th>
<th>New State</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Pread/Nread</td>
<td>R</td>
</tr>
<tr>
<td>R</td>
<td>Nread</td>
<td>R</td>
</tr>
<tr>
<td>R</td>
<td>Ndata</td>
<td>S</td>
</tr>
<tr>
<td>S</td>
<td>Pwrite/Nerase</td>
<td>W</td>
</tr>
<tr>
<td>W</td>
<td>Nexclusive</td>
<td>E</td>
</tr>
<tr>
<td>E</td>
<td>Pwrite</td>
<td>E</td>
</tr>
<tr>
<td>E</td>
<td>Pread</td>
<td>E</td>
</tr>
<tr>
<td>E</td>
<td>Nread/Ndata</td>
<td>S</td>
</tr>
<tr>
<td>S</td>
<td>Nerase</td>
<td>I</td>
</tr>
</tbody>
</table>

(6C) (4 points) Identify one disadvantage of COMA architectures compared to cc-NUMA architectures.

Overhead of tags and state that is not needed for ccNUMA

Coherence protocol is more complicated since we need to make sure last copy of a replaced block is stored somewhere in the system.
Name:

**Scratch Pad** (Not graded)