Universidad
Carlos III de Madrid

ARCOS

Computer Architecture

Group 89

Midterm exam

Name:____

NIA: ____________

1. (3 points) Given the following parallel code:

Processor 1	Processor 2
(1a) print x	(2a) x = 1
(1b) x = 2	(2b) print x

And assuming that initially x = 0 and print represents a read instruction, answer the following questions

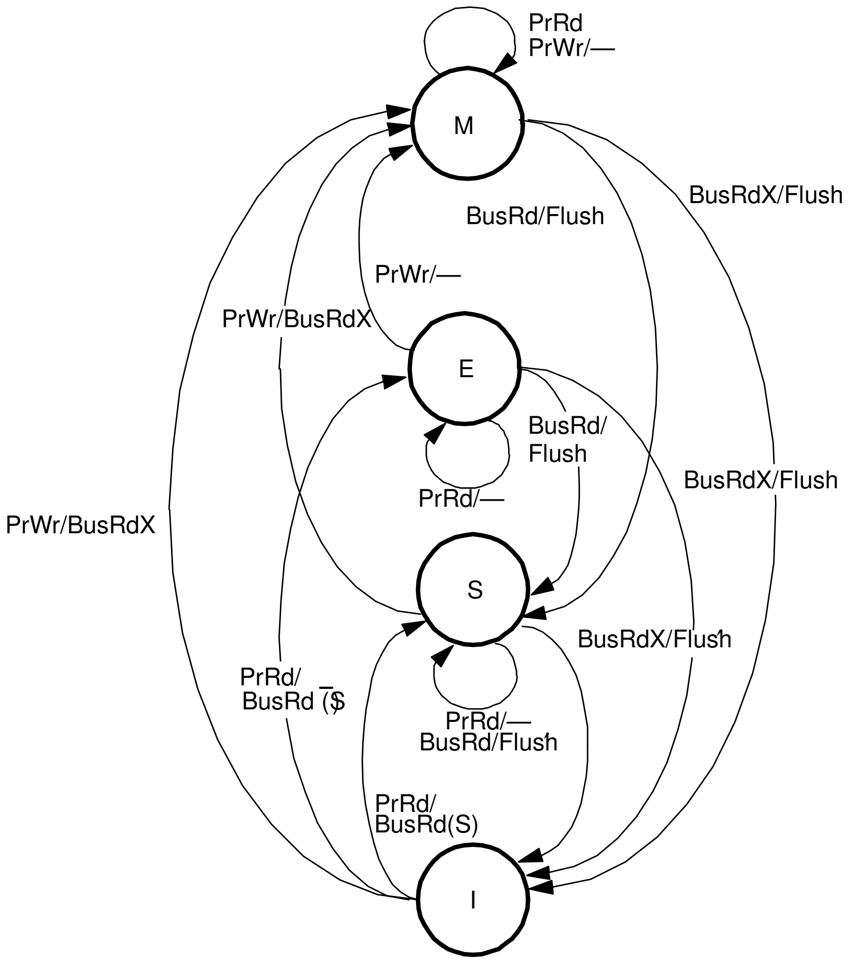
(a) Identify the dependencies in the code.
Answer: 1a-> 1b anti-dependence, 2a->2b data dependence

(b) What values will print the program under sequential consistency model? Justify your answer.
Answer: (0,1), (0,2),(1,1), (1,2)

(c) Do the dependencies identified in (a) influence the outcome of the program under the sequential consistency model?
Answer: No, the dependencies play no role, as sequential consistency model assumes no reordering.

(d) The code executes on a cache-coherent ***shared memory*** machine, employing MESI as a cache coherence protocol. Indicate in the following table the states, transitions and bus transactions of the system for the following instruction stream: 1 a, 2a, 1b, 2b.

	P1 transition	P2 transition	Bus actions
P1: print x	I->E	I	PrRd/BusRd(-S)
P2: x=1	E->I	I->M	PrWr/BusRdX BusRdX/Flush
P1:x=2	I->M	M->I	PrWr/BusRdX BusRdX/Flush
P2: print x	I->S	I->S	PrRd/BusRd(S) BusRd/Flush



(e) Assume that the code executes on a *cache-coherent distributed memory* architecture employing for cache coherence a *flat memory-based directory* scheme. The variable x is stored in the memory bank of P2 and x is not stored in any cache initially. Indicate in the following table the states of the system for the following instruction stream: 1 a, 2a, 1b, 2b.

	C1		C2		M1		M2	
	Dirty	Valid	Dirty	Valid	Dirty	Valid	Dirty	Valid
Initial	0	0	0	0	-	-	0	00
P1: print x		1						10
P2: x=1		0	1	1			1	01
P1: x = 2	1	1	0	0			1	10
P2:print x		1		1			0	11

2. A computer architecture has 16 interconnected compute nodes with the following characteristics:

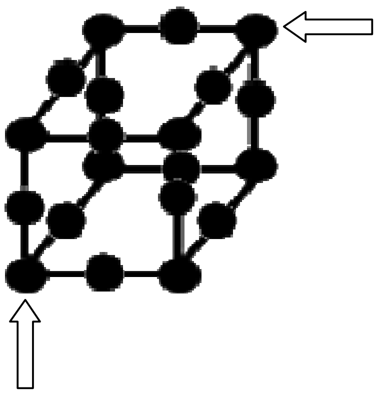
- Each node has a processor, memory and a Network Intelligent Card (NIC).
- The routing protocol is *store and forward*.
- The *routing delay* is 1 ms and the network bandwidth is 1GBit/second.
- The sending and the receiving overhead is 0.1 ms per operation.
- There is no contention in the network or at the compute nodes.

a) Make one drawing of the 16 node architecture for each of the following topologies: 3D grid, 3D torus and hypercube.

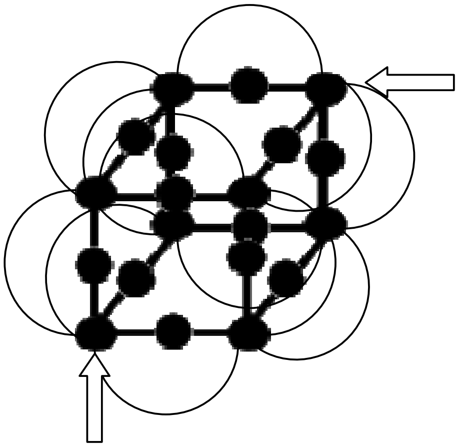
b) For each topology from a), calculate the *maximum* transfer time of a message of 1 Mbit between a pair of nodes, i.e. nodes which are farthest apart in each case.

Solution:

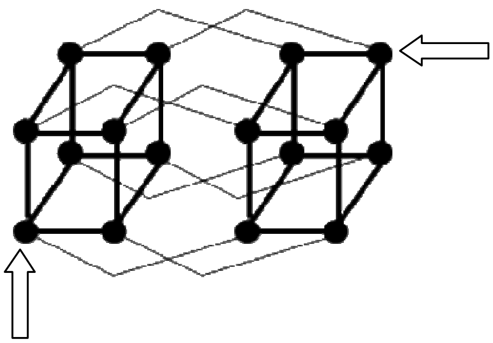
a) 3D grid



3D torus



Hypercube



b)

- **3D grid**
  T=0.1ms+6*(1ms + 1MBit/1Gbit/s)+0.1ms
- **3D torus**
  T=0.1ms+3*(1ms + 1MBit/1Gbit/s)+0.1ms
- **Hypercube**
  T=0.1ms+4*(1ms + 1MBit/1Gbit/s)+0.1ms

3. (2 points) Given a computer architecture with a two level cache hierarchy with the following characteristics:

	L1	**L2**	**RAM**
Hit time (ns)	2	8	100

A computer with this architecture executes a program, which achieves a local hit rate of L1 of 0.8, a local hit rate of L2 of 0.9 and a hit rate of RAM of 1 (i.e. the whole program resides in RAM).

a) Assuming that 100% of the memory accesses are write operations, write down the formula that computes the average memory access time for (a) *write-through* and (b) *write back for* L1 and L2.

b) Considering L1 and L2 caches as a global cache, which is the *hit rate* and *average access time* of this global cache?

c) Assuming that a double size is 8 bytes and one level cache hierarchy with a cache block of 64 bytes and given the following code:

```
double a[1016]
for (i=0;i<1000;i=i+32){
       a[i]=a[i+8]+a[i+16];
}
```

How does memory bandwidth and average access time change when using a 4 way multibanked cache?

Solution

a)

Write through:
T=2+8+100=110ns

- **Write back**

T=2+(1-0.8)(8+(1-0.9)*100)=2+0.2(8+0.1*100)=2+0.2*18=2+3.6=5.6ns

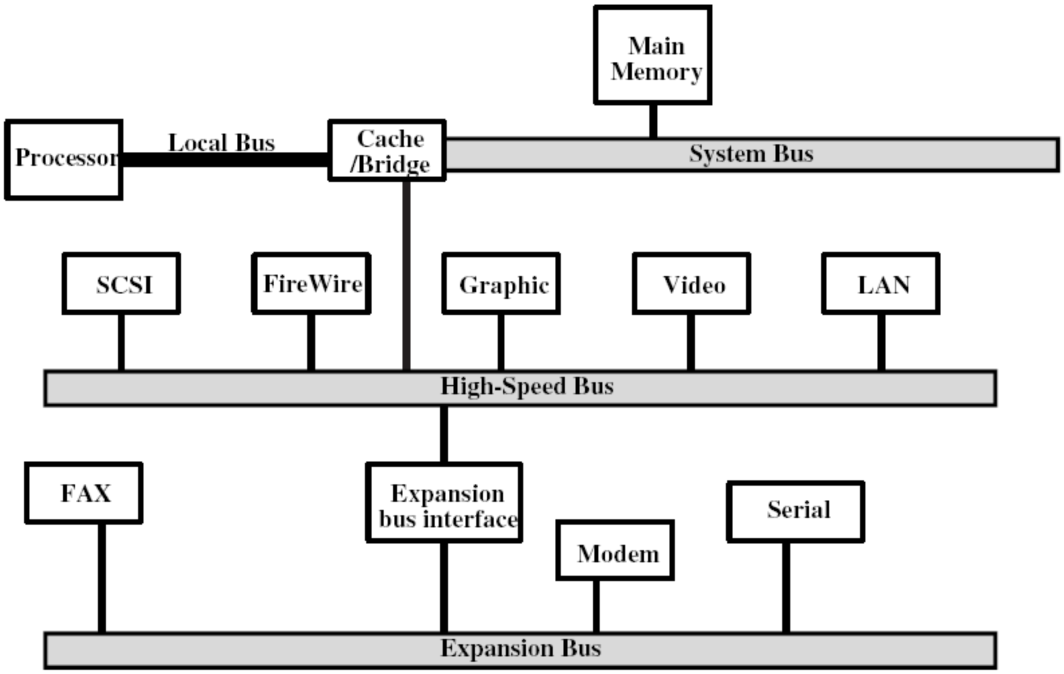
b)

T=2+(1-0.8)*8=2+0.2*8=2+1.6=3.6ns
Global HitRate=1-(1-0.8)*(1-0.9)=1-0.2*0.1=1-0.02=0.98.

c) Throughput 3 times larger for accessing cache banks in parallel. Same access time per double, but total access time 3 times smaller for the parallel accesses.

4. (1point)
   a) Draw an architecture representing the Mezzanine approach including the system bus, high-speed bus and expansion bus.
   b) Draw the devices connected to each type of bus.
   c) For each bus specify if it is located on the motherboard or on processor.

Solution:
a +b)



c) Local bus on processor, all the others on motherboard

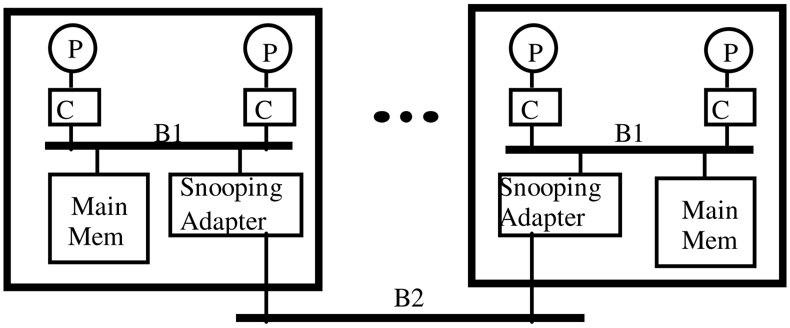
5. (1point)
   a) Describe test-and-set (t&s) lock implementation.
   b) What improvement brings test and test and set (tt&s) over t&s?
   c) What improvement brings test and set with exponential back-off over t&s?
   d) Is any of the three lock implementations fair? Justify your answer.

Solution:

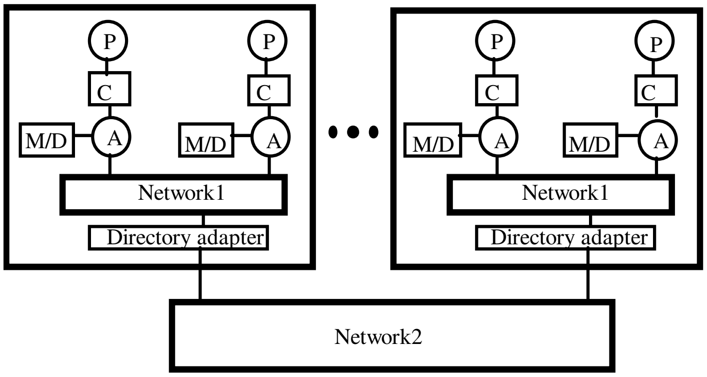
Culler 5.5.3

6. The following figures show two different cache-coherent distributed memory architectures (NUMA).
   a) Describe each of the two architectures.
   b) Compare the two architectures, emphasizing in the advantages and disadvantages of each of them.

Architecture A:



Architecture B:



Solution:
Culler chapter 8