A Decision Process Analysis of Recently Reported Implicit CO-scheduling Results

R. Poovendran

ABSTRACT

Using a series of elaborate simulation,[?,?,?] Andrea et al proposed a coscheduling algorithm based on two-phase spin blocking. By performing benchmark measurements, they derived a set of systems parameters and using heuristic arguments, they showed how a local processor can make the decision to either spin the process for some more time or to block. One of their result was based on the barrier imbalance and the other one was based on pairwise cost-benefit. They also noted that their cost derivations were for competing but identical processes that needed to be coscheduled. We note that it is possible to model the experimental setup as a binary Bayesian decision process at the local work station and choosing appropriate parameters lead to the results derived in.[?,?,?] Hence, the results derived are interpretable using well founded theoretical framework that not only only interpreted the results, may also help in studying the variations they could have obtained by analysis before experiments.

1. Introduction

1.1. Background on Binary Decisions Problem

We first describe a binary detection problem and map it to the DSM problem. Then we present the optimal decision making based on Bayes criteria.

In a binary decision problem, the observer (local processor) makes a set of decisions based on the possible two final outcomes of the problem. One of the outcomes contains some important relevant data or message and the other one contains no important data. The outcome that contains no relevant data is called the *null hypothesis* (H_0) and the other one is denoted as *alternative hypothesis* (H_1). Each hypothesis contains one or more observations which are represented by random variables. When the observer makes the decision, depending on the final actual outcomes, the following four cases can occur.

1. H_0 was decided and H_0 occurred.
2. H_0 was decided and H_1 occurred.
3. H_1 was decided and H_0 occurred.
4. H_1 was decided and H_1 occurred.

We note that the observer makes correct decisions in cases (i) and (iv), while the observer makes wrong decisions in cases (ii) and (iii). In terms of the process scheduling we can decide the set H_1 as the event that there is a coordination or message or barrier synchronization from a remote host and H_0 as no such possibility. Depending on the binary decisions, it can be inferred whether it will be cost effective to let the local process spin or go to sleep.

The four decision outcomes can be related to inferences as:

1. Decide it is better to block the process and save unnecessary spinning.
2. Decide it is better to block the process and pay a penalty by missing the message.
3. Decide it is better to spin and pay a penalty by wasting resources.
4. Decide it is better to spin and benefit by being able to respond to the incoming message.

We note that the case (ii) is related to a *miss* and the case (iv) is related to a *false alarm* in the context of decision theory.

1.2. Bayesian Criterion

In using Bayesian criteria, two assumptions are made. First, the probability of the occurrences of two outcomes are known. They are the probabilities $P_0 = P(H_0)$ and $P_1 = P(H_1)$. We note that

$$P_0 + P_1 = 1 \tag{1}$$

Each decision has a cost associated with it. For example, if the local process was allowed to spin and no message was received or barrier was completed, then there is a wasted time that could have been assigned to some other process. Similarly, if the local process was blocked and had to be woken up then there is an associated cost for missing the synchronization or coscheduling. Clearly, for a simple binary decision problem, there is no cost associated with making the right decisions. However, in a composite hypothesis problem it may matter (as in the case of fairness across multiple type of process reported i.[?])

Let $D_i; i = 0, 1$ where D_0 denotes "decide H_0" and D_1 denotes "decide H_1", we can define C_{ij} as the cost associated with the decision D_i given that the true hypothesis is H_j. In particular, the cost assignment for all the four cases of the binary hypothesis are given below

1. C_{00} for case 1
2. C_{01} for case 2
3. C_{10} for case 3
4. C_{11} for case 4

We also assume that the cost of making a wrong decision is more than cost of making a correct decision. That is,

$$C_{01} > C_{11} \tag{2}$$
$$C_{10} > C_{00} \tag{3}$$
$$\tag{4}$$

Given $(P(D_i, H_j)$, the joint probability that we decide D_i and the hypothesis H_j is true, the average cost/risk is given by

$$R = E[C] = C_{00}P(D_0, H_0) + C_{01}P(D_0, H_1) + C_{10}P(D_1, H_0) + C_{11}P(D_1, H_1) \tag{5}$$

From Bayes rule we have $P(D_i, H_j) = P(D_i|H_j)P(H_j)$. If we denote that the underlying distributions of hypotheses H_0 and H_1 are given by f_{h0} and f_{h1}, after some algebra the decision rule comes to

$$\frac{f_{h0}}{f_{h1}} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{P_0(C_{10} - C_{00})}{P_1(C_{01} - C_{11})} \tag{6}$$

Since the prior distributions are known, if we denote

$$\lambda(y) = \frac{f_{h0}}{f_{h1}} \tag{7}$$
$$\eta = \frac{P_0(C_{10} - C_{00})}{P_1(C_{01} - C_{11})} \tag{8}$$

Then, the Bayes criteria reduces to

$$\lambda(y) \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{9}$$

We will now interpret the results of the experiments reported in.[?] We denote the cost of blocking as a function of W as $f(W)$ instead of a constant.

1.3. Interpretation of the results in[?]: Local Cost-Benefit

One of the main decision made in[?] was the criterion for deciding whether to spin for more time in the presence of load-imbalance or not. In that experiment, the authors ran similar processes that were competing to be coscheduled. Since the processes are identical, their underlying distributions were identical. Hence, at any given time, there was equal probability of receiving message from any one of the competing process set. That is, the message, or barrier synchronization did not have any bias towards any particular process. Hence, $f_{h0} = f_{h1}$ was assumed in the experiments. (We note that this is not really a case in heterogeneous real world applications) Moreover, the simulations in[?] assign equal probalities to receiving or not receiving a message or load-imbalance. This reduces to $P_0 = P_1 = 0.5$ in the Bayesian decision.

From their experiments, the load-imbalance cost functions are (note that the penalty is a function proportional to time spinning in their study as it should be)

1. $C_{00} = 0$ (i.e. if blocking was correct no penalty is payed)
2. $C_{01} = (B + f(W)) + f(W)$. i.e. penalty for blocking is the spin time $B + f(W)$ and then the penalty $f(W)$ for waking up.
3. $C_{10} = f(V)$. i.e. the penalty for spinning to wait for the barrier synchronization without reaching it.
4. $C_{11} = 0$. i.e. pay no penalty by being spinning since the process arrives on time.
5. In the simulations reported in,[?] $f(W) = W$ and $f(V) = V/2$ were assumed. We note that these two quantities probably can be better measured using a benchmark. with these quantities, the decision problem reduces to

$$1 \underset{H_0}{\overset{H_1}{\gtrless}} \frac{(C_{10} - 0)}{(C_{01} - 0)} \tag{10}$$

$$\Rightarrow C_{01} \underset{H_0}{\overset{H_1}{\gtrless}} C_{10} \tag{11}$$

In this case, their decision process reduces to

$$Decide \begin{cases} \text{spin wait} & \text{if } (C_{10} = V/2) < (C_{01} = B + 2W) \\ \text{block} & else \end{cases}$$

In simple words, if $v < 2(B + 2W)$ it is beneficial to spin waiting for the load-imbalance to close.

1.4. Pairwise: Cost-Benefit: Incoming Messages

In the second experiment, authors developed their results based on the intuition that a process handling messages should continue to spin. H_0 here represent the event that there is no-message and H_1 here is message arrival. As in the previous case, identical but competing jobs were used for coscheduling. Probabilities on receiving or not receiving a message were set equal implicitly. The correct decisions were not assigned any penalties. Round trip and the processing overhead at the end-to-end pairwise processing time was $2L + 4O$. The costs, derived from various time requirements are enumerated below:

(a) $C_{00} = 0$ (i.e. if blocking was correct no penalty is payed)
(b) $C_{01} = 2L + 4O + 5W$. i.e. penalty for blocking is the flight latency, processing time, "short spin" time and three more block related penalties.
(c) $C_{10} = 2L + 4O + T$. i.e. the penalty for latency + processing + additional wait.
(d) $C_{11} = 0$. i.e. pay no penalty by being spinning since the process arrives on time.

With these quantities, the Bayes decision reduces to the their decision process as

$$Decide \begin{cases} \text{spin wait} & \text{if } (C_{10} = 2L + 4O + T) < (C_{01} = 2L + 4O + 5W) \\ \text{block} & else \end{cases}$$

This leads to the result obtained by the authors[?] as decide to spin more time if $T < 5W$.

2. M-ary Decision Process and the Issues of Fairness and Scalability of Co-scheduling

Derivations given in the original paper by Culler etal does not address the issues of multiple process case and fairness. We use m-ary decision process to analyze these cases. In m-ary analysis,, there are M outputs each corresponding to one of the M hypotheses. There are M decisions and corresponding to each of the decisions, there are M outcomes, forcing us to consider a total of M^2 possibilities in the decision process. As before, we will assign the cost of deciding that the hypothesis i is chosen and the outcome hypothesis is j is true by $C_{i,j}$. Then the average risk/cost in this case is given by

$$R = E[C] = \sum_{i=0}^{M-1}\sum_{j=0}^{M-1} P_j C_{i,j} \int_{z_i} f_{r|H_j}(R|H_j)dR \tag{12}$$

$$= \sum_{i=0}^{M-1} P_i C_{i,i} \int_{z_i} f_{r|H_i}(R|H_j)dR + \sum_{i=0}^{M-1}\sum_{j=0;i\neq j}^{M-1} P_j C_{i,j} \int_{z_i} f_{r|H_j}(R|H_j)dR \tag{13}$$

$$= \sum_{i=0}^{M-1} P_i C_{i,i} + \sum_{i=0}^{M-1}\sum_{j=0;i\neq j}^{M-1} P_j (C_{i,j} - C_{j,j}) \int_{z_i} f_{r|H_i}(R|H_i)dR \tag{14}$$

In this formulation, we note that the term corresponding to $\sum_{i=0}^{M-1} P_i C_{i,i}$ is a positive constant and hence, the minimization of the cost function is decided by checking which of the regions of integration yields the minimal cost function. If we denote

$$I_i = \sum_{j=0;i\neq j}^{M-1} P_j (C_{i,j} - C_{j,j}) \int_{z_i} f_{r|H_i}(R|H_i)dR, \tag{15}$$

the decision process is given by decide process i if $I_i < I_j; 0 \leq j \leq M; i \neq j$.

If we define the likelihood ratio as

$$\lambda_i = \frac{f_{r|H_i}(R|H_i)}{f_{r|H_0}(R|H_0)}, \tag{16}$$

the decision process reduces to the familiar form as in the case of binary process as in equation(1.2). We note that that the equation(**??**) can be used to show how to select the appropriate thresholds for different competing processes.

In a multiprocess environment, different processes have different granularities. Some processes may perform relatively more computations than communications. If one were to measure the amount of time each process spends on computing and communicating, and dynamically compute the relative frequencies or the probabilities of the message arrivals,the expected message arrivals for different processes will be different. Especially, a process that spends more time computing will have relatively low probability of message arrival compared to a process that spends relatively more time communicating.

These observations mean, in the m-ary decision process, the message arrival probabilities for different processes will be different. Hence, one can not expect the same spin wait threshold time for all different processes for the case of cost-benefit analysis of the pairwise message arrival.

We first show that if all the processes have the same spin-wait threshold as used by the authors, after some algebra

$$C_{i,0} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{C_{0,i} P_i \lambda_i}{P_0} \tag{17}$$

where $C_{i,j}$, denotes the cost of deciding in favor of event i but the j is true. The λ_i is given by the ratios of the densities of hypotheses i and 0.

$$\lambda_i = \frac{f_i}{f_0} \tag{18}$$

This equation is same as the one in the case of single process due to the following reason. If all the processes are assumed to be identically distributed, to obtain the results used by the authors, we have to choose the cost $C_{i,j} = C_{i,0} = 2L + 4O + T$, and

$$\frac{\lambda_i P_i}{P_0} = \frac{\lambda_i P_i}{P_0} = \eta \text{for all } i, j = 1, \cdots, n \tag{19}$$

with $\eta = 1$. This in turn reduces the detection formula to

$$(2L + 4O + 5W) \underset{<}{\overset{H_1, H_2 : H_n}{>}} H_0, H_2 : H_n \ (2L + 4O + T) + (n-1)(T - 5W)\eta. \tag{20}$$

$$(2L + 4O + 5W) \quad H_0, H_2 \underset{<}{\overset{H_1, H_2 : H_n}{>}} H_n \quad (2L + 4O + T) + (n-1)(T - 5W) \tag{21}$$

$$5nW \quad H_0, H_2 \underset{<}{\overset{H_1, H_2 : H_n}{>}} H_n \quad nT. \tag{22}$$

This formulation reduces to the decision to spin block if $T \leq 5W$. This equation has some problems. This is the formula authors used in spin blocking the processes independent of whether the processes are identical or not. We also note that in deriving the formula, following implicit (incorrect) assumptions had to be made

(a) All the processes have same value of η_i set to 1.

(b) The cost of deciding to spin-wait a process i and finding that the message is from j is set to $2L + 4O + 5W$. This is really not the case.

(c) All processes have the same relative frequency of message arrivals

We note that even in the case of identical processes, application of single process results as done by authors is quite ad-hoc. We explain the reasons below.

In order to form the correct formulation, we need to make te distinction that the average spin time for different processes are different. We also note that the maximum amount of time a process is allowed to spin wait is given by $2L + 4O + 5W$ by the earlier section. Actual cost of allowing process i to spin wait and then finding the message comes from process j leads to a cost of $C_{i,j} = T_i + (2L + 4O + 5W)$, and not just $(2L + 4O + 5W)$.

We now derive the correct formula for n identical processes with (no loss of generality) the threshold being decided for process # 1.

$$\eta_1(2L + 4O + 5W) \quad H_0, H_2 \underset{<}{\overset{H_1, H_2 : H_n}{>}} H_n \quad (2L + 4O + T_1) + \sum_{i=2}^{i=n} \eta_i (C_{1,i} - C_{0,i}) \tag{23}$$

$$\eta_1(2L + 4O + 5W) \quad H_0, H_2 \underset{<}{\overset{H_1, H_2 : H_n}{>}} H_n \quad (2L + 4O + T_1) + \sum_{i=2}^{i=n} \eta_i (T_1 + 2L + 4O + 5W - 2L - 4O - 5W) \tag{24}$$

$$\frac{\eta_1 5W + (\eta_1 - 1)(2L + 4O)}{1 + \sum_{i=2}^{i=n} \eta_i} \quad H_0, H_2 \underset{<}{\overset{H_1, H_2 : H_n}{>}} H_n \quad T. \tag{25}$$

Hence, for n identical processes, the threshold for pair-wise spin block reduces to $T \leq \frac{\eta_1 5W}{1+(n-1)\eta_1}$. Setting the value of $\eta_1 = 1$ to derive the corrected threshold for the case of authors', we get $T < \frac{5W}{n}$, where n is the number of competing jobs.

From this derivation, if the processes are identical: *If n competing identical jobs are to be co-scheduled, at any time, one can expect with probability $\frac{1}{n}$ next message will be for the currently executing process. With probability $\frac{n-1}{n}$, the next message will be for a competing process. Hence, if there are several identical competing jobs,we can reduce the spin blocking time by a factor equal to the number of processes without hurting the net progress.*

observations: If the number n is very large there can be problem similar to process thrashing. So after a certain value of n we also need to set a set a lower limit to the threshold for spin blocking?

Identical Processes We now take a specific example of $n = 3$ and analyze the results. If all the processes are identical and $\eta_1 = 1$, then we have $T \leq 5W/3$. i.e. spiing for

Different Processes For illustration we have $n = 3$, table 1 summarizes a set of values of η_i's and the corresponding values of the thresholds

η_1	η_2	η_3	T_1	T_2	T_3
1	1	1	$5W/3$	$5W/3$	$5W/3$
1	1/2	1/2	$5W/2$	$W - L/5 - 2O/5$	$W - L/5 - 2O/5$
2	1	1	$10W/3 + (2L + 4O)/3$	5W/4	5W/4
10	1	1	5W (upper bound)	$5W/12$	$5W/12$
100	1	1	5W (upper bound)	$5W/102$	$5W/102$

In the table, if we get the estimated threshold greater than $5W$, it is cost effective to spin upto $5W$ and block. This isreflected in the last two rows of the table given. *There must be a way to find a non-zero lower bound too!*

2.1. Interpretation in terms of fairness

We are in fact able to explain (thanks to lottery paper in the course material) that our results have interpretation from the fairness point of view as well. In the earlier example, we noted that if $\eta_1 = 2\eta_2 = 2\eta_3$, the proportional allocation of performing communications is $2 : 1 : 1$. i.e. process 1 spends a lot of time communicating than computing compared to processes 2 and 3.If we convert the times of communicating to computing they are approximately in the ratios of $1 : 2 : 2$ (more communication implies proportionally less computation). i.e. process 1 spends only $1/5$ time as the rest of the processes computing. Our thresholds gave the values of spin waits $\approx 2.5W : W : W$: which gives the advantage ratios of receiving messages to processes as $2.5 * 1/5 : 1 * 2/5 : 1 * 2/5 = 2 : 2 : 2 = 1 : 1 : 1$. Hence, we note that the Bayesian decision process suggests that the process that spends a lot of time communicating and less time computing should be allowed to spin wait for the time proportional to the probability of its message arrivals. In fact, from Bayesian analysis, we have threshold proportional to the probability of the message arrival rates for any process if the $\lambda_i = 1$.

From the fairness point of view, the *Lottery scheduling paper of CMSC 818K* suggests that the time quantum given for any process for any resource should be inversely proportional to the amount of time it uses the resource. Although we have not uses the currency approach, our approach of threshold selection (or allowing the processor to be busy for an additional short time occupied with a process) suggests that, in an environment with multiple, competing jobs, a process that spends a lot of time communicating should be allowed to have proportioal amount of time for spin wait to reduce the overall system cost measured in terms of the peanlty paid by missing an incoming message of by false waiting on a message that will have low probability of arrival.

We note however, unlike the lottery scheduling scheme where the only quantity used in decision making is the realtive amount of usage of the resource, we can not simply use the arrival probabilities. We also need to consider the ratios of the process densities (λ_is). This explains the non exact proportionality of the thresholds.

Some more thinking required here.

3. Discussions

We note that the scope of the experiment is limited in the following sense.

(a) Authors assumed that the competing jobs are all identical.

(b) They also assumed that the probability of receiving or not receiving a request for a particular process is identical.

(c) They assumed that there is no penalty for making the right decision with respect to one process.

(d) They assumed that W is already known fixed quantity.

(e) We were able to show that using the Bayesian decision approach, we can derive the thresholds derived by the authors for load-imbalance, pairwise cost benefit for single process, pairwise cost benefit for competing processes case by noting that the authors implicitly assume that the processes are identical even if they are competing.

(f) We also showed how to compute the thresholds for some examples.

(g) Although we have used the values of $2L + 4O + 5W$, given by the authors, our approach can be used for other thresholds as well.

4. Open Questions

More relevant issues to be discussed later.