NORTEL NETWORKS

How the world shares ideas.

On-switch Algorithm Adjustment for XA-CORE

Nicolae Santean
Traffic Engineering Discipline, 0G60

Performance Analysis

In this paper we discuss the impact of partial de-serialization of collision handlers in XA-CORE. The present strategy allows handlers of isolated collisions to be executed in parallel. This change is referred to as partial de-serialization since handlers of grouped collisions are still serialized.

In this sections we show that the processing power gain induced by this change depends on the clustering of collisions and that there are cases where the gain is not substantial.

Consider a set of k processors running in parallel. A singleton is a processor that is not involved in any collision at a given time. Considering the transitive closure of collision relationship among processors we can define a collision cluster as a maximal group of processors in collision at a given time. One can observe that singletons can be viewed as clusters with one processor (trivial clusters). All the other clusters are called non-trivial clusters. The size of a cluster is the number of its processors.

The average number of collision handlers present in the system at an arbitrary given time can be expressed as

$$n_b = k \cdot B_u - \alpha_{\geq 2}$$

where B_u is the data contention rate per PE at 100% occupancy and $\alpha_{\geq 2}$ is the average number of clusters of size at least 2. Furthermore, the average number of collision handlers involved in collisions of an arbitrary cluster is given by

$$n_a = \frac{k \cdot B_u}{\alpha_{\geq 2}} - 1.$$

Then one can express the waiting on serialization in both before and after change cases

$$w_b = \frac{n_b \cdot (n_b - 1)}{2} \quad , and \quad wa = \frac{n_a (n_a - 1)}{2} \cdot \alpha_{\geq 2}$$

where by w_b we denote the waiting on serialization before the change and by w_a the waiting on serialization after the change. Then, after some formal calculus we obtain the gain in procession power due to partial de-serialization, given by

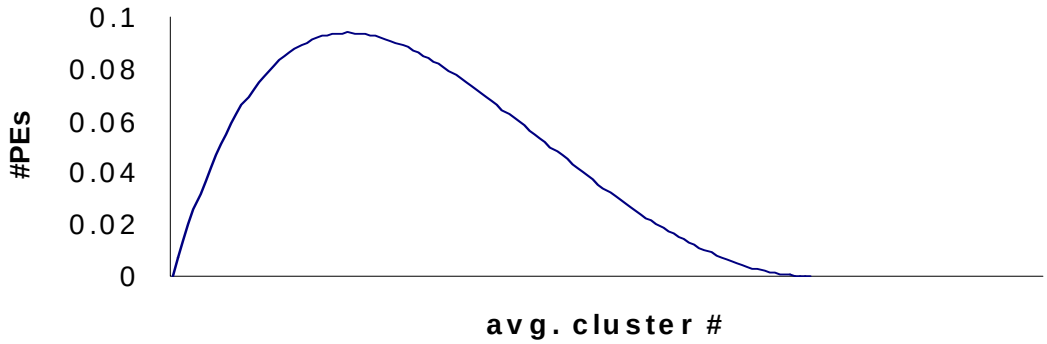
$$\Delta_w (\alpha_{\geq 2}) = w_b - w_a = \frac{(k \cdot B_u - \alpha_{\geq 2})^2 (\alpha_{\geq 2} - 1)}{2 \cdot \alpha_{\geq 2}}$$

Recalling, $\alpha_{\geq 2}$ is the average number of clusters of size greater than 2 (non-trivial clusters) and therefore $1 \leq \alpha_{\geq 2} \leq k \cdot B_u$. In the interval $[1, kB_u]$ Δ_w is positive, null at extremities and it has a unique critical point(maximal) in

$$A_{\geq 2} = \frac{1 + \sqrt{1 + 8 \cdot k \cdot B_u}}{4}.$$

A particularly interesting case is when $\alpha_{\geq 2}$ is exactly half of the contention rate (k B_u /2). In this case, surprisingly, although the change completely eliminates waiting times due to collision handler serialization (since we have only separate pairs of processors in collision) the obtained gained is not necessarily maximal. As mentioned before, only the neighborhood of the critical point $A_{\geq 2}$ gives a maximal difference.

One can therefore conclude with the following interpretation: if $\alpha_{\geq 2}$ is situated in the neighborhood of $A_{\geq 2}$ then the partial de-serialization induce a highest processing power gain whereas in the other cases smaller gain and even no gain at all are induced. The following graph plots the gain obtained in an 11 processor system at 20% contention rate (B_u).



If the average number of non-trivial clusters in this system is about 1.32 then we can count on the maximal gain of about 0.093 PEs per unit of time processing power.

Formula Adjustment

Given the probability of contention between two parallel processes, one can theoretically estimate $\alpha_{\geq 2}$. Since this is a difficult task that will eventually be addressed in the future, we first analyze the case of a maximal gain since this case triggers the greatest adjustment. In this optimistic case, the value of w_s from step 6 of the on-switch algorithm has to be adjusted by subtracting

$$\frac{\left(k \cdot B_u - \alpha_{\geq 2}\right)^2 \left(\alpha_{\geq 2} - 1\right)}{2 \cdot \alpha_{\geq 2}} \quad , \quad where \quad \alpha_{\geq 2} = \frac{1 + \sqrt{1 + 8 \cdot k \cdot B_u}}{4}.$$

Oppositely, the worst case situation is reached when $\alpha_{\geq 2}$ is either equal to 1 – case in which we always have an unique cluster in the system – or asymptotically equal to $k \cdot B_u$ - case in which they are both infinitely small and practically no contention occurs. Observe that $\alpha_{\geq 2}$ is equal to $k \cdot B_u$ only when the contention rate is null, case in which both values are null. In these two extreme cases no adjustment is needed.

Other two possibilities to estimate the formula adjustment are: either to obtain $\alpha_{\geq 2}$ from measurements, or to experimentally “tune-up” the value for $\alpha_{\geq 2}$.