Review comments:

Reviewer #1:

~~It is very well written paper. One thing that I can suggest to authors is talking about the related work in a separate section or part of introduction or background section.~~

~~Comentario: No lo vamos a separar. Decir amablemente. Decir que no obstante se ha extendido el related work.~~

Reviewer #2:

~~The authors provide a very interesting contribution on perishable inventory control for heterogeneous platforms. The manuscript is well-written, with an interesting contribution and adequate experimentation.~~

~~My only recommendation is to perform a more extensive literature review on related topics in heterogeneous computing. Other than this minor issue, the paper is ready for publication in my opinion and will be an interesting contribution to the state of the art in this area.~~

Comentario: algunas cosas se han puesto…

Reviewer #6:

~~1. The variable "t" is used tyo represent time on page 4 and threads on page 17. You ought to use something different, One suggestion is below:~~

~~2. On the graph on page 17, I strongly suggest not using a "legend" to explain which curve is which, I suggest putting text next to each line instead, i.e. "1 thread", "2 threads" . "16 threads".~~

~~Comentario:~~

~~3. If i read correctly, if we count the number of CPU chips and GPU chips used in each evaluation case, the performance for 1 CPU chip plus 8 GPU chips is about the same as 8 CPU chips.~~

~~I'd hope that the GPU's could do better, but you got what you got.~~

Comentario: que 1cpu+8gpu es como 8cpu??? No… Sobre el performance de las GPU’s nada que decir, ygwyg, correcto.

~~4. One question on the parallel speedup: what is the ideal calculation time target for this kind of problem? I assume you want to generate an order request for the next time interval, and maybe a pre-order forecast for the interval after that. For a journal on Inventory Control, the audience may already know the answer, but for a parallel computing journal, I think the extra context would be good.~~

Comentario: Comentado en el mismo parrafo en el que se ha puesto la referencia al IJPR.

Reviewer #7:

~~The contribution describes the solution of an economical problem "perishable inventory control" in a massive parallel way. The article is well written. However, I would like to remark a few points:~~

~~The solution discussed can definitely be considered as parallel or distributed computing. Whether the computations made can be called high performance computing has to be reconsidered.~~

~~The algorithm contains a Monte Carlo method, which is parallelised. Samples of different size are packed into bins according to different heuristics in order to balance the workload. At the end of Sec. 4 it is stated that the balance of the workload is almost perfect. This is a nice result meaning that further improvements are not necessary. Otherwise one could think about the alternative solution of a dynamic approach like a master slave process in order to distribute the load.~~

~~The scaling of the whole problem is very good -- for both, pure CPU runs and runs with GPUs -- but not perfect. What is the reason for this? The load balance is perfect. The communication cost have not been discussed. They might be not negligible, the whole load balancing problem becomes more complex and the bin packing approach is an approximation only. If the communication costs are negligible the problem is embarrassingly parallel. For this case I suggest to avoid the usage of the term high performance computing.~~

Comentario: ~~decir que hemos cambiado el termino hpc por parallel and distributed computing. Decir que las heurísticas del bin packing problem si que tienen un coste computacional muy bajo y son negligible time. Decir que había una función que constituia un lower bound para el coste asociado a cada vector Y de pedidos(o como lo hayamos llamado en el paper) que permitia descartar calcular ciertos casos comunicando por pthreads y MPI que hilo había encontrado hasta el momento el coste más bajo, pero que al considerar muchos procesos o hilos no suponía ninguna ventaja de tiempo apreciable…~~

The concurrent execution model within a single GPU warp should be addressed. It is not very clear to which extent the different samples can be handled with single instruction multiple data instructions. This can be a reason for a limited efficiency on the GPUs.

Minor points:

Section 5, line 276: The Intel Xeon E5 2650 has 8 cores. Presumably hyperthreading was used -- 16 virtual cores.

~~Section 5, line 280: "mpicc compilers" mpicc is not a compiler but a compiler wrapper of the MPI library, which compiler (and version) has been used (gnu, intel, ...) and which MPI library (mvapich, ...)?~~

~~Section 4.2, line 231 to line 235: The choice of the block size (BS) is mentioned: What is the best block size?~~

~~Section 2, line 64: Matlab is a commercial tool. For reproducibility it would have been better to use an open source software like Octave.~~

Comentario: pero no ha sido el caso… De todas formas es portable.