TPDS—reviews

Reviewer：1

Recommendation：Reject

Comments：
Dear authors,

thank you very much for submitting your work to TPDS. You work focuses on a timely problem, but also one that has been extensively studied by prior work. Hence, the standards are quite high, given the vast availability of baselines. I appreciated the paper structure of the paper that tries to build an intuition based on certain observation and the evaluation on both simulation and actual deployment. However, I have several concerns with the current submission.

First and foremost, I think the paper suffers from a lack of clarity in certain very important parts of it. For example, the main aspect of the design is the "self—attention" mechanism. It is even part of the paper title. However, this is never explained thoroughly. There is a reference in Section V.D. However, given its importance, there should be a description of the mechanism for a reader that might not be familiar with the topic.

Although I did appreciate the intend behind the observations in Section IV, I had a hard time understanding and believing them, as I believe that these are only correlations that do not necessarily lead to causation. Hence, they cannot be used a motivation to design a resource allocation scheme. Specifically, observation 1 says that there is a correlation between utilization and latency. However, this does not mean that there are no cases of services with utilization and low latency. In fact, this is the ideal operation point, i.e. at the knee of the latency—throughput curve. Observation 2 correlates the CPU utilization of a latency—critical service and its neighbors. Although this might hold true in the particular deployment scenario, it is not a property of any deployment. CPU utilizations of collocated services have to do with the services themselves, their load level, and the placement decision. Unless there is some dependency relationship between the collocated services, I cannot see how the second observation can be used.

The description of the server filter mentions that Hestia allocates a continuous set of hyper—threads. It is not clear to me why this is necessary. I understand that it simplifies the way of thinking about the problem, but also at the same time, it can lead to high inefficiency due to fragmentation. Imagine a scenario in which there are available CPU resources but they are not contiguous. Will Hestia accept this as a candidate set of hyper—threads?

I also, have several concerns about the set of features in the prediction model. Unlike previous work on the same topic, Hestia does not consider the workload type or other lower—level metrics, e.g. performance counters to predict interference. Instead, it uses RPS. How do temporal variations of RPS affect the result? Do you assume a consistent load over the duration of the experiment? What happens if RPS changes? In the same model explanation, you never describe what are the

LS embeddings, how did you come up with them, and what they capture. You only mention that these are learnable parameters.

The paper lacks an implementation section that explains the core implementation of Hestia and how it can be integrated to existing cluster management systems, e.g. Kubernetes or Docker Swarm.

Finally, the evaluation needs to be improved. First, it does not include relevant baselines. I acknowledge reading the justification of why the authors decided to do so, but I was not convinced. Including a comparison with Parties or Paragon would strengthen the contribution. Second, it does not include an evaluation of the Hestia predictor itself. It is not clear how long it takes to make a placement decision and what are the necessary resources. Including only end-to-end metrics can be misleading. Third, the simulation results are not clear. How is it possible to evaluate interference when simulating latency-critical instances? Fourth, I found the performance benefits quite marginal compared to the gains prior work has shown with such complicated ML-based approaches.

Additional Questions:
1. Which category describes this manuscript?: Research/Technology

2. How relevant is this manuscript to the readers of this periodical? Please explain your rating under Public Comments below.: Relevant

1. Please explain how this manuscript advances this field of research and/or contributes something new to the literature.: The paper proposes a new way to allocate server resources to latency-critical applications that considers interference at the level of hyperthreads, hence it operates a lower granularity compared to the state of the art.

2. Is the manuscript technically sound? Please explain your answer under Public Comments below.: Partially

1. Are the title, abstract, and keywords appropriate? Please explain under Public Comments below.: Yes

2. Does the manuscript contain sufficient and appropriate references? Please explain under Public Comments below.: References are sufficient and appropriate

If you are suggesting additional references they must be entered in the text box provided. All suggestions must include full bibliographic information plus a DOI.

If you are not suggesting any references, please type NA.: NA

3. Does the introduction state the objectives of the manuscript in terms that encourage the reader to read on? Please explain your answer under Public Comments below.: Could be improved

4. How would you rate the organization of the manuscript? Is it focused? Is the length appropriate for the topic? Please explain under Public Comments below.: Could be improved

5. Please rate the readability of the manuscript. Explain your rating under Public Comments below.: Difficult to read and understand

6. Should the supplemental material be included? (Click on the Supplementary Files icon to view files): Does not apply, no supplementary files included

7. If yes to 6, should it be accepted:

8. Would you recommend adding the code/data associated with this paper to help address your concerns and/or strengthen the paper?: No

Please rate the manuscript. Please explain your choice.: Poor

Reviewer: 2

Recommendation: Author Should Prepare A Minor Revision

Comments:
This paper tackles an interesting problem, of scheduling services with different performance and latency requirements in complex server architectures.

It presents Hestia, an interference-aware scheduler, that leverages an attention mechanism to capture interference (core and socket) creating a model to predict performance of long running services.

Though both simulations and real world experiments it demonstrates that it can effectively decrease CPU utilization of large scale clusters and tail latency of long running applications with several instances.

Even though it solves an interesting problem, with a smart and flexible idea, I like to see Hestia integrated with an existing open-source solution.
At the same time, a comparison with an existing interference-aware scheduler in a real world cluster, would make the paper so much stronger.

==========

Detailed Notes:

Abstract:

* a server usually deploys multiple latency-sensitive (LS) instances. -> Containers? Servives? Deploys or Collocates? Sentence a bit unclear
* there is currently a lack of fine-grained binding core strategy -> core-binding strategy
* to effectively guide production scheduling -> production workload scheduling?
* self-attention mechanism -> Explain?
* as well as workload and hardware heterogeneity -> Typo
* revealing an achievement in the performance of LS services of up to 20.89%. -> improvement instead of achievement -> what is considered performance here?

Interference-aware scheduling:
* Leverage an attention mechanism to capture interference (core and socket) creating a model to predict performance
* Interference scoring mechanism to proactively address interference
* Minimal metric overhead
* Seamless integration with existing schedulers

Related Work:
* 32K LS instances — how many actual services where there? And what kind of characteristics?
* User-facing service performance reduction is not necessarily translated to increased tail latency
* LS instances execute within long-lived containers2 — if containers and instances are used interchangeably, what does this sentence say?
* Would be nice for Figure 1 to show Caching too as described in the text
* As there are a significant number of LS instances associated with various LS services —> what is the estimated number of container per LS service? Use a reference
* How do you define maximum potential interference? Ref section

Design:
* HT scalability is just a 2x increase from Cores — How much of a challenge is that in practice?
* How would Hestia work for LS with no prior RPS or CPU usage information
* What about services that are already deploy? Does Hestia support migrations
* Is Hestias code or experiments publicly available

Evaluation:
* I understand the challenge of comparing existing interference-aware scheduler on simulation mode(using their own metrics) but would be nice to have that in a real cluster evaluation
* How does the predictor scale with the number of cores, applications, etc

Nit:
* Fig2-Fig4, captions misaligned
* Use HyperRef to make Section refs more user friendly

Additional Questions:

1. Which category describes this manuscript?: Research/Technology

2. How relevant is this manuscript to the readers of this periodical? Please explain your rating under Public Comments below.: Relevant

1. Please explain how this manuscript advances this field of research and/or contributes something new to the literature.: This paper describes Hestia, an interference-aware scheduler, leveraging an attention mechanism to capture interference (core and socket) and create a model to predict performance of long running services in shared compute cluster. It uses an interference scoring mechanism to proactively address interference of such long running services and with minimal metric overhead warrants seamless integration with existing schedulers like kubernetes.

Paper tackles an interesting problem, of effectively scheduling services with different performance and latency requirements in complex server architectures. By modeling the socket and core depedencies, manages to improve the CPU efficiency of large scale clusters, while reducing interference between applications.

2. Is the manuscript technically sound? Please explain your answer under Public Comments below.: Yes

1. Are the title, abstract, and keywords appropriate? Please explain under Public Comments below.: Yes

2. Does the manuscript contain sufficient and appropriate references? Please explain under Public Comments below.: References are sufficient and appropriate

If you are suggesting additional references they must be entered in the text box provided. All suggestions must include full bibliographic information plus a DOI.

If you are not suggesting any references, please type NA.: NA

3. Does the introduction state the objectives of the manuscript in terms that encourage the reader to read on? Please explain your answer under Public Comments below.: Could be improved

4. How would you rate the organization of the manuscript? Is it focused? Is the length appropriate for the topic? Please explain under Public Comments below.: Satisfactory

5. Please rate the readability of the manuscript. Explain your rating under Public Comments below.: Readable — but requires some effort to understand

6. Should the supplemental material be included? (Click on the Supplementary Files icon to view

files): Does not apply, no supplementary files included

7. If yes to 6, should it be accepted: After revisions. Please include explanation under Public Comments below.

8. Would you recommend adding the code/data associated with this paper to help address your concerns and/or strengthen the paper?: Yes

Please rate the manuscript. Please explain your choice.: Good

Reviewer: 3

Recommendation: Reject

Comments:
1. The paper conclude that lowering CPU usage leads to better performance. They discuss the linear relationship of CPU and RPS (request per second) and conclude that lowering CPU would improve performance as you can get higher throughput. They do not address workload-specific internal contentions or contention of resources other than CPU.
2. The paper makes the claim of "minimal overhead" in section F but does not actually measure the overhead of recording additional metrics and additional computing necessary for predictive CPU interference patterns.
3. Schedulers in the real world don't have the luxury of measuring workload-specific performance without any interference. The authors assume that this is possible.
4. While the authors optimize CPU performance successfully, they don't account for the second-order effects - which component take additional hit? For example, did this lead to increase in memory bus utilization?
5. The claim "Hestia can be plugged into another scheduler" needs to be backed by details, especially when the other scheduler may choose to optimize for unrelated metrics that directly work against Hestia's concerns.

Additional Questions:
1. Which category describes this manuscript?: Research/Technology

2. How relevant is this manuscript to the readers of this periodical? Please explain your rating under Public Comments below.: Very Relevant

1. Please explain how this manuscript advances this field of research and/or contributes something new to the literature.: The authors have implemented a new scheduler that attempts to minimize resource usage of large compute clusters, specifically CPU utilization of servers, as well as the performance of a distributed workload by reducing interference among workloads sharing physical resources such as memory, CPU caches, and inter-socket communication. The paper details the math behind designing a predictive algorithm for scheduling workload that accounts for CPU

interference. It then goes on to compare the scheduler with other workload schedulers.

2. Is the manuscript technically sound? Please explain your answer under Public Comments below.: Partially

1. Are the title, abstract, and keywords appropriate? Please explain under Public Comments below.: Yes

2. Does the manuscript contain sufficient and appropriate references? Please explain under Public Comments below.: References are sufficient and appropriate

If you are suggesting additional references they must be entered in the text box provided. All suggestions must include full bibliographic information plus a DOI.

If you are not suggesting any references, please type NA.: NA

3. Does the introduction state the objectives of the manuscript in terms that encourage the reader to read on? Please explain your answer under Public Comments below.: Could be improved

4. How would you rate the organization of the manuscript? Is it focused? Is the length appropriate for the topic? Please explain under Public Comments below.: Poor

5. Please rate the readability of the manuscript. Explain your rating under Public Comments below.: Difficult to read and understand

6. Should the supplemental material be included? (Click on the Supplementary Files icon to view files): Does not apply, no supplementary files included

7. If yes to 6, should it be accepted:

8. Would you recommend adding the code/data associated with this paper to help address your concerns and/or strengthen the paper?: No

Please rate the manuscript. Please explain your choice.: Poor