**[Review #180A](https://socc20.hotcrp.com/review/180A)**

**Overall merit**

**1.** Reject

**Paper summary**

This paper tackles the challenges of (static) resource allocation in HPC clouds. The authors propose CPU and memory over-commitment as techniques towards improving cloud utilization. The paper demonstrates the feasibility of resource over-commitment by running popular HPC workloads on a TORQUE cluster.

**Novelty**

**1.** Published before or openly commercialized

**Writing quality**

**3.** Adequate

**Reviewer expertise**

**3.** Knowledgeable

**Comments for author**

SoCC has not seen a host of HPC-focused papers, which creates a good opening for this work. However, the paper fails to advocate effectively for its own cause. Sure, it has some interesting tidbits but on the whole, it lacks both depth and technical contributions needed for a full paper. I will list my concerns below (under three broad categories) in the hope that it will help in revising the paper.

Uniqueness of HPC requirements is not highlighted: Despite spending much of the sections-1 and 2, the paper fails to make a clear case for why the requirements of HPC are distinct from other workloads, and why the current resource managers used in public/private clouds are insufficient to serve them. For example, you mention workloads from the Johns Hopkins Applied Physics Lab but do not explain how its performance, tail-latency, and fault-tolerance requirements are different from those of web/data analytics workloads running on public clouds. Sec-2.1 seems dated, and section-2.3 misses a large body of related work on resource over-commitment (for instance, Resource Central [SOSP17], Harvesting Spare Cycles [OSDI16], and Resource Deflation [Eurosys19]). It is foundational for this paper to establish the uniqueness of HPC requirements; without that, the rest of the paper falls apart.

Meeting the standard for an industry paper: The authors might have misunderstood the category of industry/systems papers. Evaluating a commercial product in a simulated lab environment does not meet the expected standard, especially if no novel insights are revealed. Industry papers typically describe experiences from systems deployed in the real-world, often at scale and over long periods of time. For reference, please check Hotspot Mitigation for the Masses [SoCC19].

Lack of new systems insights (in mechanisms, policies, evaluation methodologies, or deployment experiences): Resource over-commitment is great when time-varying workloads align well; however, for a system that promulgates SLA guarantees, the scheduler should be explicit about its guarantees and must incorporate fail-safe mechanisms like preemptive termination of low priority jobs. Simply moving VMs around in a reactive manner (as suggested in section-3.3) neither guarantees performance nor improves utilization. There are a couple of ways to restructure this work. One option is to concretely define multiple classes of SLAs for the HPC cloud, then design and evaluate scheduling policies/mechanisms that would meet these SLA. An example of such work is Long-Term SLOs for Reclaimed Cloud Computing Resources [SoCC14]. Second option is to analyze workloads from a real HPC cloud, then propose scheduling policies to increase its utilization. I can think of two examples: Preparation and Optimization of a Diverse Workload for a Large-Scale Heterogeneous System [SC19] and Harvesting Spare Cycles [OSDI16].

[**Review #180B**](https://socc20.hotcrp.com/review/180B)

**Overall merit**

**1.** Reject

**Paper summary**

* Paper's motivation is to improve efficiency of HPC workloads by running them in a cloud environment
* The paper evaluates several HPC scenarios using VMWare ESXi as a hypervisor to measure the throughput improvement afforded by resource overcommit

**Novelty**

**1.** Published before or openly commercialized

**Writing quality**

**3.** Adequate

**Reviewer expertise**

**3.** Knowledgeable

**Comments for author**

* Real experiments on workloads and hardware
* The motivation of improving cluster utilization for efficiency is an important and relevant topic
* The paper is very light on insights, analysis, and novelty. As written, it reads like a benchmarking report from running HPC workloads on a commercially available hypervisor. There isn't solid conclusions derived from how HPC workloads are more/less sensitive to overcommit or if the results are more broadly applicable to other HPC workloads or system configurations.

[**Review #180C**](https://socc20.hotcrp.com/review/180C)

**Overall merit**

**1.** Reject

**Paper summary**

This paper presents an implementation and evaluation of a HPC cluster that leverages CPU and memory overcommitment techniques offered by a VMWare hypervisor.

**Novelty**

**1.** Published before or openly commercialized

**Writing quality**

**3.** Adequate

**Reviewer expertise**

**3.** Knowledgeable

**Comments for author**

Sorry but I failed to see what was novel in this paper. Work on overbooking resources for improving utilization abounds and I saw neither a new methodology in how to overlook nor in terms of implementing it in a system. The claim that the overcommittment problem is somehow fundamentally different for HPC workloads was not substantiated by any type of workload characterization.