Compute System Simulator: Modeling the Impact of Allocation Policy and Hardware Reliability on HPC Cloud Resource Utilization

Jarrod Leddy
Huseyin Yildiz
jarrodleddy@microsoft.com
hyildiz@microsoft.com
Microsoft Corporation
Redmond, Washington, USA

CCS CONCEPTS

• **General and reference** → *Design*; *Reliability*; • **Computer systems organization** → *Cloud computing*; • **Computing methodologies** → *Agent / discrete models*; *Discrete-event simulation.*

KEYWORDS

Cloud Computing, Hardware Reliability, Resource Allocation, System Utilization

ACM Reference Format:
Jarrod Leddy and Huseyin Yildiz. 2025. Compute System Simulator: Modeling the Impact of Allocation Policy and Hardware Reliability on HPC Cloud Resource Utilization. In *Proceedings of The International Conference for High Performance Computing, Networking, Storage, and Analysis (SC '25).* ACM, New York, NY, USA, 1 page. https://doi.org/XXXXXXX.XXXXXXX

1 POSTER SUMMARY

As cloud computing systems grow in complexity and scale, optimizing resource allocation while maintaining reliability becomes increasingly critical. The paper introduces the Compute System Simulator, a Python-based tool designed to model and analyze how HPC workloads are launched, managed, and interrupted within cloud clusters. It simulates the interaction between allocation policies and hardware failures, offering insights into system utilization and workload delays.

The simulator operates on a modular framework, allowing researchers and engineers to test emerging allocation strategies under realistic failure conditions. It models a cloud cluster as a hierarchical tree structure—such as clusters composed of racks, nodes, and GPUs—where each tier is uniformly defined. This flexible architecture supports simulation of current and future systems, including large-scale deployments. At its core, the simulator uses discrete time steps to evolve the system state. Each hardware component is initialized with a failure time sampled from a Weibull distribution, which realistically models aging and reliability degradation. When a failure occurs, affected VMs are interrupted and reallocated based on the defined policy. The simulator supports three allocation types: Single (flexible placement), Whole (dedicated resources), and Whole Adjacent (performance-optimized placement). These policies can be applied at any tier of the system hierarchy. The simulator also tracks key diagnostics such as GPU utilization, workload interruption frequency, and effective runtime increases due to failures. These metrics are crucial for both cloud providers and users—idle GPUs represent cost inefficiencies, while job delays impact user productivity and budget.

To demonstrate its capabilities, the paper presents a detailed case study using Microsoft Azure's GB200 NVL72 cluster. This cutting-edge infrastructure features racks with 18 nodes, each containing four Blackwell GPUs and two Grace CPUs, interconnected via NVLink for high bandwidth and low latency. The study simulates workloads requesting either 72 or 68 GPUs, with and without hardware failures, over a two-year period.

Results show that in ideal conditions, GPU utilization matches expectations (100% for 72 GPUs, 94% for 68 GPUs). However, when failures are introduced, utilization drops significantly—down to 69% for 72 GPUs and 89% for 68 GPUs. More importantly, the runtime of workloads increases due to interruptions and reallocation delays. For 72-GPU jobs, the average runtime extension is 32 hours (19% increase), while 68-GPU jobs see only a 6-hour increase (4%). These findings highlight the trade-off between maximizing resource usage and maintaining resilience.

Further simulations explore the impact of varying mean time to failure (MTTF), mean time to repair (MTTR), and workload size. A sweep of 270 simulations reveals that optimal GPU utilization does not always occur at full rack allocations. Instead, smaller workloads often yield better performance under realistic failure conditions. These results underscore the importance of flexible workload sizing and checkpointing strategies.

The Compute System Simulator provides a powerful platform for analyzing the interplay between resource allocation and hardware reliability in cloud environments. It enables data-driven decisions for infrastructure design, workload management, and maintenance strategies. By simulating realistic scenarios, it helps cloud providers and users balance performance, cost, and resilience in high-performance computing systems.

Received 18 August 2025

SC '25, Nov 16–21, 2025, St. Louis, MO

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/XXXXXXX.XXXXXXX