Method and Apparatus for Optimized Job Scheduling in the High-Performance Computing Cloud Environment

Authors : Subodh Kumar, Amit Varde

Abstract : Typical on-premises high-performance computing (HPC) environments consist of a fixed number and a fixed set of computing hardware. During the design of the HPC environment, the hardware components, including but not limited to CPU, Memory, GPU, and networking, are carefully chosen from select vendors for optimal performance. High capital cost for building the environment is a prime factor influencing the design environment. A class of software called “Job Schedulers” are critical to maximizing these resources and running multiple workloads to extract the maximum value for the high capital cost. In principle, schedulers work by preventing workloads and users from monopolizing the finite hardware resources by queuing jobs in a workload. A cloud-based HPC environment does not have the limitations of fixed (type of and quantity of) hardware resources. In theory, users and workloads could spin up any number and type of hardware resource. This paper discusses the limitations of using traditional scheduling algorithms for cloud-based HPC workloads. It proposes a new set of features, called “HPC optimizers,” for maximizing the benefits of the elasticity and scalability of the cloud with the goal of cost-performance optimization of the workload.

Keywords : high performance computing, HPC, cloud computing, optimization, schedulers
Conference Title : ICHPCA 2023 : International Conference on High Performance Computing Architectures
Conference Location : Dubai, United Arab Emirates
Conference Dates : January 30-31, 2023

Open Science Index, Computer and Systems Engineering Vol:17, No:01, 2023 publications.waset.org/abstracts/155107.pdf