Contribution ID: **107** Type: **Talk**

Resource Scheduling and Allocation in HPC Cyberinfrastructure

High-Performance Computing (HPC) systems play a pivotal role in modern scientific research, enabling complex simulations, data analysis, and large-scale modelling across disciplines such as climate science, genomics, physics, and engineering. As these systems grow in scale and sophistication, the efficient scheduling and allocation of computational resources become crucial for ensuring optimal system performance, maximising resource utilisation, and meeting the needs of diverse user communities. In HPC environments, resource scheduling and allocation determine how tasks are assigned to hardware resources such as CPUs, GPUs, memory, storage, and network bandwidth. Effective scheduling strategies are critical for maintaining fairness among users, optimising job throughput, reducing waiting times, and enhancing energy efficiency.

Cyberinfrastructure, the integration of advanced computing platforms with large-scale data storage and high-speed networks, addlayers of complexity to resource management. The heterogeneity of hardware, dynamic workload demands, and multi-user environments require advanced resource scheduling algorithms. Traditional approaches like First-Come, First-Served (FCFS), Shortest Job First (SJF), and Backfilling have evolved to meet these challenges, while more advanced strategies like Priority Scheduling, Gang Scheduling, and Hybrid Scheduling offer increased flexibility and efficiency. Energy-aware scheduling has also gained importance, given the significant portion of operational costs that energy consumption in large HPC systems can account for.

Despite significant advancements, resource allocation in HPC systems continues to face challenges. The increasing complexity of workloads, the need for energy-efficient computation, and the evolving demands of users necessitate more sophisticated resource management techniques. Emerging trends such as machine learning-driven scheduling, serverless computing, and the integration of cloud-based HPC infrastructures are opening new avenues for improving the scalability, flexibility, and efficiency of HPC systems. Machine learning algorithms, for instance, offer the potential to predict workload patterns and optimize resource scheduling dynamically, while serverless computing models and cloud integration promise more adaptable and scalable resource provisioning.

This Article explores the current state of resource scheduling and allocation in HPC cyberinfrastructure, addressing key algorithms, challenges, and emerging trends shaping the future of high-performance computing.

Student or Postdoc?

Email address

Co-Authors

CHPC User

No

CHPC Research Programme

N/A

Workshop Duration

Primary author: WAYI-MGWEBI, Ntombovuyo (University of Mpumalanga)

Co-author: OGUNLEYE, Olalekan Samuel (University of Mpumalanga)

Presenter: WAYI-MGWEBI, Ntombovuyo (University of Mpumalanga)

Session Classification: HPC Technology

Track Classification: HPC Technology