JUWELS Booster – Early User Experiences

Andreas Herten
a.herten@fz-juelich.de
Jülich Supercomputing Centre, Forschungszentrum Jülich GmbH
Jülich, Germany

ABSTRACT

Over the last few years, GPUs became ubiquitous in HPC installations around the world. Today, they provide the main source of performance in a number of Top500 machines – for example Summit, Sierra, and JUWELS Booster. Also for the upcoming Exascale era, GPUs are selected as key enablers and will be installed numerously. While individual GPU devices already offer plenty of performance ($O(10)$ TFLOP/s$_{\mathrm{FP64}}$), current and next-generation supercomputers employ them in the thousands. Using these machines to the fullest extend means not only utilizing individual devices efficiently, but using the entire interconnected system of devices thoroughly.

JUWELS Booster is a recently installed Tier-0/1 system at Jülich Supercomputing Centre (JSC), currently the 7th-fastest supercomputer in the world, and the fastest in Europe. JUWELS Booster features 936 nodes, each equipped with 4 NVIDIA A100 Tensor Core GPUs and 4 Mellanox HDR200 InfiniBand HCAs. The peak performance of all GPUs together sums up to 73 PFLOP/s and it features a DragonFly+ network topology with 800 Gbit/s network injection bandwidth per node.

During installation of JUWELS Booster, a selected set of applications were given access to the system as part of the *JUWELS Booster Early Access Program*. To prepare for their first compute time allocation, scientific users were able to gain first experiences on the machine. They gave direct feedback to the system operations team during installation and beyond. Close collaboration was facilitated with the application support staff of JSC, giving unique insights into the individual processes of utilizing a brand-new large-sale system for a first time. Likewise, performance profiles of applications could be studied and collaboratively analyzed, employing available tools and methods. Performance limiters of the specific application on the platform were identified and proposals for improvement developed.

This talk will present first experiences with JUWELS Booster and the applications utilizing the system during its first months. Applied methods for onboarding, analysis, and optimization will be shown and assessed. Highlights of the state of the art of performance analysis and modeling for GPUs will be presented with concrete examples from the JUWELS Booster Early Access Program.

CCS CONCEPTS

KEYWORDS

BIOGRAPHY

Andreas Herten is the group leader of the *Accelerating Devices* Lab at Jülich Supercomputing Centre in which he investigates usage of accelerators by HPC applications. Within the *NVIDIA Application Lab at Jülich* he supports scientific users in analyzing, porting, and optimizing their workloads to GPUs. He was the coordinator of the *JUWELS Booster Early Access Program* and has extensive knowledge of the machine. He is involved with various international collaborations for improving performance and scalability on GPUs.

Before working in HPC, Andreas did his PhD in experimental particle physics, developing methods for real-time data suppression on GPUs.

Apart from massively parallel computing with GPUs, Andreas is interested in making HPC more accessible. He teaches different programming courses on GPU computing and beyond.