# Accelerating High Performance and AI Workloads with Slurm and NVIDIA DGX Systems

High-Performance & AI Computing Solution Brief

Today's competitive landscape and disruptive forces put organizations under huge pressure to bring safe, effective products and solutions to market quickly, and at a lower cost, while satisfying rapidly evolving requirements. Current existing computing systems and resources alone can no longer cope with the tighter time constraints and sheer volumes of data needed to generate those solutions and results. Leaders throughout multiple industries are looking to expand available high-performance computing (HPC) resources as well as enable new artificial intelligence (AI) capabilities to address some of the most challenging priorities. Faster processing speeds and AI algorithms are cutting product delivery cycles, reducing treatment and diagnosis times, accelerating fraud detection, and speeding up innovation.

While 84% of executives view AI as critical or extremely critical to their organization's success, only 15% have AI projects fully implemented beyond planning or pilots.[1] Many organizations are in the early stages of HPC and AI adoption or expansion as they face barriers moving experiments to full-scale production operation, including technology complexity and expertise access. Challenges also arise in expanding to additional, integrated cloud resources and meeting demanding requirements for specific HPC and AI workloads. Overcoming these challenges is essential as 79% of executives believe AI will increase efficiency and lower costs, and 73% believe it will increase their ability to introduce new products/services or enter new market areas.[1] The SchedMD and NVIDIA DGX solutions address these complexities to accelerate the desired outcomes.

**Key Go-to-Market Partners to Accelerate Your Success Faster & Easier**

Expanding HPC & AI capabilities with the flexible, purpose-built infrastructure in NVIDIA DGX systems enables organizations to meet exploding data and workload challenges. Delivery times can be reduced for key projects where time is most critical. As a partner in the NVIDIA DGX-Ready Software program, the SchedMD and NVIDIA solution also provides the industry-leading open-source workload manager Slurm, to manage large-scale, complex workloads for faster processing and optimal utilization of the specialized high-performance and AI resource capabilities needed for each workload. Slurm combined with SchedMD expert services maximizes data throughput, reliability, and results in the fastest possible time. SchedMD and NVIDIA have worked together to deliver reference architectures and solutions optimized for key use cases, enabling simplified management and integration so organizations can focus on results and get to them faster and easier.

Market-leading open-source workload manager for the most complex and demanding HPC systems for 10+ years.

+

Slurm developer and services provider supporting and optimizing the speed, throughput, and resource consumption for organizations' unique workload mix.

+

Global leader in AI hardware & software; purpose-built for AI, making adoption faster and simpler for organizations of all sizes.

# Optimize HPC & AI Workloads on NVIDIA DGX

## High-Performance & AI Computing Capabilities Brief

**Slurm provides key scheduling to NVIDIA GPUs**
Slurm manages GPUs similar to CPUs with flexible control for requesting GPUs and binding tasks to the GPU (GPU=first-class resource). Slurm also supports NVIDIA Multi-Instance GPU (MIG), auto-detecting GPU resources and constraining workloads to only the specific allocated GPUs disallowing processes from using more than requested. Slurm sets CUDA_VISIBLE_DEVICES environment variable allowing the job to know the allocated GPU.

**Massive scalability to handle 10-50K+ GPU+node clusters and increase throughput 5-10x.**
Slurm workload manager enables unmatched workload throughput across massive numbers of jobs and massive scale NVIDIA HPC & AI infrastructure resources to deliver innovation and insights faster for a competitive edge. Slurm's scheduling and resource management capabilities handle both effortlessly, including job arrays to submit millions of tasks in milliseconds and ultra granular task allocation by specialized resources (cores, GPUs, threads, etc.). The SchedMD team fine-tunes configurations to workload mix and priorities while improving resource consumption efficiency by 30-40%.

**Proven workload reliability for 20-30% improved uptime.**
Slurm provides the highest levels of dependability to run critical workloads proven by customers who have seen reliability and uptime improvements of 20-30%. Uptime & productivity is further enhanced by SchedMD top support experts who help ensure your HPC system continually processes workload at levels needed as the mix and scale evolve.

**Tailored Slurm & workload expertise knowledge transfer to empower users.**
The SchedMD experts create tailored training based on specific workload use cases, resource and cloud use, and organization needs along with key Slurm fundamentals. Transferring knowledge in hands-on workshops empowers users to harness all the capabilities of Slurm in achieving project priorities.

**Uniquely qualified team expertise improves resolution speed and quality 3-5x.**
The uniquely qualified SchedMD team provides the fast, quality support that top priority projects demand. Our expertise spans software engineering, Slurm development, computer engineering, and systems administration, as well as 5-10x more complex HPC scheduling experience than most organizations can assemble to self-support. 10+years of real-world, large-scale HPC & AI scheduling experience, including half of the biggest systems in the TOP500 and five years working with NVIDIA, enable quick resolution of complex challenges to ensure success. It also means resolutions are done without expertise escalation delays and without causing additional issues.

**An active community & robust development investment delivers market-leading innovation.**
The robust feature set in Slurm provides all the leading capabilities organizations need in one workload management tool with continual open source innovation to address new critical requirements driven by a vibrant user community around the world. SchedMD's strong development investment in Slurm ensures this community innovation is reflected in the latest proven Slurm releases.

SchedMD

**Schedmd.com**
**Slurm.schedmd.com**

**Contact the SchedMD solution team** for more information on how SchedMD and Slurm can optimize and accelerate your specific workloads and results for AI & HPC on NVIDIA DGX systems.

Sales@schedmd.com, Slurm Solutions Specialist: jess.arrington@schedmd.com

*1 Greenstein, B., October 2018, "Making AI Responsible and Effective". Retrieved November 9, 2020.*