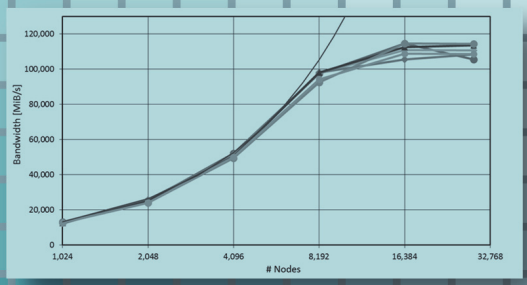

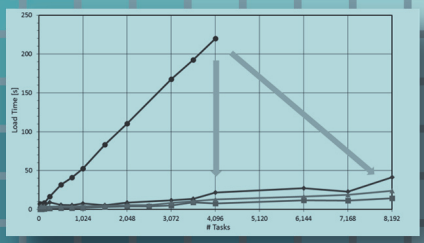

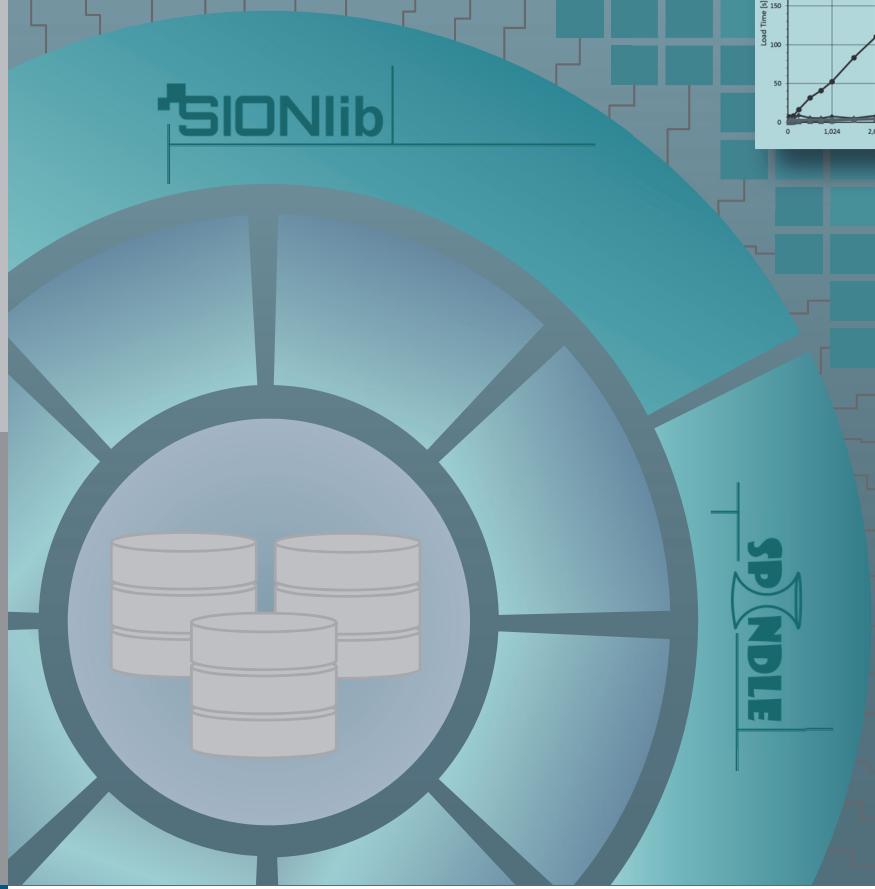

Efficient Task-Local I/O Operations of Massively Parallel Applications

Wolfgang Frings

Member of the Helmholtz Association

JÜLICH
FORSCHUNGSZENTRUM

Forschungszentrum Jülich GmbH
Institute for Advanced Simulation (IAS)
Jülich Supercomputing Centre (JSC)

Efficient Task-Local I/O Operations of Massively Parallel Applications

Wolfgang Frings

Schriften des Forschungszentrums Jülich
IAS Series Band 30

ISSN 1868-8489 ISBN 978-3-95806-152-1

Contents

On current large-scale HPC systems often occur I/O patterns that produce a high load on the file system during access to checkpoint and restart files. Applications running on systems with distributed memory will often perform such I/O individually by creating task-local file objects on the file system. At large scale, these task-local I/O patterns impose substantial stress on the metadata management components of the I/O subsystem. Such metadata contention occurs also at the startup of dynamically linked applications while searching for library files.

The reason for these limitations is that the serial I/O components of the operating system do not take advantage of application parallelism. To avoid the above bottlenecks, this work describes two novel approaches which exploit the knowledge of application parallelism, the underlying I/O subsystem structure, the parallel file system configuration, and the network between HPC-system and I/O system to coordinate and optimize access to file-system objects. The underlying methods are implemented in two tools, SIONlib and Spindle, which add layers between the parallel application and the corresponding POSIX-based standard interfaces of the operating system, eliminating the need for modifying the underlying system software.

SIONlib is already applied in applications to implement efficient checkpointing and is also integrated in the performance-analysis tools Scalasca and Score-P to efficiently store trace data. Latest benchmarks on the Blue Gene/Q in Jülich demonstrate that SIONlib solves the metadata problem at large scale by running efficiently up to 1.8 million tasks while maintaining high I/O bandwidths of 60-80% of file-system peak with a negligible file-creation time. The scalability of Spindle could be demonstrated by running a benchmark on a cluster of Lawrence Livermore National Laboratory at large scale. The results show that the startup of dynamically linked applications is now feasible on more than 15000 tasks, whereas the overhead of Spindle is nearly constantly low.

With SIONlib and Spindle, this work demonstrates how scalability of operating system components can be improved without modifying them and without changing the I/O patterns of applications. In this way, SIONlib and Spindle represent prototype implementations of functionality needed by next-generation runtime systems.

This publication was edited at the Jülich Supercomputing Centre (JSC) which is an integral part of the Institute for Advanced Simulation (IAS). The IAS combines the Jülich simulation sciences and the supercomputer facility in one organizational unit. It includes those parts of the scientific institutes at Forschungszentrum Jülich which use simulation on supercomputers as their main research methodology.

IAS Series
Volume 30
ISBN 978-3-95806-152-1

JÜLICH
FORSCHUNGSZENTRUM