JÜLICH
FORSCHUNGSZENTRUM

I/O at JSC

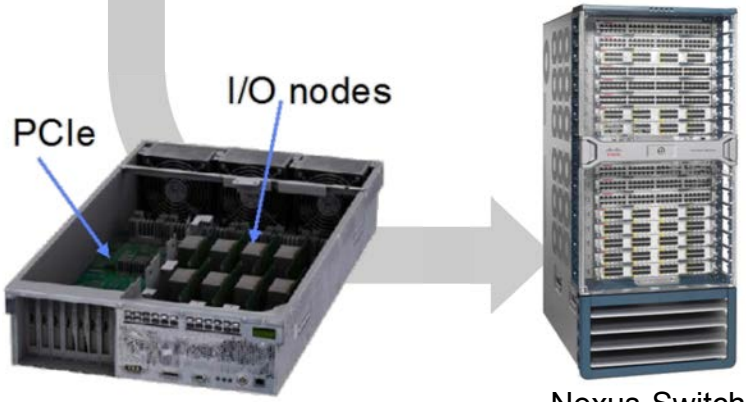
- → I/O Infrastructure
- → Workloads, Use Case
- → I/O System Usage and Performance
- → SIONlib: Task-Local I/O

Wolfgang Frings
W.Frings@fz-juelich.de
Jülich Supercomputing Centre

HPC I/O in the Data Center (HPC-IODC), ISC, Frankfurt, July 16th, 2015

JUQUEEN: Jülich's Scalable Petaflop System

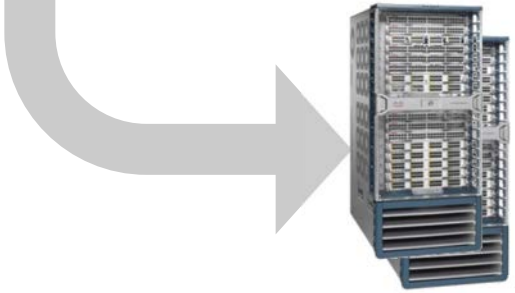
- IBM Blue Gene/Q JUQUEEN
- IBM PowerPC® A2 1.6 GHz,
  16 cores per node
  28 racks (7 rows à 4 racks)
  28,672 nodes (**458,752 cores**)
- 5D torus network
- 5.9 Pflop/s peak
  5.0 Pflop/s Linpack
- Main memory: **448 TB**
- **I/O Nodes**: **248** (27x8 + 1x32)
- **Network**: 2x CISCO Nexus 7018
  Switches (connect I/O-nodes)
  Total ports: **512 10 GigEthernet**



Nexus-Switch

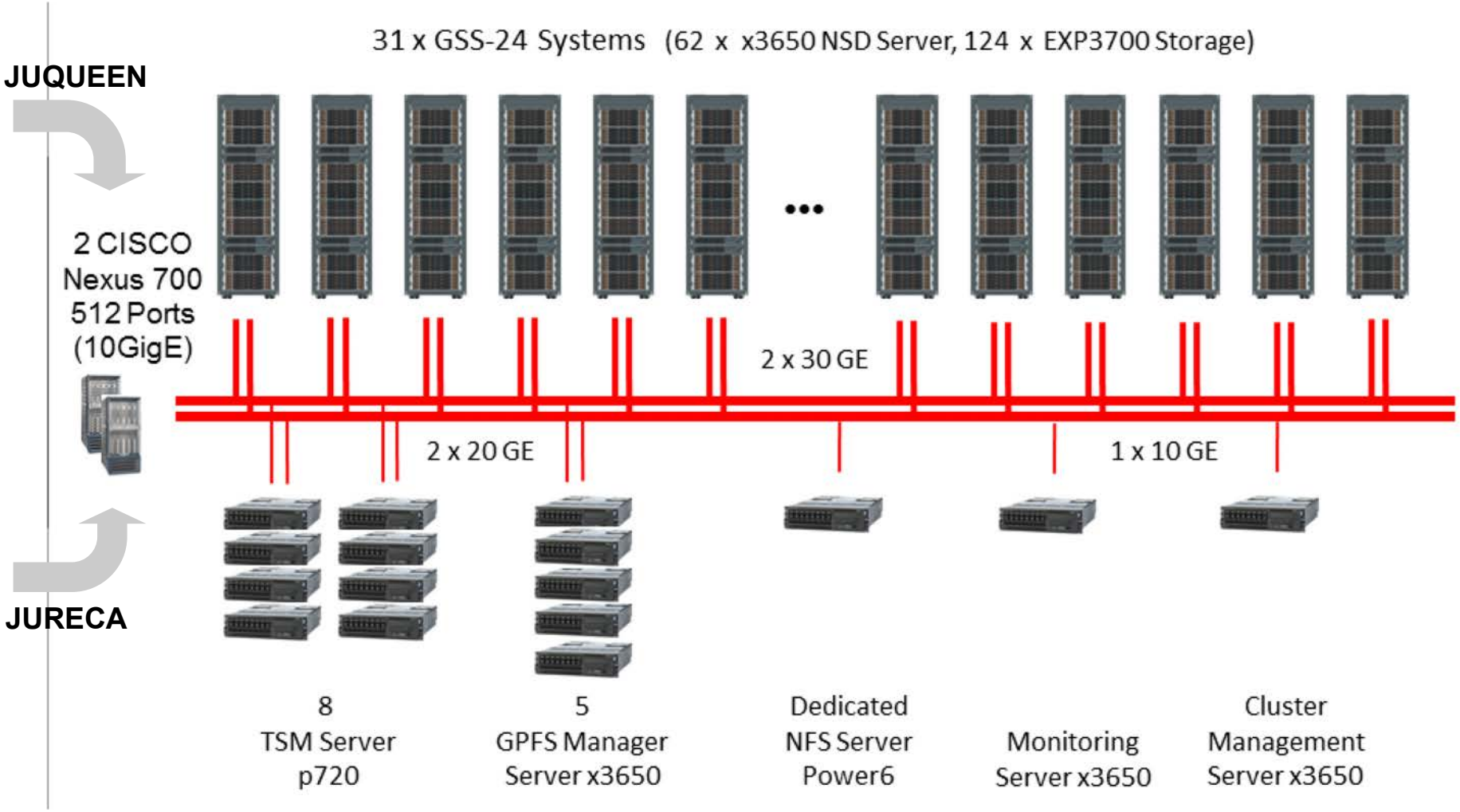
JURECA: Jülich Research on Exascale Cluster Architectures

- 2 Intel Haswell 12-core processors, 2.5 GHz, SMT, 128 GB main memory
- **1,884 compute nodes** or 45,216 cores, thereof
  - 75 nodes with 2 K80 NVIDIA graphics cards each and
  - 12 nodes with 512 GB main memory and 2 K40 NVIDIA graphics cards each for visualisation
- 2.245 Petaflop/s peak (with K80 graphics cards)
- **281 TByte memory**
- Mellanox Infiniband EDR
- Connected to the GPFS file system on JUST (IB/10GigE)
- Installation: 2015



CISCO Nexus 700 512 Ports (10GigE)

JUQUEEN and JUST I/O-Network



JUST4-GSS

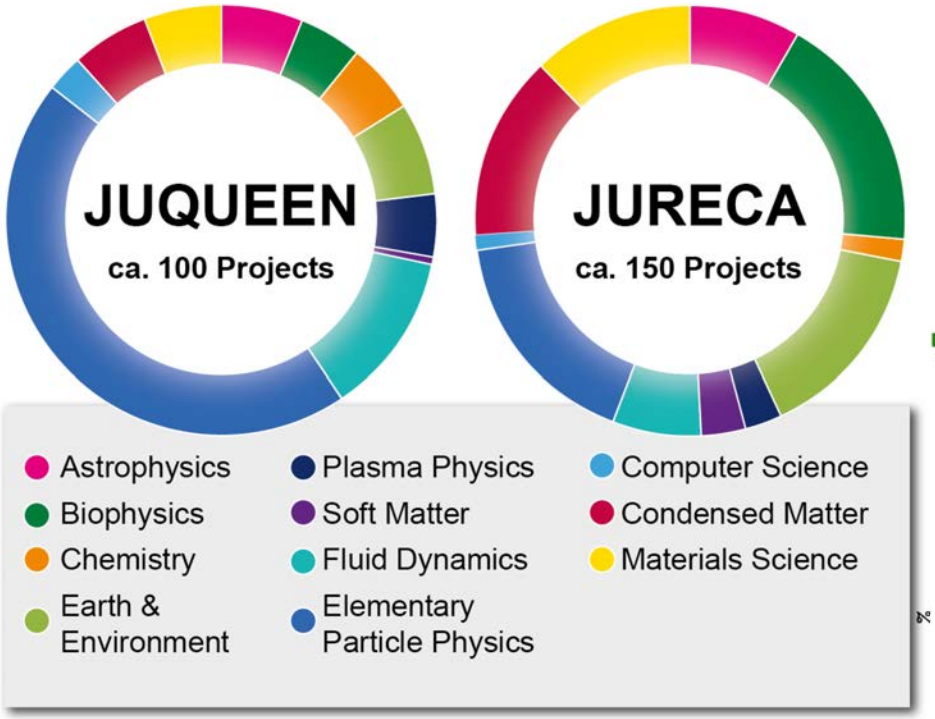
Parallel I/O Hardware at JSC (Just4, GSS)

- **Ju**elich **St**orage Cluster (JUST)
  - GPFS Storage Server (GSS/ESS)
  - End-to-End integrity
  - Fast rebuild time on disk replacement
  - GPFS + TSM Backup + HSM
- **Just4-GSS**
  - Capacity: **12.6 Pbyte**
    I/O Bandwidth: up to **200 GB/sec**
  - Hardware: IBM System x® GPFS™ Storage Server solution, GPFS Native RAID
  - 31 Building blocks: each 2 x X3650 M4 server, 232 NL-SAS disks (2TB), 6 SSD

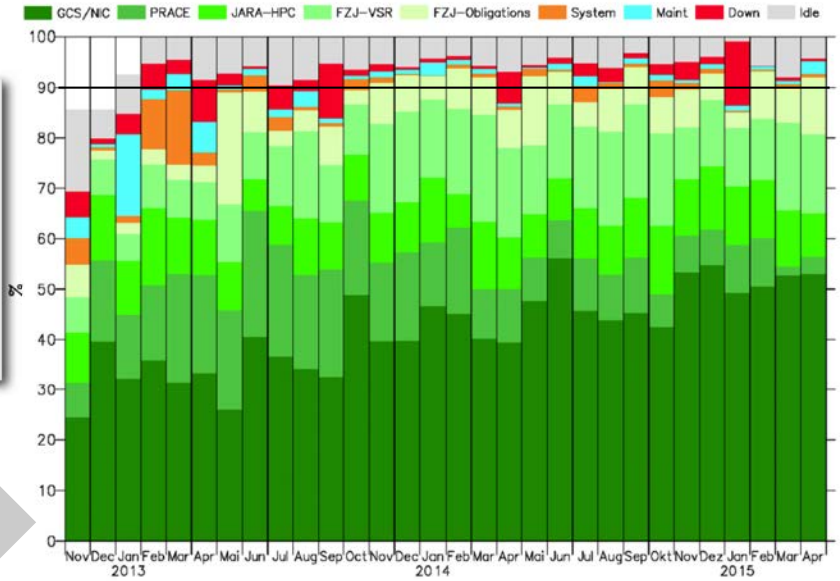
Workload: Applications

Leadership-Class System

General-Purpose Supercomputer

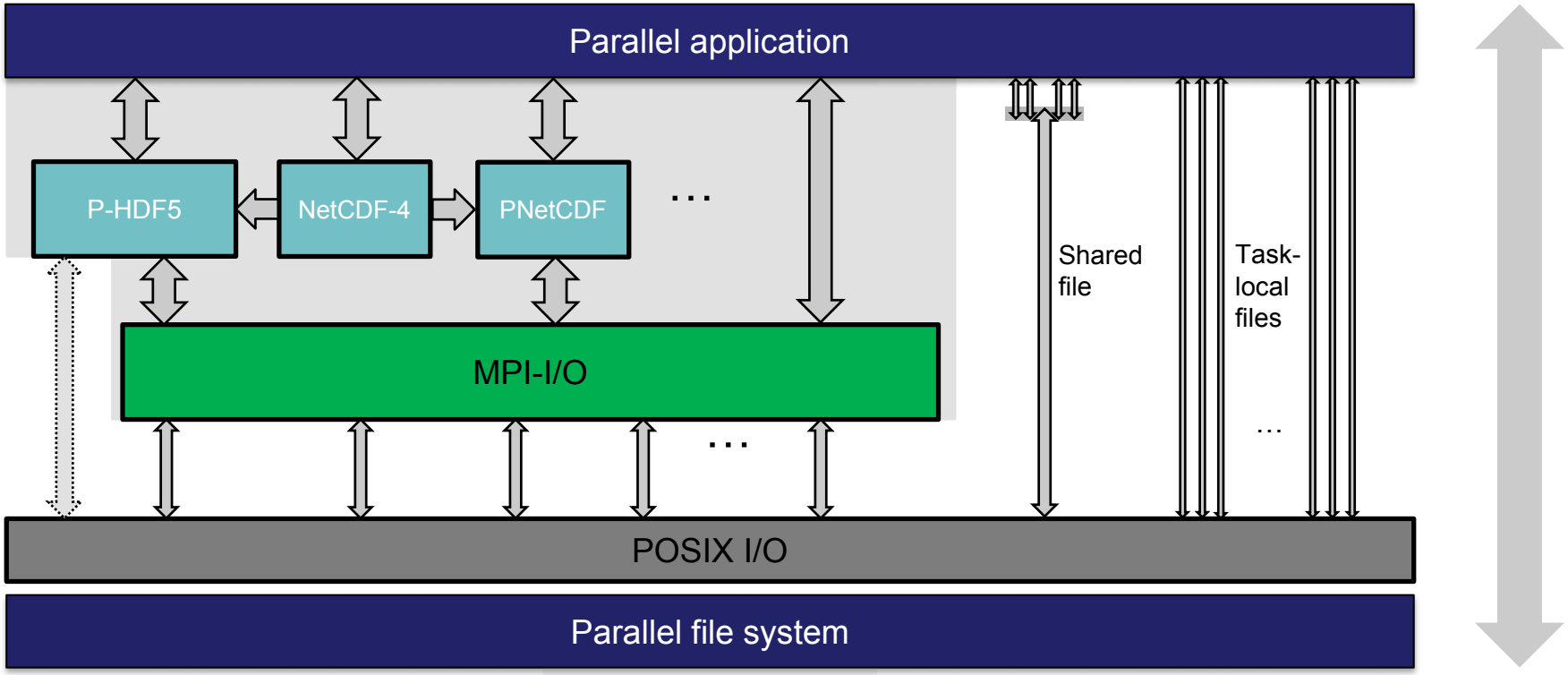


- Granting periods
- 05/2015 – 04/2016
- 11/2014 – 10/2015



JUQUEEN

Workload: I/O Libraries



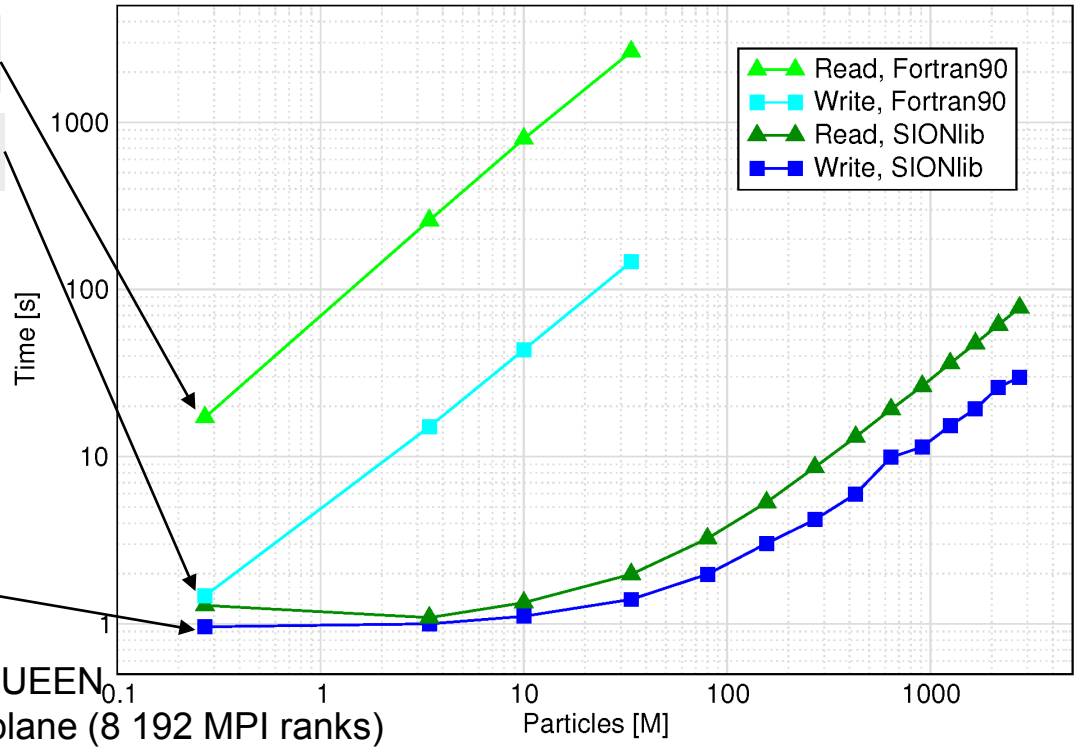
Use Case: MP2C

Parallel MD-simulation: *couples multiple particle collision dynamics with molecular dynamics to implement mesoscale simulation of hydrodynamic media → particle I/O*

Fortran read (data sieving)

Fortran write from one task

Shared File parallel I/O with SIONlib

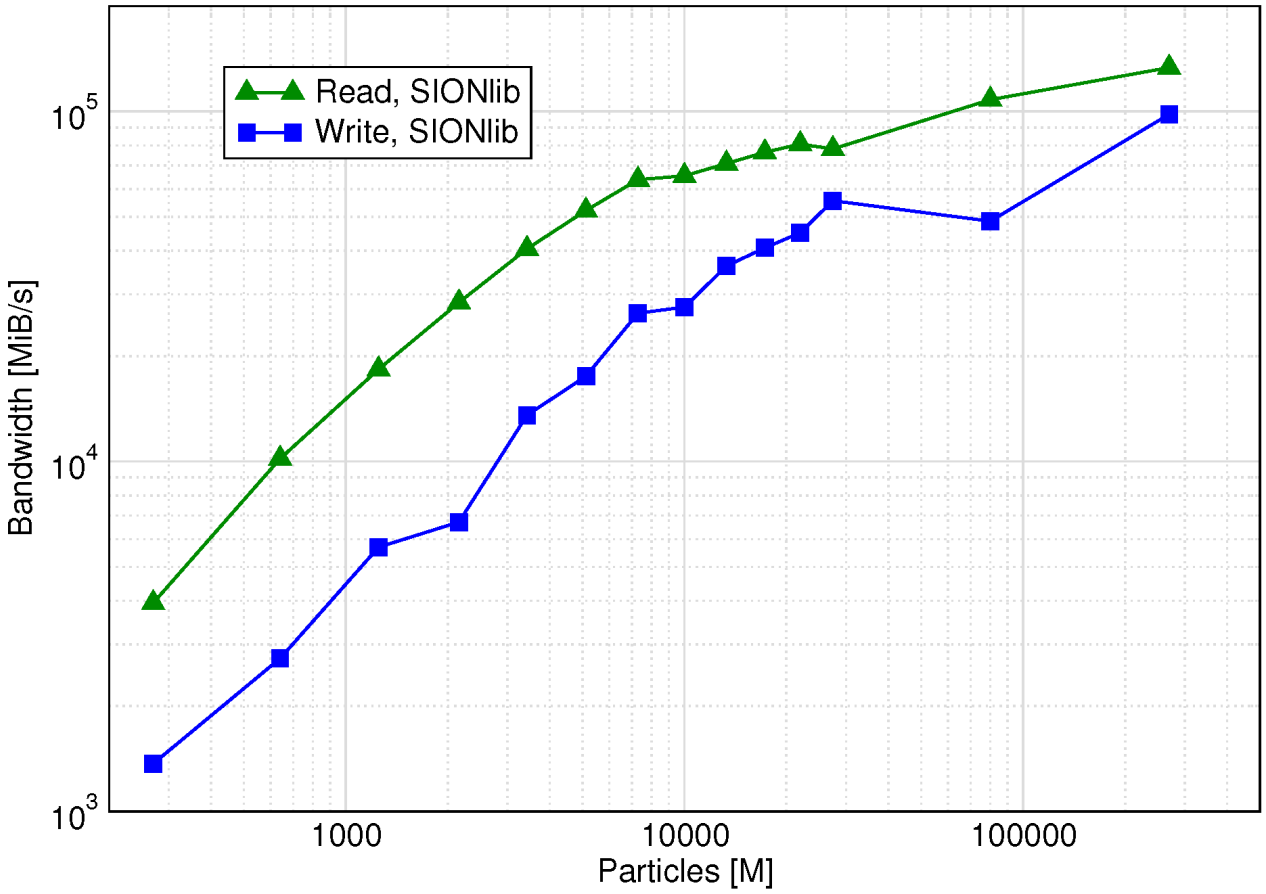


JUQUEEN midplane (8 192 MPI ranks)

http://www.fz-juelich.de/ias/jsc/EN/Expertise/High-Q-Club/MP2C/_node.html

Use Case: MP2C (full scale)

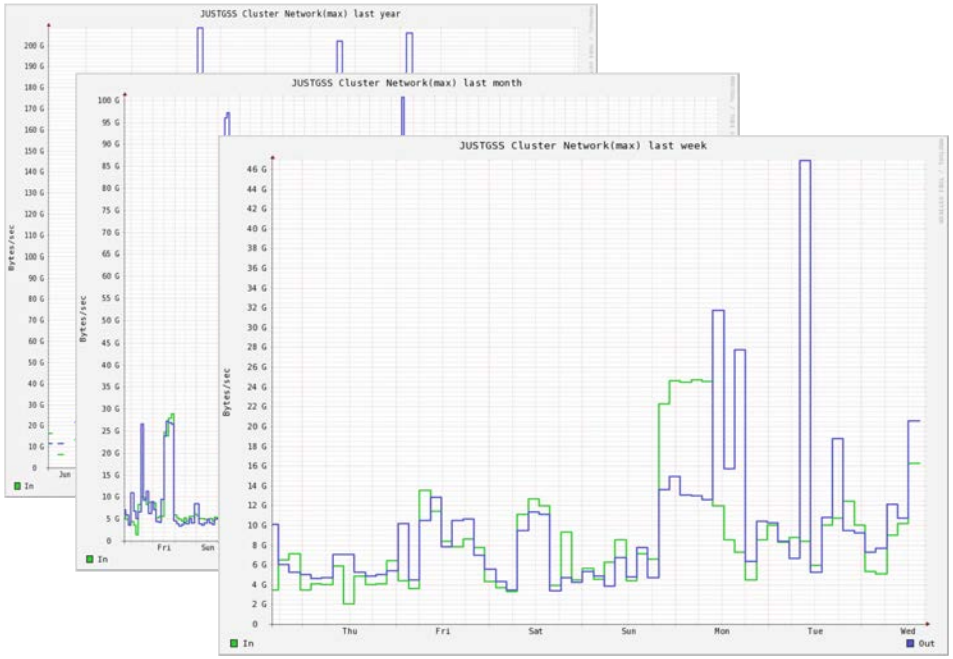
→ Shared File parallel I/O with SIONlib (coalescing I/O, multi-file)



MP2C: I/O bandwidth for checkpointing executions with 1.8 million MPI ranks on 28 racks of JUQUEEN using SIONlib for reading and writing particle data

I/O Monitoring: Ganglia on JUST

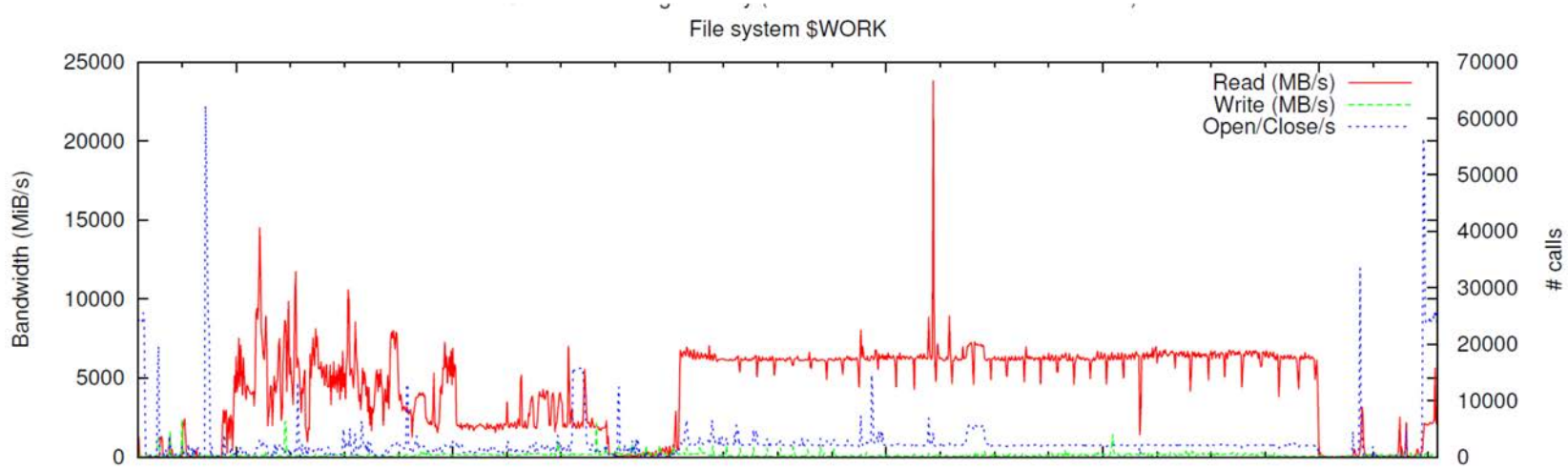
- Ganglia network monitoring on JUST GPFS cluster
  - Overall network load
  - Per file-server network load



I/O Monitoring: GPFS MMPMON

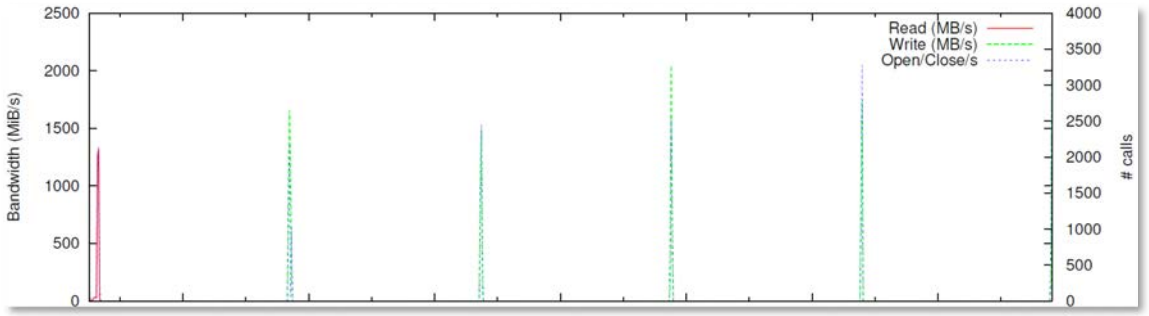
- GPFS daemon log: one snapshot/min,
  metrics → bytes written/read, open/close operations
- I/O-activity on JUQUEEN (I/O nodes) and JURECA

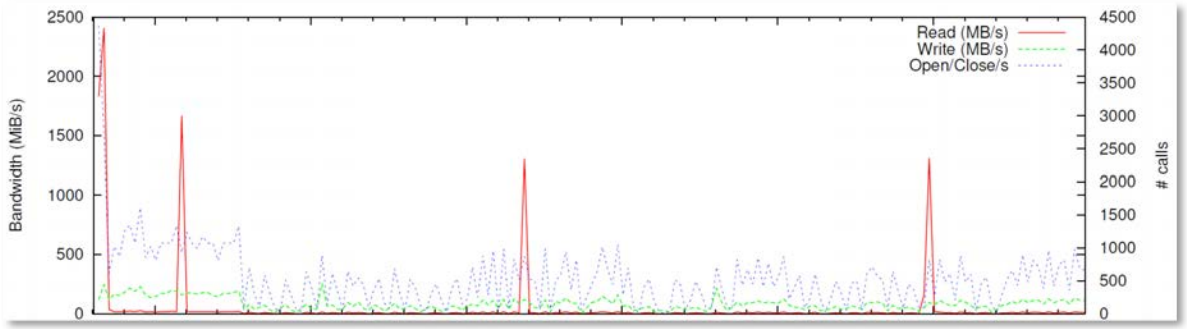
→ Example: Aggregate I/O Usage JUQUEEN

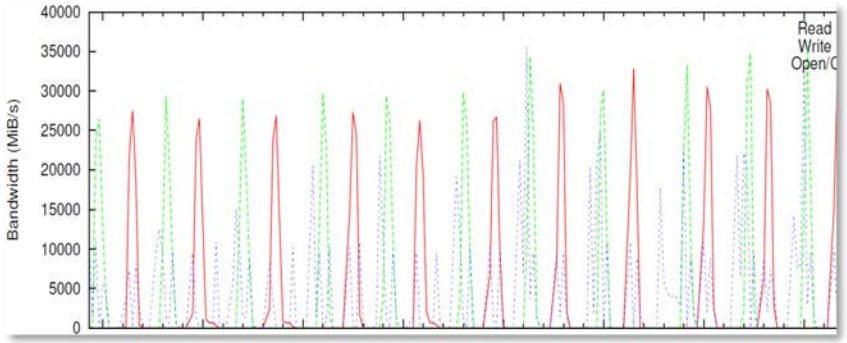


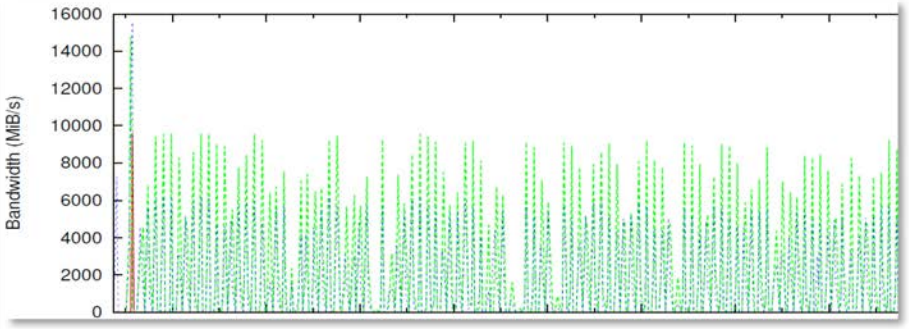
I/O Monitoring: GPFS MMPMON

→ Examples:
Job-based
I/O-Usage
on JUQUEEN

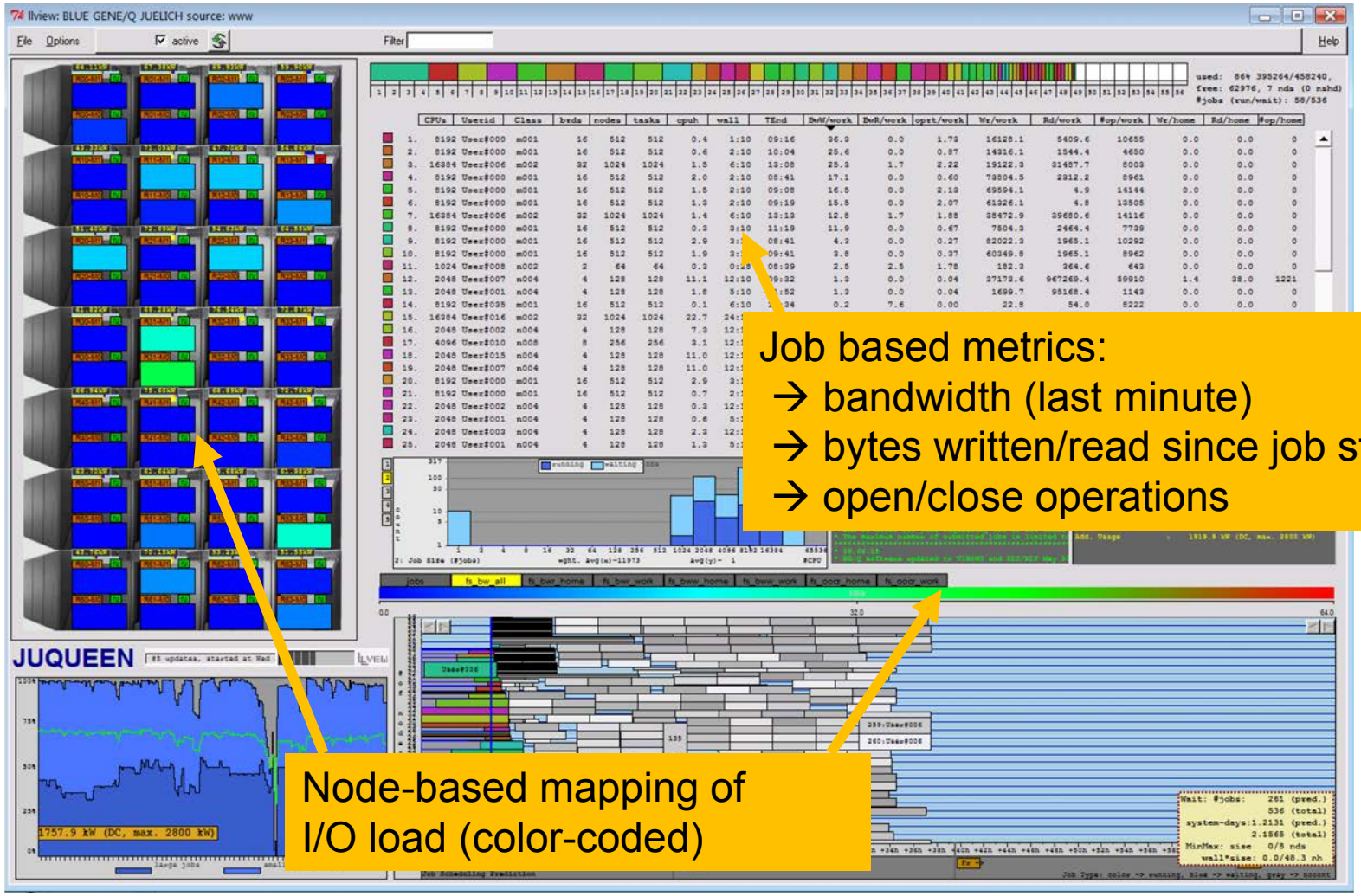






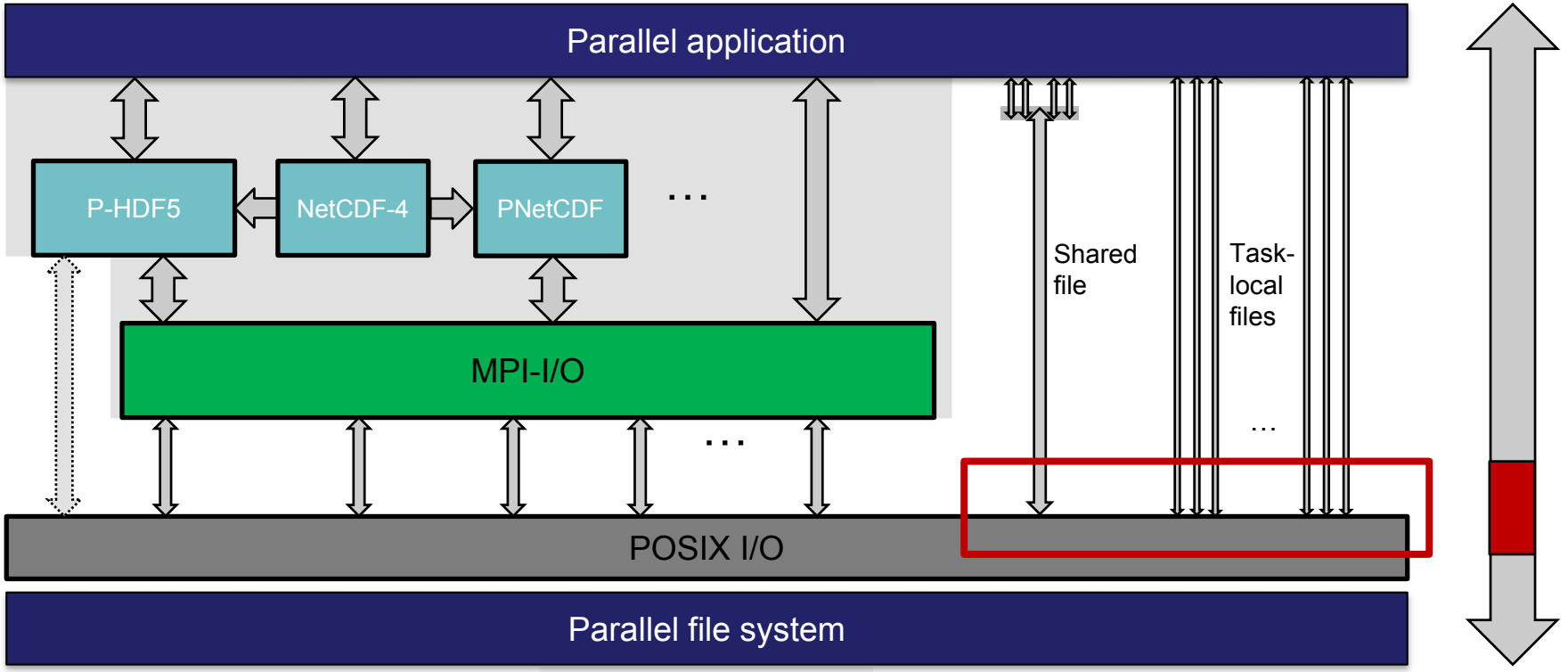


I/O Monitoring: LLview & GPFS mmpmon

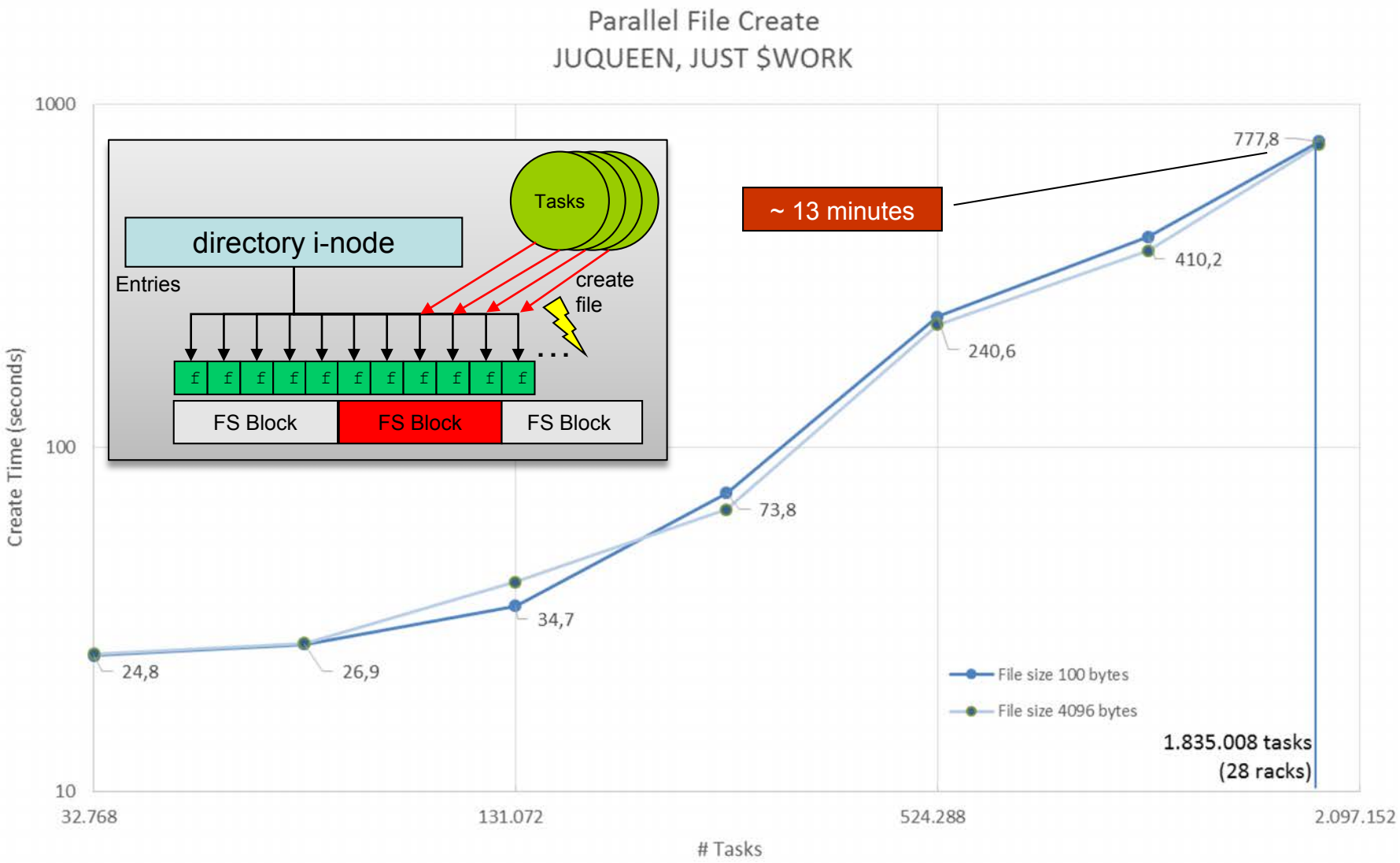


http://www.fz-juelich.de/jsc/llview

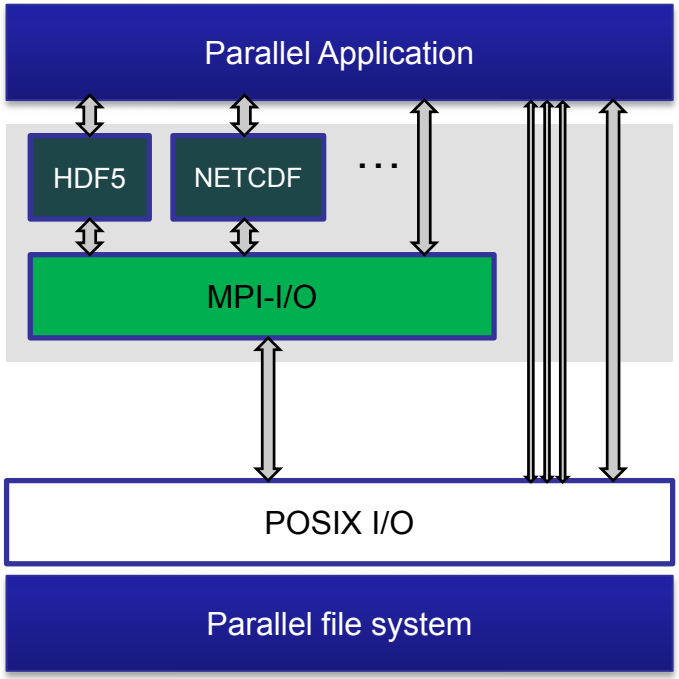
Workload: I/O Libraries

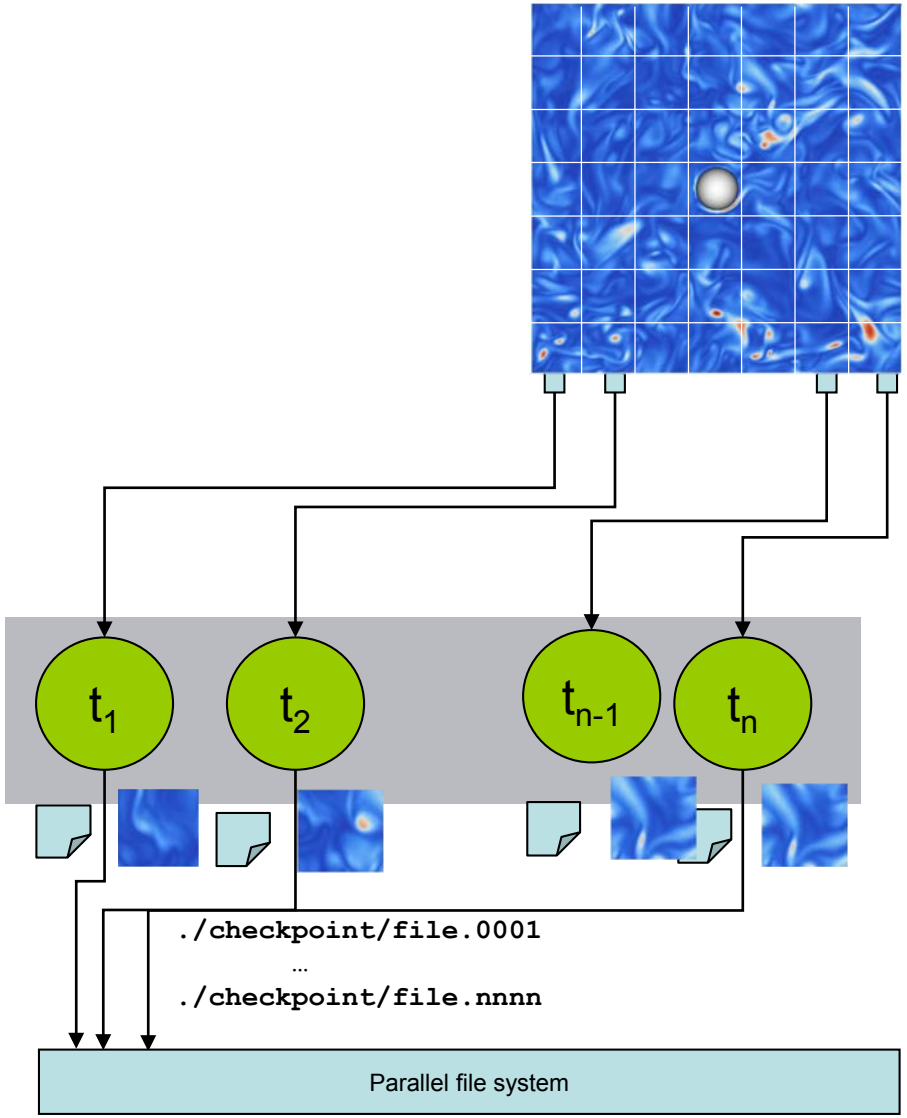


I/O Bottleneck: File Creation



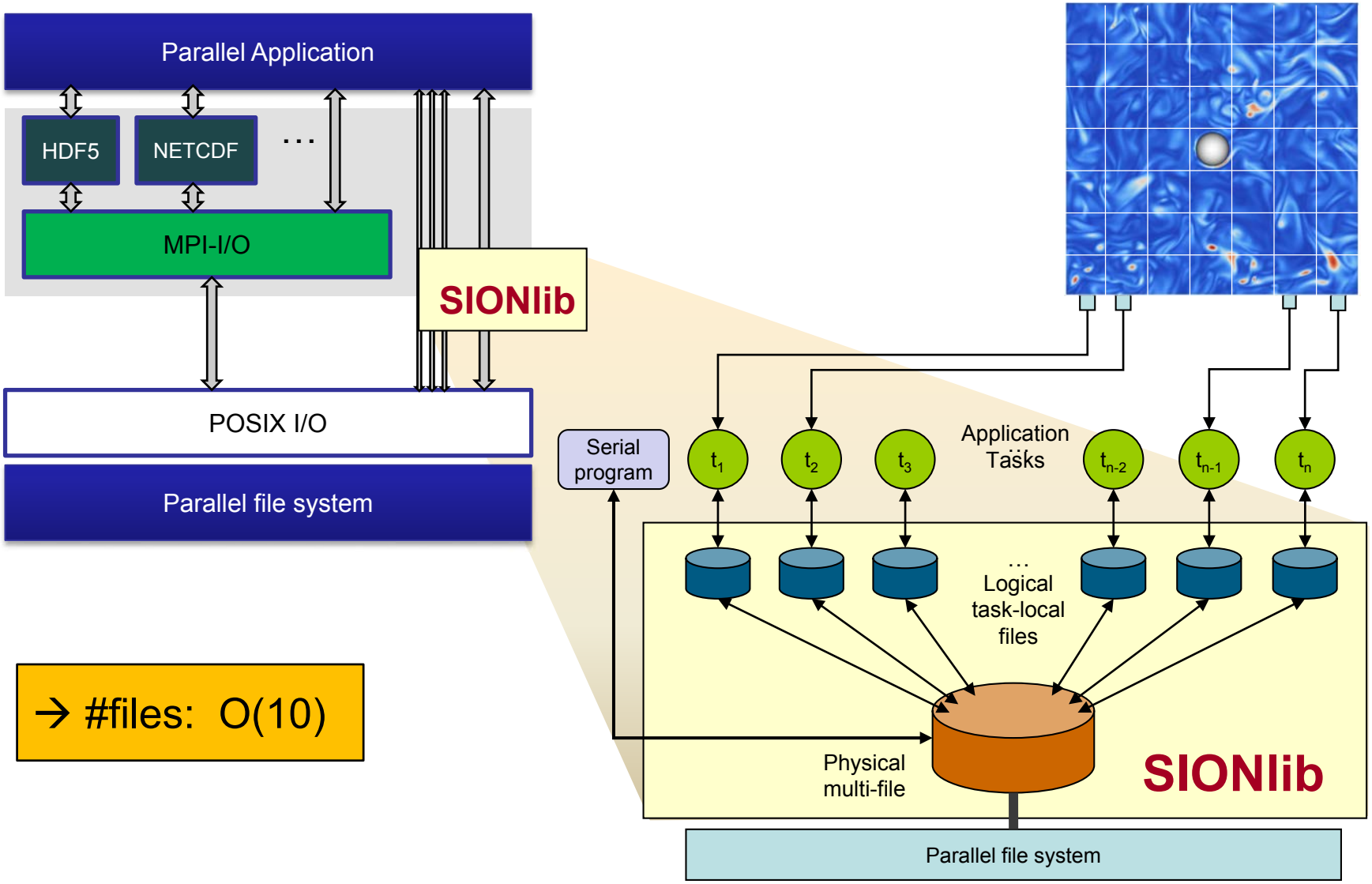
SIONlib: Shared Files for Task-local Data





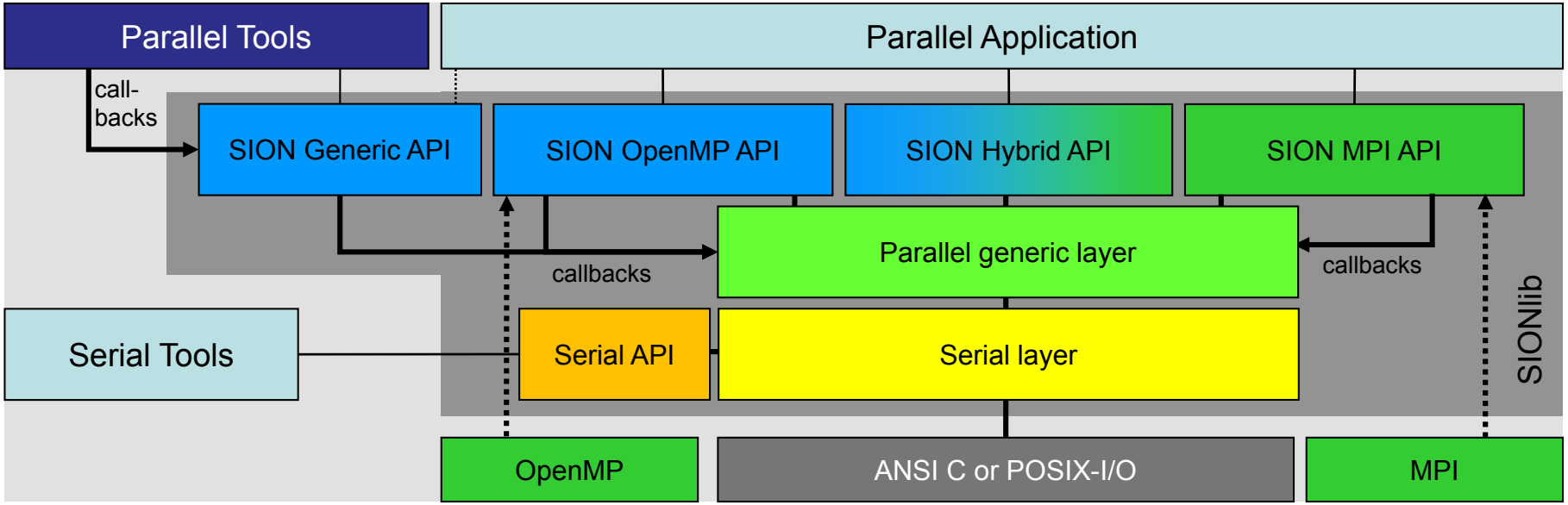
→ #files: O(10)

SIONlib: Shared Files for Task-local Data



→ #files: O(10)

SIONlib: Architecture & Example



- Extension of I/O-API (ANSI C or POSIX)
- C and Fortran bindings, implementation language C
- Current versions: 1.5.5, 1.6rc
- Open source license: http://www.fz-juelich.de/jsc/sionlib

```
/* fopen() → */
sid=sion_paropen_mpi( filename , "bw",
                      &numfiles, &chunksize,
                      gcom, &lcom, &fileptr, ...);

/* fwrite(bindata,1,nbytes, fileptr) → */
sion_fwrite(bindata,1,nbytes, sid);

/* fclose() → */
sion_parclose_mpi(sid)
```

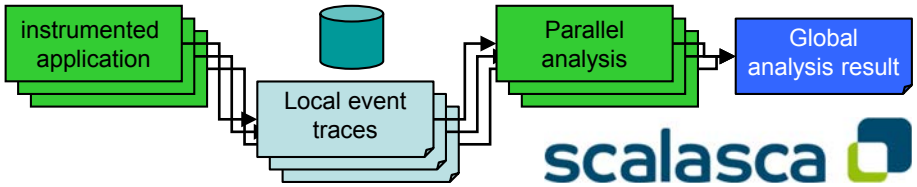
SIONlib: Applications

- **Applications**

DUNE-ISTL (Multigrid solver, Univ. Heidelberg) **ITM** (Fusion-community),
LBM (Fluid flow/mass transport, Univ. Marburg), **PSC** (particle-in-cell code),
OSIRIS (Fully-explicit particle-in-cell code), **PEPC** (Pretty Efficient Parallel C. Solver)
Profasi: (Protein folding and aggr. simulator) **NEST** (Human Brain Simulation)
MP2C: (Mesoscopic hydrodynamics + MD)

- **Tools/Projects**

Scalasca: Performance Analysis



Score-P: Scalable Performance Measurement Infrastructure for Parallel Codes

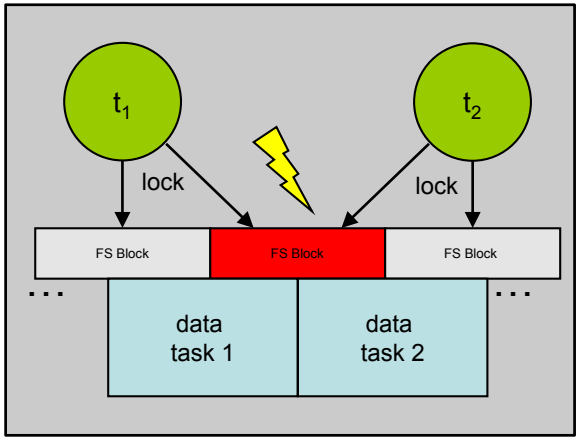
DEEP-ER: Adaption to new platform and parallelization paradigm
Buddy-Checkpointing

I/O-Benchmarking: Concurrent Access & Contention

File System Block Locking → Serialization

SIONlib: Logical partitioning of Shared File:

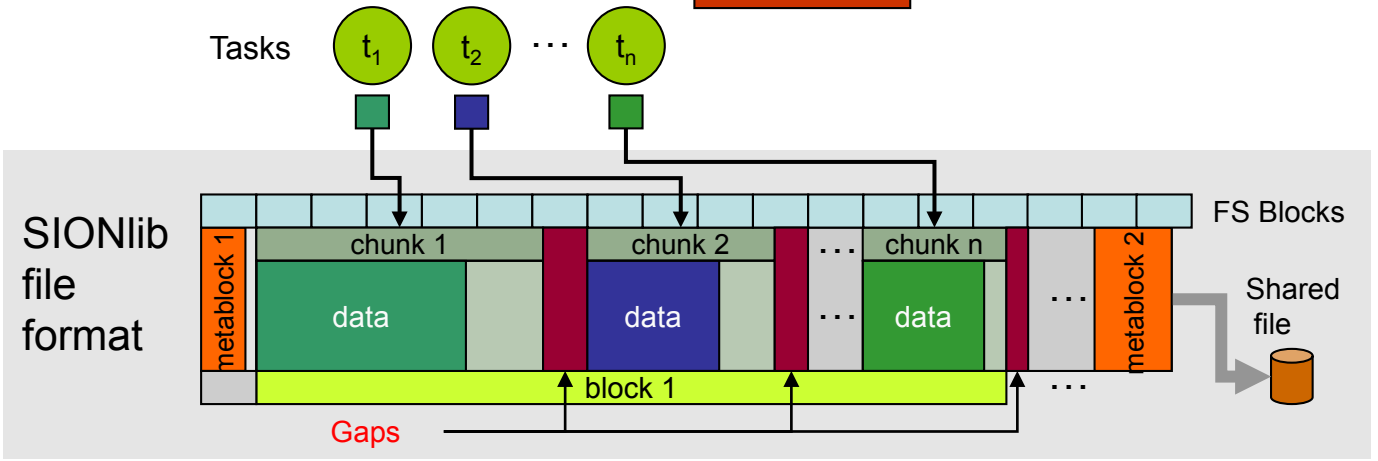
- Dedicated data chunks per task
- Alignment to boundaries of file system blocks → no contention



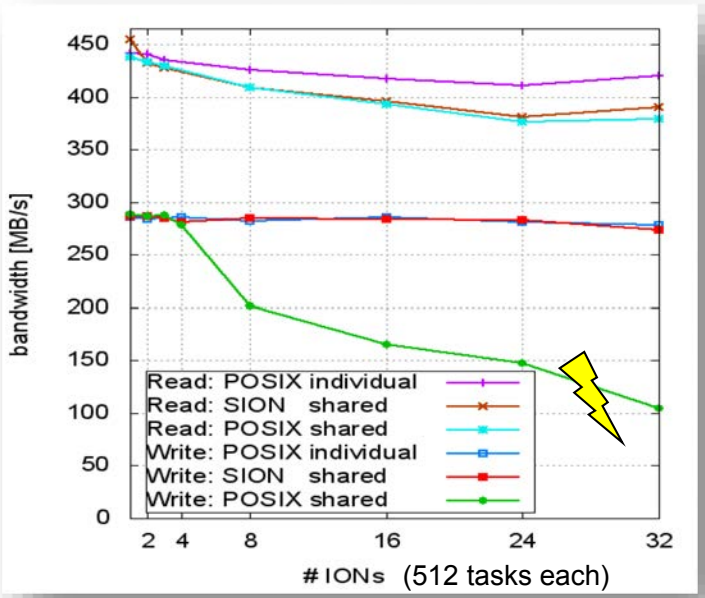
#tasks	data size	blksize	write bandwidth
32768	256 GB	aligned	**3650.2** MB/s
32768	256 GB	not aligned	**1863.8** MB/s

Jugene (JSC, IBM Blue Gene/P, GPFS, fs:work)

2x faster

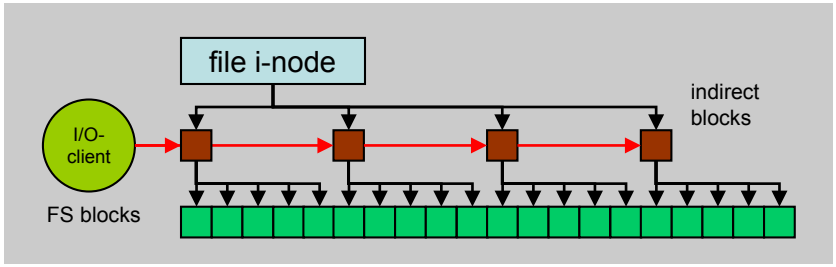


I/O-Benchmarking: Increasing #tasks further …



JUGENE: Bandwidth per ION, comparison individual files (POSIX), one file per ION (SION) and one shared file (POSIX)

- Bottleneck: file meta data management
- by first GPFS client which opened the file

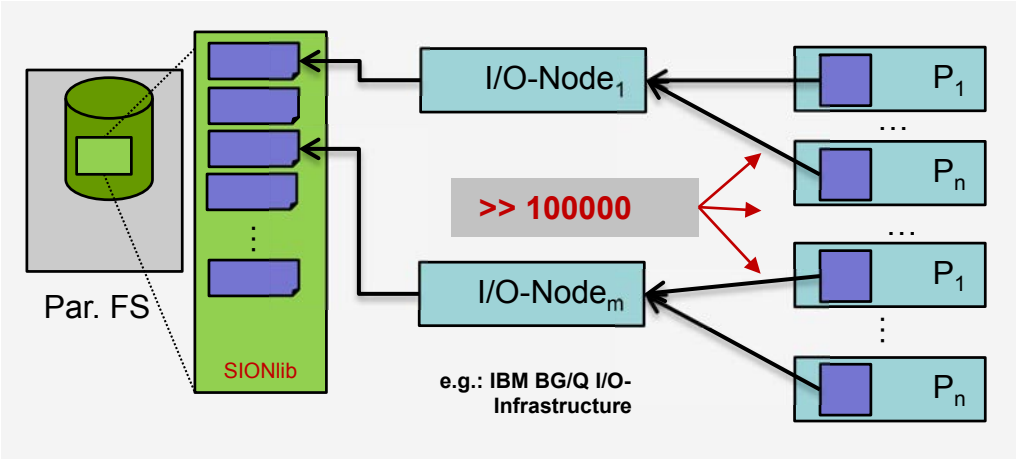


→ Parallelization of file meta data handling using multiple physical files

→ Mapping: Files : Tasks

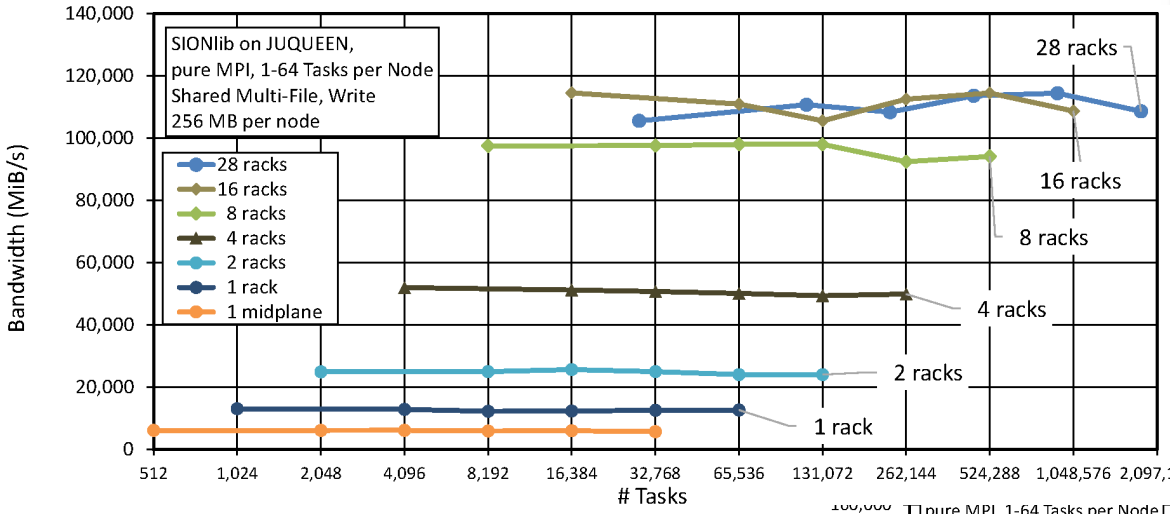
1 : n ← **p : n** → n : n

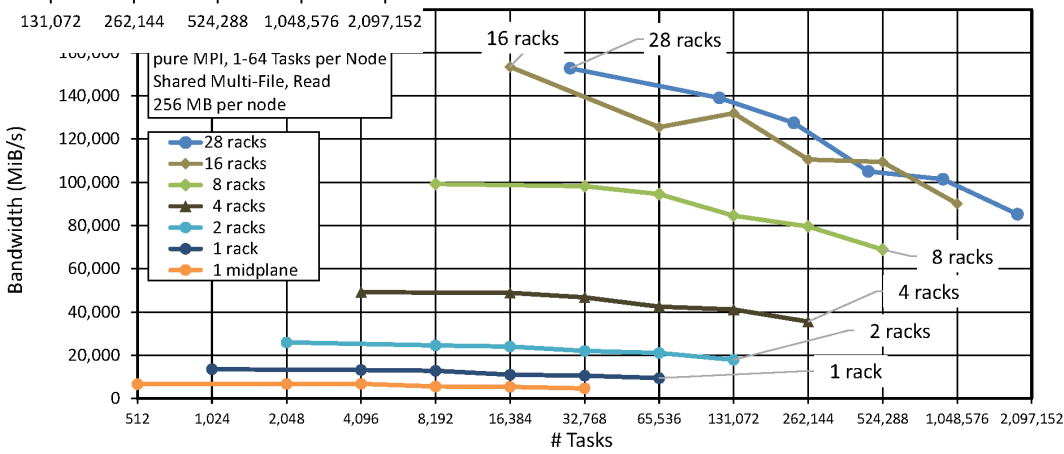
→ IBM Blue Gene:
One file per I/O-node (locality)



JUQUEEN SIONlib Scaling

- JUQUEEN (BG/Q) → JUST (GPFS/GSS)
- Benchmark: 1.8 million tasks, ~7 TiB
- → 50 – 70% of peak I/O bandwidth
- Multi-file approach: one file per I/O-bridge





Conclusion

I/O workloads

- Large number of applications → diversity of I/O usage
- Optimization of parallel I/O library on systems

Hardware & I/O infrastructure

- JUQUEEN → Hierarchical I/O infrastructure
- JUST → Shared file system for multiple HPC systems

I/O monitoring

- Combination of information from different sources

  →LLview + GPFS mmpmon

I/O challenges

- Large number of tasks/threads in parallel I/O
- Support task-local I/O → SIONlib