The Effective I/O Bandwidth Benchmark (b_eff_io)

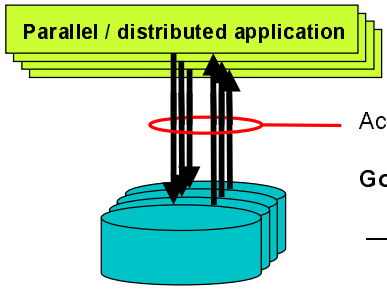
Rolf Rabenseifner
High-Performance Computing-Center Stuttgart (HLRS), University of Stuttgart,
rabenseifner@hlrs.de *www.hlrs.de/people/rabenseifner*
Alice E. Koniges
Lawrence Livermore National Laboratory,
koniges@llnl.gov *www.rzg.mpg.de/~ack*

Message Passing Interface and High-Performance Clusters
Developer's and User's Conference (MPIDC 2000)
Cornell University, Ithaca, March 20–23, 2000
www.mpidc.org

Content

- Our goals
- Starting-points
- Definition of the b_eff_io benchmark
- Results
- Outlook
- Summary

Effective I/O Bandwidth Benchmark



Accumulated I/O bandwidth ?

Goal: one characteristic value!

— like Rmax in TOP500 —

Starting-Points

- Application benchmarks
  - using real, I/O-intensive applications
- Filesystem benchmarks
  - measuring several parameters around the most friendly disk-usage-pattern
- Hardware benchmarks
  - maximum bandwidth of the disk — special-benchmark
- Let's meet in the middle
  - MPI-I/O based benchmark
    - portable
    - platform independent

Starting-Points — the I/O Parameter Space

- How to define and measure one characteristic I/O bandwidth value?
- The I/O parameter space — 20 orthogonal parameters:
  - **Application parameters:** (a) the size of contiguous chunks in the memory, (b) the size of contiguous chunks on disk, which may be different in the case of scatter/gather access patterns, (c) the number of such contiguous chunks that are accessed with each call to a read or write routine, (d) the file size, (e) the distribution scheme, e.g., segmented or long strides, short strides, random or regular, or separate files for each node, and (f) whether or not the chunk size and alignment are wellformed, e.g., a power of two or a multiple of the striping unit.
  - **Usage aspects:** (a) how many processes are used and (b) how many parallel processors and threads are used for each process.
  - **I/O interface:** Posix I/O buffered or raw, special filesystem I/O of the vendor filesytem, or MPI-I/O.
  - **MPI-I/O aspects:** (a) access methods, i.e., first writing of a file, rewriting or reading, (b) positioning method, i.e. explicit offsets, individual or shared file pointers, (c) coordination, i.e., accessing the file collectively by (all) processes or noncollectively, (d) synchronism, i.e., blocking or nonblocking. Additional aspects are: (e) whether or not the files are open unique, i.e., that the file will not be concurrently opened elsewhere, and (f) which consistency is chosen for conflicting accesses, i.e., whether or not atomic mode is set.
  - **Filesystem parameters:** (a) how many nodes or processors are used as I/O servers, (b) how much memory is used as bufferspace on each application node, (c) the disk block size, (d) the striping unit size, and (e) the number of parallel striping devices that are used.

Starting-Points — the I/O Parameter Space

- Measuring 10 values in each orthogonal dimension
  —> 10^{20} exciting measurements
    —> your benchmark will run the next decade

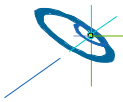
- Using ***one basic, central parameter set***
  —> varying each single parameter
    —> only 10*20 measurements
      —> fine, you get your results in ~ 1 hour
        —> but, vendors *Xray* and *Ybm* do not like the ***one basic, central parameter set*** preferred by *Zun*, or ?–)

Interrupt — b_eff

- b_eff — the effective communication bandwidth benchmark

  Authors: Karl Solchenbach, Hans-Joachim Plum and Gero Ritzenhoefer, Rolf Rabenseifner
- www.hlrs.de/mpi/b_eff
- One characteristic accumulated communication bandwidth number
  := average bandwidth on several communication patterns*)
  average on different message sizes
  maximum over different MPI programming techniques

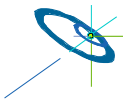
  *) each process sends a message to another one, each pattern is defined as a set of processor-pairs
- needs a few minutes



Interrupt — Existing I/O Benchmarking Papers

- Existing I/O benchmarking papers:

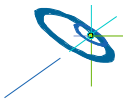
  e.g., Terry Jones, Alice Koniges, R. Kim Yates,
  Performance of the IBM General Parallel File System,
  to be published in Proceedings of the International Parallel and Distributed Processing Symposium, May 2000. Also available as UCRL JC135828
- many hours of dedicated benchmarking time is used
- characterizing a specific system
- not portable
- Rule: Balanced HPC systems should be able to write the total memory in 10 minutes to disk



Starting-Points

— back to b_eff_io —

- What can we learn from the communication benchmark b_eff?
- How can we reduce I/O benchmarking to 15 minutes?
- How can b_eff_io determine
  - how one system work
  - with respect to another?
- Comparing parallel I/O on many systems!



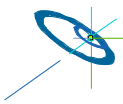
b_eff / b_eff_io

b_eff:

- the repeating factors are controlled between:
  - large values, to reduce time measurement errors
  - small values, to reduce the total execution time

b_eff_io:

- time driven approach
  - how many bytes can I write/read in 5 minutes, using different access patterns



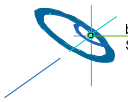
b_eff / b_eff_io

b_eff:

- Major results: **b_eff** *and* **b_eff / number of processors**
- Minor results: accumulated bandwidth for different communication patterns, message sizes, ...

b_eff_io:

- Major results: **b_eff_io** *and* **total memory size / b_eff_io**
- Minor results: different partition sizes (PEs), pattern types, buffer sizes, ...



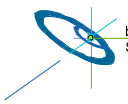
b_eff / b_eff_io

b_eff:

- Maximum over
  - MPI configuration parameters
  - MPI communication methods (Isend, Sendrecv, Alltoallv)

b_eff_io:

- Maximum over
  - Usage aspects:
    - **how many processes are used and**
    - **how many parallel processors and threads are used for each process.**
  - Filesystem parameters
  - Scheduled execution time (≥ 10 minutes)



Definition of b_eff_io

(Release 0.5)

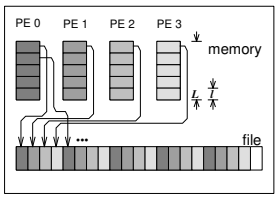
b_eff_io := Maximum over all usage and filesystem parameter } manually

Average on write, rewrite, read
Average on five access pattern types
Average on several patterns*)
of bandwidth of each pattern
} auto-matically, in time T

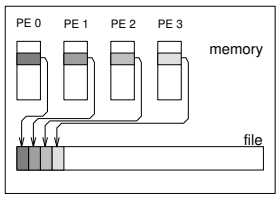
*) defined by the size of contiguous chunks written from memory to disk and the contiguous chunk in memory written by each MPI call

The average is weighted by time units defined for each pattern.

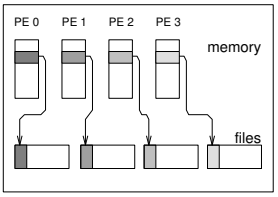
Definition of b_eff_io — the Pattern Types



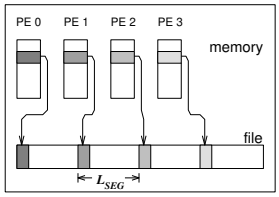
type 0: scattering *)



type 1: strided coll.



type 2: noncoll., separated



type 3 noncoll. / 4 coll. segmented

Pattern that can be optimized

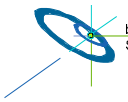
Chunk sizes on disk:
- max (2MB, memory of one node/128) *)
- wellformed: 1MB, *) 32 kB, 1 kB,
- non-wellformed: 1MB+8B, *) 32 kB+8B, 1kB+8B

*) double weighted

Definition of b_eff_io — Bandwidth measurement

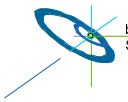
- Bandwidth measurement

```
MPI_Barrier()
start_time = MPI_Wtime()   at root only
repeat
   MPI_File_write() or MPI_File_read()
   MPI_Barrier()
   conti = (MPI_File_read()- start_time) < time_unit
   MPI_Bcast(conti)
while conti
if (write access) MPI_File_sync
MPI_Barrier()
end_time = MPI_Wtime()   at root only
bandwidth = (accumulated data size)
            / (end_time - start_time)
```

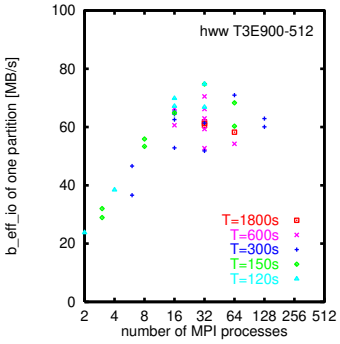


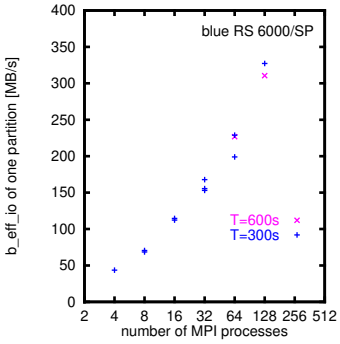
First Results — Comparing two systems

- Cray T3E 900-512 at HLRS/RUS, Stuttgart
  - 512 processors
  - 10 striped Raid-disks, connected via GigaRing
  - mpt.1.3.0.2 with ROMIO, modified: using asynchronous I/O
    - **www.hlrs.de/mpi/mpi_t3e.html#StripedIO**
    - **www.hlrs.de/mpi/ufs_t3e/**
  - theoretical peak throughput = 300 MB/s
- IBM RS 6000/SP at LLNL, called "blue"
  - 336 SMP nodes with each 4 processors
  - benchmark: using 1 processor per node
  - IBM General Parallel File System (GPFS) with 20 VSD I/O server
  - ROMIO
  - measured peak performance: 950 MB/s read, 690 MB/s write (on 128 nodes)



First Results — Comparing b_eff_io (#processes)

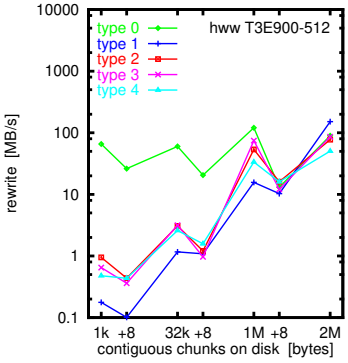


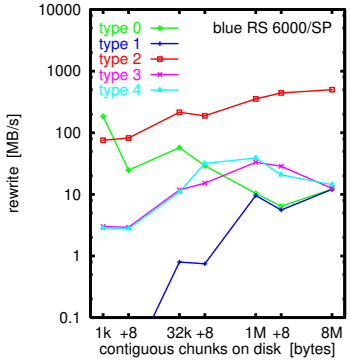


Full bandwidth on **Cray T3E**:
- about 30% of peak performance
- reached already with 8 processors!
  => optimal for any load:
  many small jobs ... one large job

Full Bandwidth **IBM SP**:
- about 35% of peak performance
- reachable only with high-CPU-count jobs

First Results: "rewrite" on Cray T3E and IBM SP

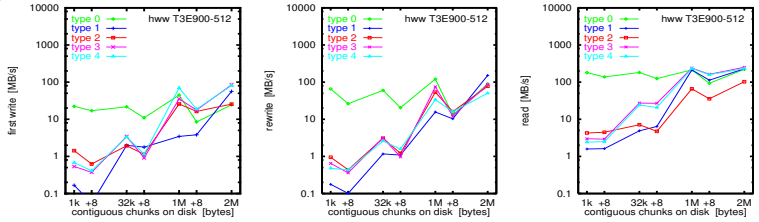




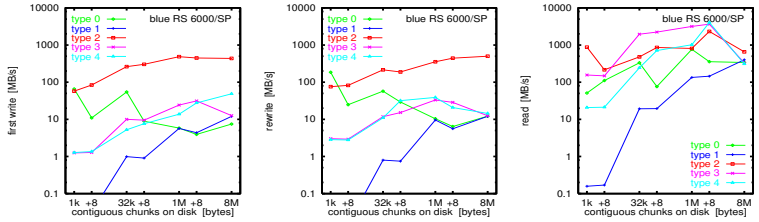
Cray T3E, 32 PEs, T=10 min.

IBM SP, 128 nodes, T=10 min.

First Results on Cray T3E and IBM SP

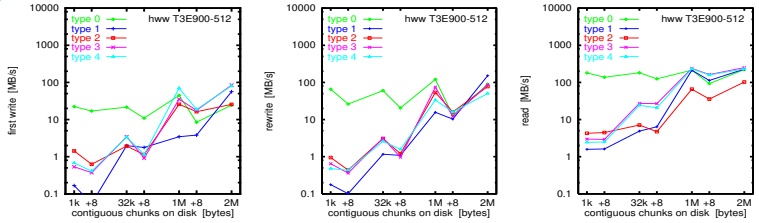


Cray T3E, HLRS, 32 PEs, T=10 min.: b_eff_io = 71 MB/s

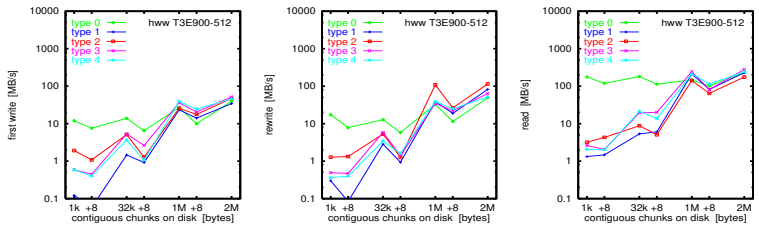


IBM SP, LLNL "blue", 128 nodes, T=10 min.: b_eff_io = 311 MB/s

First Results on Cray T3E and IBM SP



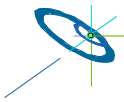
Cray T3E, HLRS, 32 PEs, **T=10 min**.: b_eff_io = 71 MB/s



Cray T3E, HLRS, 32 PEs, **T=30 min**.: b_eff_io = 62 MB/s

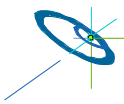
First Results — Interpretation

- maximum bandwidth / partition sizes
- small influence of T
- benchmarked platforms: MPI-I/O is optimal only for one pattern type
- but different optimal type on each platform
- non-wellformed data sizes: worse I/O bandwidth
- (re)write bandwidth << read bandwidth
- no chance to predict bandwidth for other patterns



Outlook

- www.top500cluster.org
- Issues
  - collecting hardware characteristics on clusters
  - several benchmark results
  - stored in a database
  - web-interface
  - each reader can define his own weights, and
  - can receive a personal weighted (ranked) list of all clusters
- Status
  - hardware information: some clusters already stored in database
  - benchmarks and web-interface: under discussion
  - b_eff and b_eff_io are under evaluation

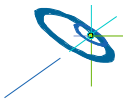


Summary

Effective I/O Bandwidth Benchmark (b_eff_io)

- characteristic average number for I/O bandwidth
- detailed information about several patterns:
  - access pattern types,
  - buffer sizes,
  - access methods (initial write, rewrite, read)
- 15 minutes for a first pass on any platform

Results

- Cray T3E900-512 at HLRS
- IBM RS 6000/SP at LLNL (“blue”)

Usable on MPP systems, SMP systems, and on clusters of SMPs