Effective Communication and File-I/O Bandwidth Benchmarks

b_eff and 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*

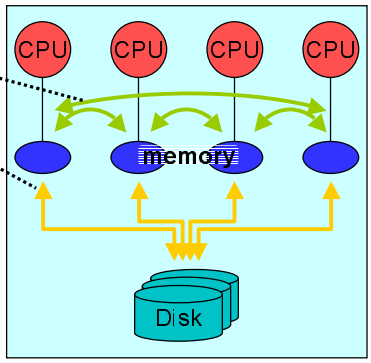
European PVM / MPI Users' Group Meeting 2001
Santorini, Greece, April 18-20, 2002 (postponed from Sep. 23-26, 2001)

H L R S

Effective Communication & I/O Bandwidth Benchmarks

Goals

- Parallel **Communication** Benchmark
- Parallel **File-I/O** Benchmark
  - each process is involved!
- Detailed insight
  - bandwidth experiments of several
    - I/O or communication patterns
    - chunk or message sizes
- One characteristic value
  - based on experiments above
  - averaging
- Appropriate execution time for rapid benchmarking



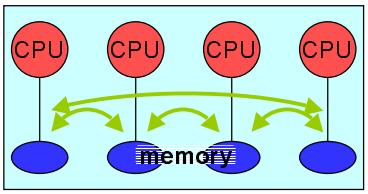
H L R S

b_eff

the
effective communication bandwidth benchmark

H L R S

Definition of the Effective Communication Bandwidth Benchmark: b_eff



- www.hlrs.de/mpi/b_eff/
- Authors: Karl Solchenbach, Hans-Joachim Plum, and Gero Ritzenhoefer (Pallas), Rolf Rabenseifner (HLRS)

- 6 ring patterns
- 30 random patterns
- 13 additional patterns
- 21 message sizes
- 3 communication methods
- 3 times repeated

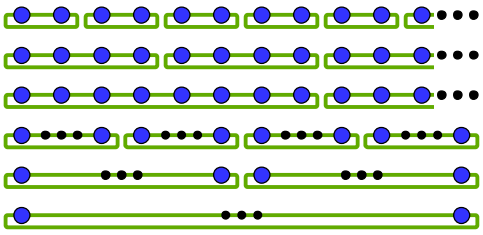
(6+30+13) x 21 x 3 x 3 = **9261 experiments**

- automatically controlled measurement-loop length, i.e., time driven approach
- 5 - 20 msec / experiment —> benchmark completes in a **few minutes**

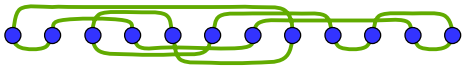
H L R S

Definition of b_eff — communication patterns and sizes

- **6 ring patterns**
  - ring size = 2
  - 4
  - 8
  - max(#PE/4, 16)
  - max(#PE/2, 32)
  - #PE



- **30 random patterns**



- **21 message sizes**
  - 1, 2, 4, 8, 16, 32, 64, 128, 256, 512 byte, 1kB, 2kB, (12 sizes)
  - 9 logarithmic equidistant sizes: 4kB, ..., L_{max}= memory per PE / 128

Definition of b_eff — averaging

One characteristic accumulated communication bandwidth number
:= average bandwidth on several communication patterns
average on different message sizes
maximum over different MPI programming methods

$$b_eff = \mathrm{logavg}\Big(\mathrm{logavg}_{\mathbf{ring}pat}\ \big(\mathrm{avg}_{\mathbf{L}}\big(\max_{\mathbf{method}}\big(\max_{\mathbf{rep}}(b_{\mathbf{pat,L,method,rep}}) \big)\big)\big),\ \mathrm{logavg}_{\mathbf{random}pat}\ \big(\mathrm{avg}_{\mathbf{L}}\big(\max_{\mathbf{method}}\big(\max_{\mathbf{rep}}(b_{\mathbf{pat,L,method,rep}}) \big)\big)\big) \Big)$$

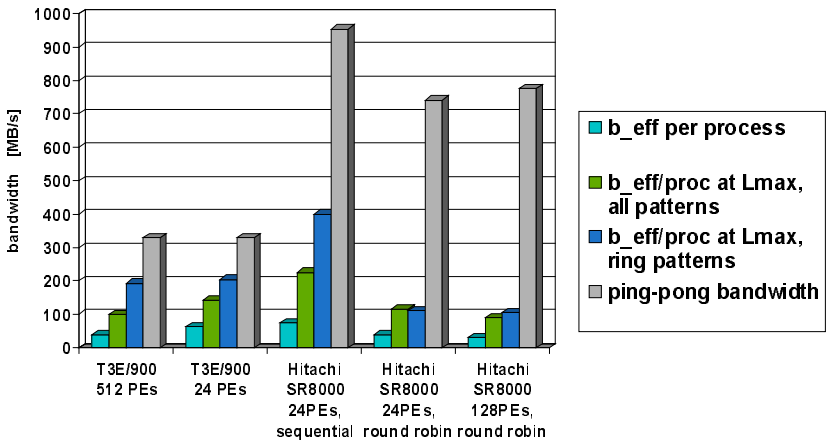
with

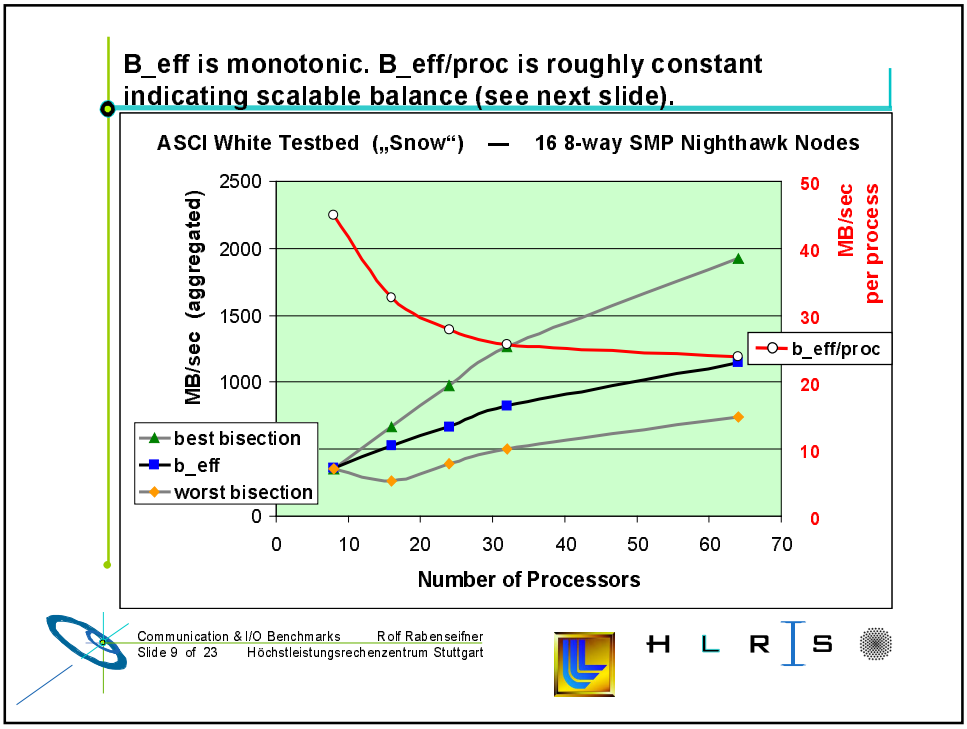
- $b_{\mathbf{pat,L,method,rep}}$ = accumulated bandwidth of each experiment
  over all processes
- methods: MPI_Sendrecv, MPI_Alltoallv, and nonblocking Irecv&Isend&Waitall
- pat & L: patterns and message sizes, see previous slide
- rep: repetition number = 1..3
- avg: arithmetic mean
- logavg: geometric mean

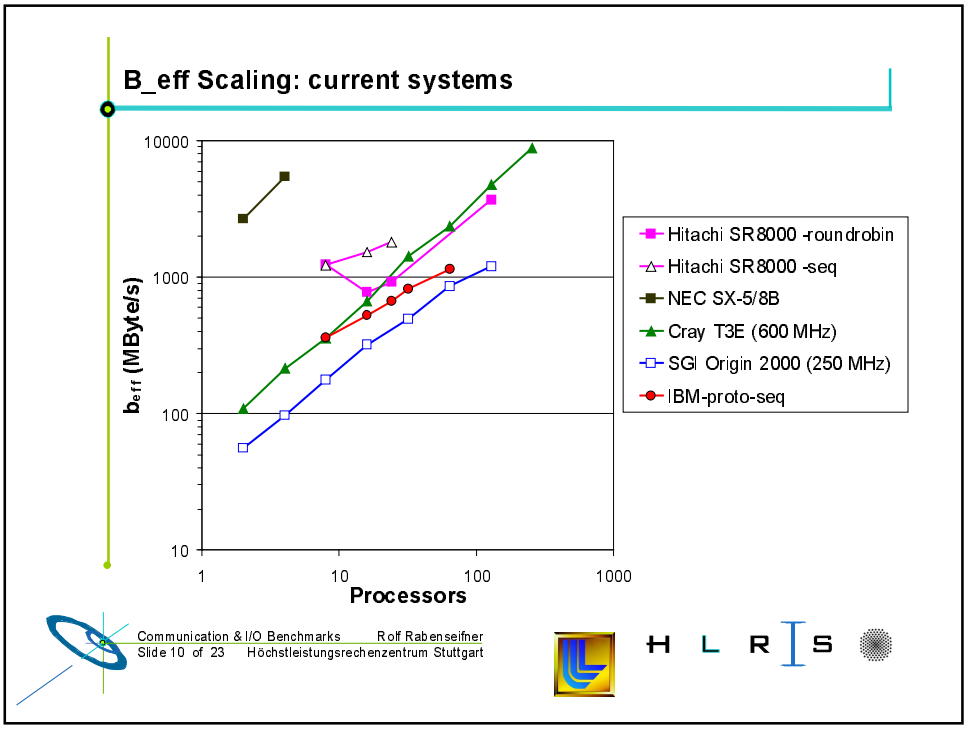
Features of Effective Bandwidth benchmark

- Based on MPI, source code is available
- Measures total architecture, not only point-to-point
- Checks performance of architecture and not the quality of the MPI implementation
- Suited for MPP-architectures and clusters
- Runs on any number of processors
- Results are easy to understand
- Generates a single number b_{eff} (like LINPACK R_{max})
- Aggregate bandwidth of the total system

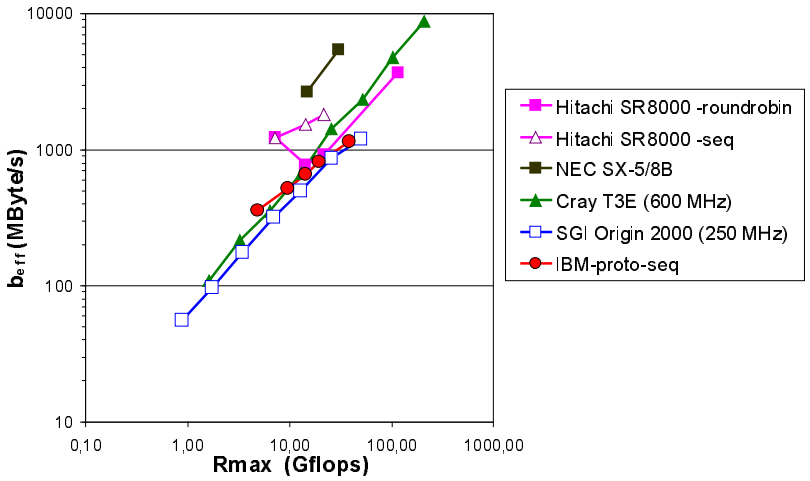
b_eff per process and ping-pong bandwidth




B_eff is monotonic. B_eff/proc is roughly constant indicating scalable balance (see next slide).
ASCI White Testbed („Snow“) — 16 8-way SMP Nighthawk Nodes
MB/sec (aggregated)
MB/sec per process
b_eff/proc
best bisection
b_eff
worst bisection
Number of Processors
Communication & I/O Benchmarks Rolf Rabenseifner
Slide 9 of 23 Höchstleistungsrechenzentrum Stuttgart
H L R S


B_eff Scaling: current systems
b_eff (MByte/s)
Processors
Hitachi SR8000 -roundrobin
Hitachi SR8000 -seq
NEC SX-5/8B
Cray T3E (600 MHz)
SGI Origin 2000 (250 MHz)
IBM-proto-seq
Communication & I/O Benchmarks Rolf Rabenseifner
Slide 10 of 23 Höchstleistungsrechenzentrum Stuttgart
H L R S

B_eff Scaling: current systems



H L R S pallas

b_eff_io

the effective MPI-I/O bandwidth benchmark

H L R S

Goals for b_eff_io

- portable
- characterizing parallel I/O capabilities of a system
- many patterns
- ... that can be optimized
- going beyond the caching capabilities
  --> measuring the real disk I/O

Rule: Balanced HPC systems should be able to write the total memory in 10 minutes to disk

==> **An I/O benchmark should not need hours!**
— 10 minutes may be enough to overrun any cache!

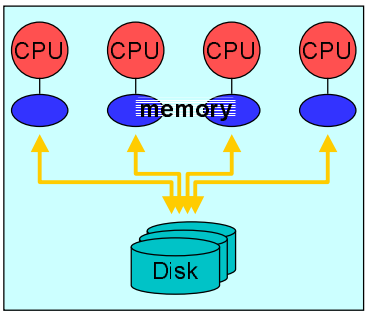
- time driven approach / automatically controlled repetition factors
- rapid benchmarking (30-60 min.)

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) on disk, (c) ... (f)
  - **Usage aspects:**
    - (a) how many processes are used
    - (b) how many parallel processors and threads are used for each process.
  - **I/O interface:**
    - (a) Posix I/O buffered or (b) raw,
    - (c) special filesystem I/O of the vendor filesystem,
    - (d) MPI-I/O.
  - **MPI-I/O aspects:**
    - (a) access methods, i.e., first writing of a file, rewriting or reading, (b) ...
    - (c) coordination, i.e., collectively or noncollectively, (d) ... (f)
  - **Filesystem parameters:**
    - (a) which filesystem is used,
    - (b) how many nodes are used as I/O servers, (c) ... (f)

  (full list, see paper)

Definition of the Effective File-I/O Bandwidth Benchmark: b_eff_io

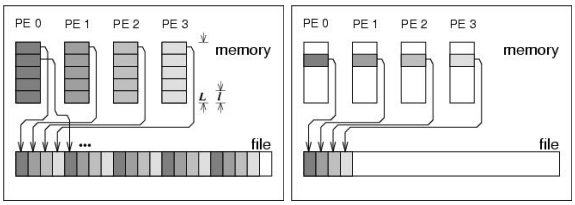


- 5 I/O patterns
- 7 chunk sizes
- 3 accesses (initial write, rewrite, read)
- 3 compute partition sizes (number of parallel benchmark processes)

315 different measurements

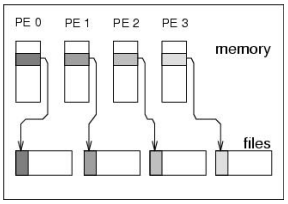
- benchmark completes in **30-60 minutes**
- www.hlrs.de/mpi/b_eff_io/

Definition of b_eff_io — the Pattern Types

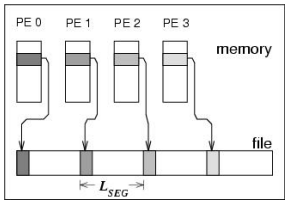


type 0: scattering *)

type 1: strided coll.



type 2: noncoll., separated



type 3 noncoll. / 4 coll. segmented

Pattern that can be optimized

Chunk sizes on disk:

- L_{max} = 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

(Release 1.0)

b_eff_io := Maximum over all usage and filesystem parameters } manually
Average on write, rewrite, read
Average on five access pattern types
Average on several chunk size values*)
of measured bandwidth
} auto-matically, in time T=30 min.

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

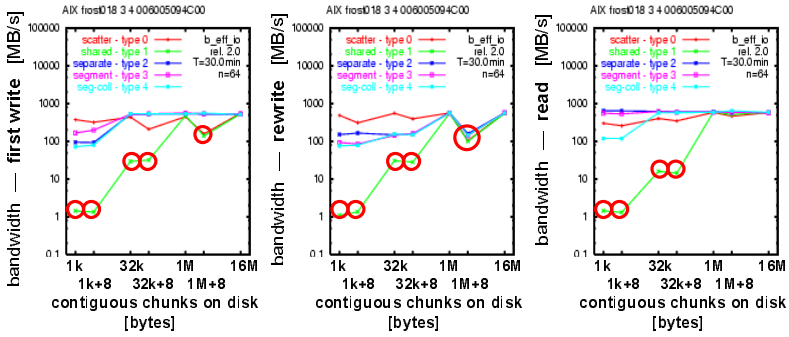
Output of the b_eff_io benchmark program

- the b_eff_io value

weighted average bandwidth for **write** : **422.**388 **MB/s** on 64 processes
weighted average bandwidth for **rewrite** : **331.**833 **MB/s** on 64 processes
weighted average bandwidth for **read** : **473.**441 **MB/s** on 64 processes
Total amount of data written/read with each access method: **208558.**575 MBytes
= 39.8 percent of the total memory (**524288** MBytes)
b_eff_io of these measurements = 451.727 **MB/s**
on 64 processes with 2048 MByte/PE, scheduled time=30.0 Min,
on AIX frost018 3 4 006005094C00
total memory / b_eff_io = 524288 Mbytes / 451.727 MB/s **= 19.3 min.**

- detailed results
  - as ASCII table
  - one page with 3+5 plots
    - **all measurements sorted by access:** write / rewrites / reads
    - **and same sorted by pattern types:** type-0 / type-1 / type-2 / type-3 / type-4

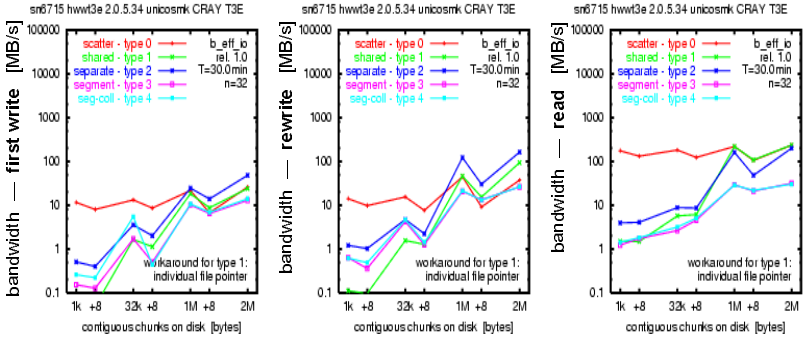
Results on ASCI White testbed with optimal hints



- each pattern may be optimized with hints *)
  - here: IBM_large_block_io=false & IBM_io_buffer_size=2MB (instead of 8MB) for "*scatter*" and "*shared*" pattern type with chunk sizes <= 32k+8
  - 86 from 105 patterns quite fast ! (19) candidates for further optimization

*) R. Rabenseifner, A. E. Koniges, J.-P. Prost, and R. Hedges: The Parallel Effective I/O Bandwidth Benchmark: b_eff_io. To be published in Special Issue of Calculateurs Parallèles Journal on Parallel I/O for Cluster Computing.

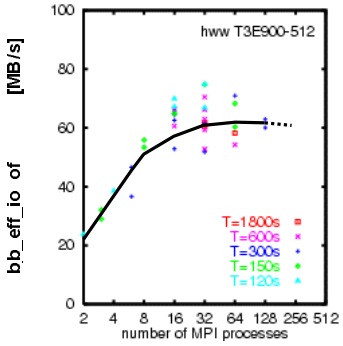
Detailed Results – Cray T3E-900/512 at HLRS *)

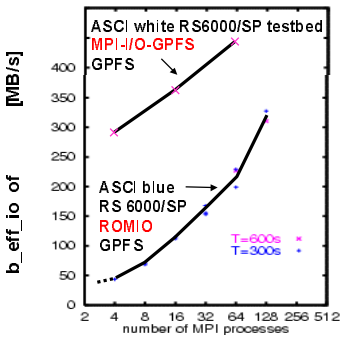


weighted average bandwidth for write: 12 MB/s on 32 processes (25%)
weighted average bandwidth for rewrite: 21 MB/s on 32 processes (25%)
weighted average bandwidth for read: 98 MB/s on 32 processes (50%)
b_eff_io of these measurements = 57 MB/s on 32 processes

*) with disk striping, patches at www.hlrs.de/mpi/mpi_t3e.html

First Results — Comparing b_eff_io (#processes)





Cray T3E:
- reached already with 8 processors!
  => optimal for any load:
  many small jobs ... one large job

IBM SP:
- with optimized MPI-I/O-GPFS
  - high bandwidth for any load
- with ROMIO
  - I/O capability bound to application
  - full bandwidth only with many processes

Acknowledgements

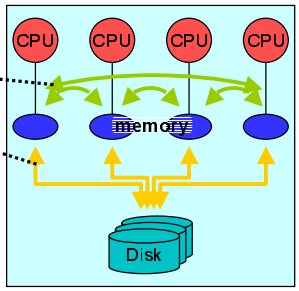
Thanks to pallas.
They initiated this project with their bi-section based b_eff benchmark.

Work by Lawrence Livermore National Laboratory
is performed under the auspices of the U.S. Department
of Energy by the University of California
under Contract W-7405-ENG-48, UCRL-VG-143637.

Summary

Two parallel benchmarks:

- b_eff: Communication
- b_eff_io: File-I/O
- Detailed insight
  - many patterns, chunk sizes, ...
- One characteristic value
  - averaging
- Appropriate execution time for rapid benchmarking
  - time driven approach
  - b_eff: 3–5 min.
  - b_eff_io: 30–60 min.
- ASCI white compared with ...



Further information:

www.hlrs.de/mpi/b_eff & .../b_eff_io
www.hlrs.de/mpi/mpi_t3e.html & .../ufs_t3e
www.hlrs.de/people/rabenseifner/publ/publications.html