25 Years of MPI Symposium

A View on MPI's Recent Past, Present, and Future

Argonne National Lab/EuroMPI/USA Conference, Chicago, IL

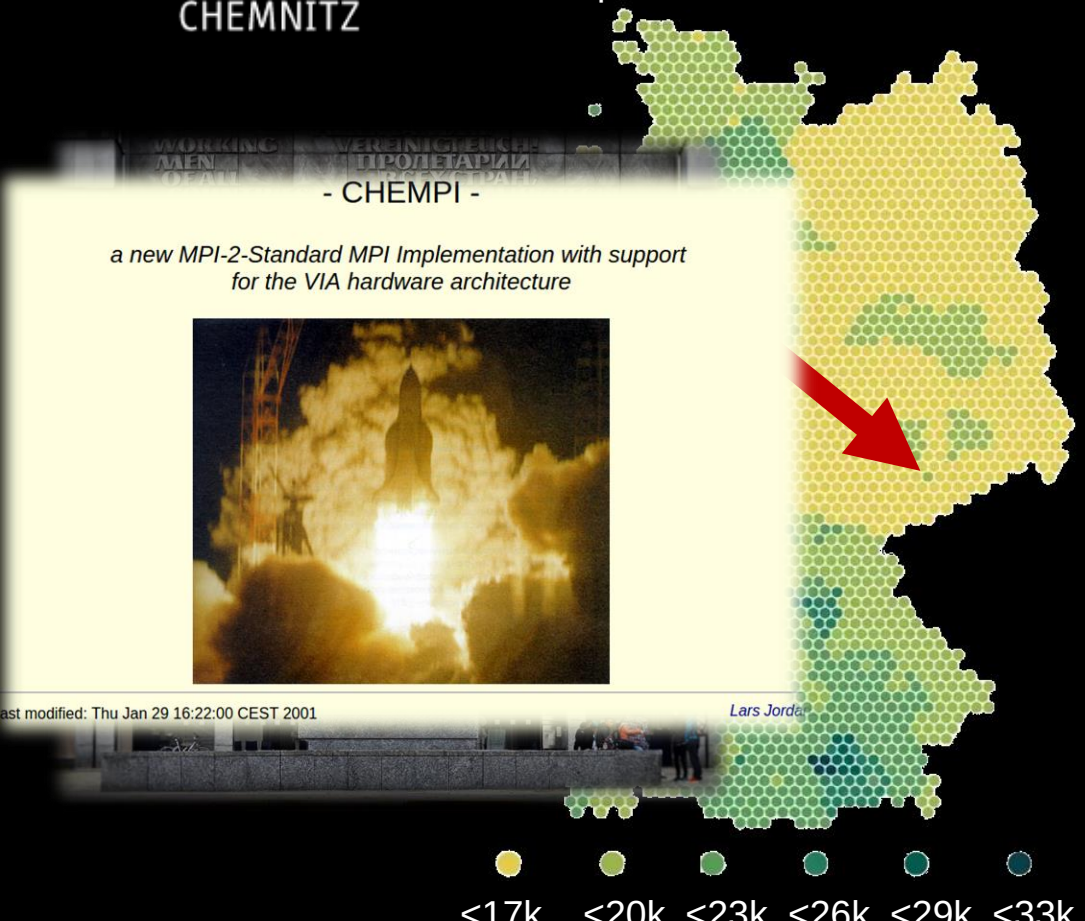
Torsten Hoefler (on behalf of nobody, neither my institution, nor myself, not MPI collectives WG!)

Thanks for organizing this!

My personal journey with MPI

TECHNISCHE UNIVERSITÄT CHEMNITZ

Disposable income distribution



CHEMNITZ UNIVERSITY OF TECHNOLOGY

Diploma Thesis

Evaluation of publicly available Barrier–Algorithms and Improvement of the Barrier–Operation for large–scale Cluster–Systems with special At[illegible]tion on InfiniBand™ Networks

Torsten Höfler
htor@informatik.tu-chemnitz.de

Advisers: Dipl.-Inf. T. Mehlan, Dipl.-Inf. F. Mietke

Supervisor: Prof. Dr.-Ing. W. Rehm

OPEN MPI

Nonblocking collective operations – first discussed in MPI-1!

- **MPI_I<collective>(args, MPI_Request *req);**

Implementation and performance analysis of non-blocking collective ...
ieeexplore.ieee.org/document/5348811
by T Hoefler - 2007 - **Cited by 162** - Related articles
Implementation and performance analysis of non-blocking collective operations for MPI. Abstract: ... LibNBC provides **non-blocking** versions of all **MPI collective operations**, is layered on top of **MPI**-1, and is portable to nearly all parallel architectures.

EuroPVM/MPI'06 Speedup for Jacobi/CG

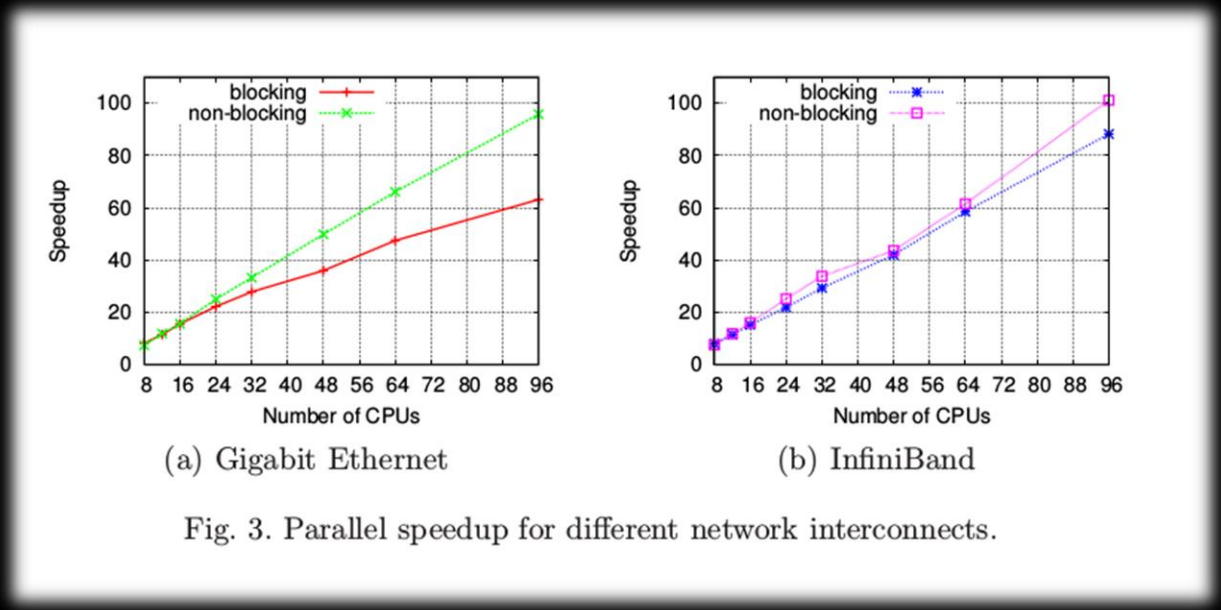


Fig. 3. Parallel speedup for different network interconnects.

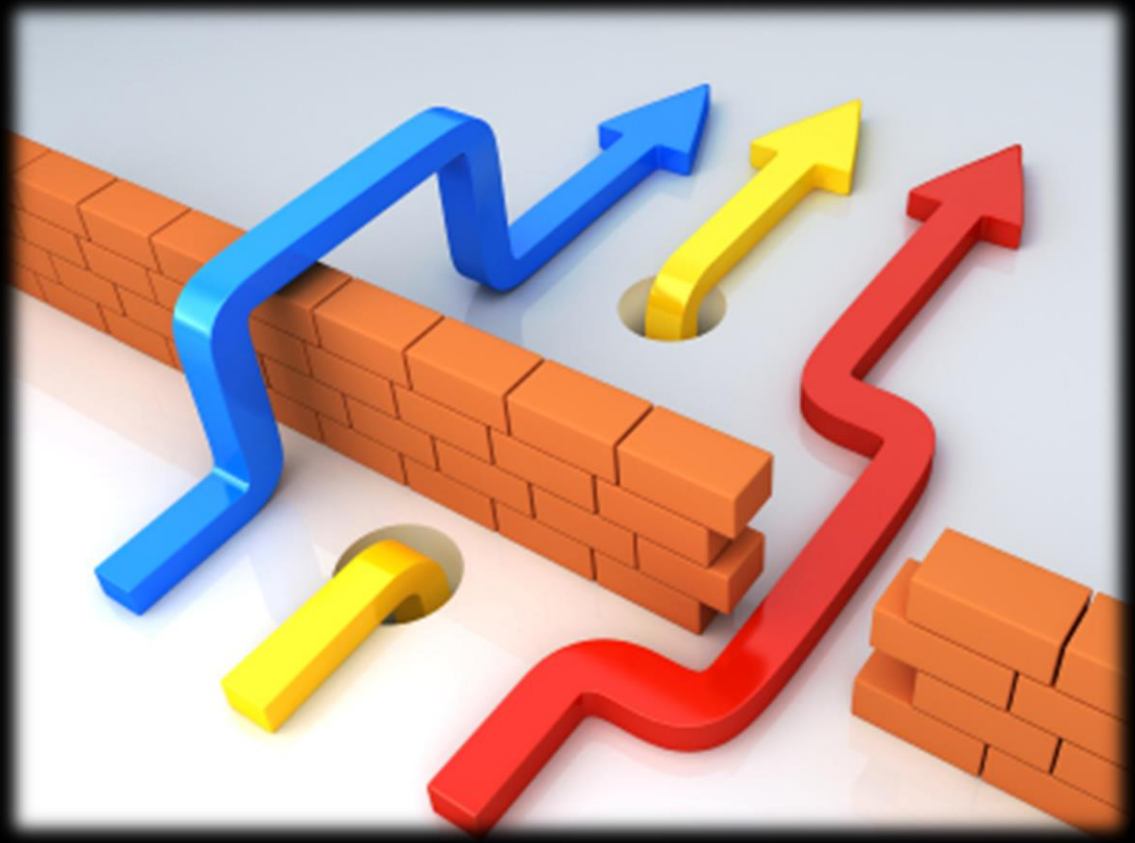
six years later

ten years later

Message progression in parallel computing - to thread or not to thread ...
ieeexplore.ieee.org/document/4663774/
by T Hoefler - 2008 - **Cited by 66** - Related articles

But wait, nonblocking barriers, seriously?

… turns out to be very useful after all:

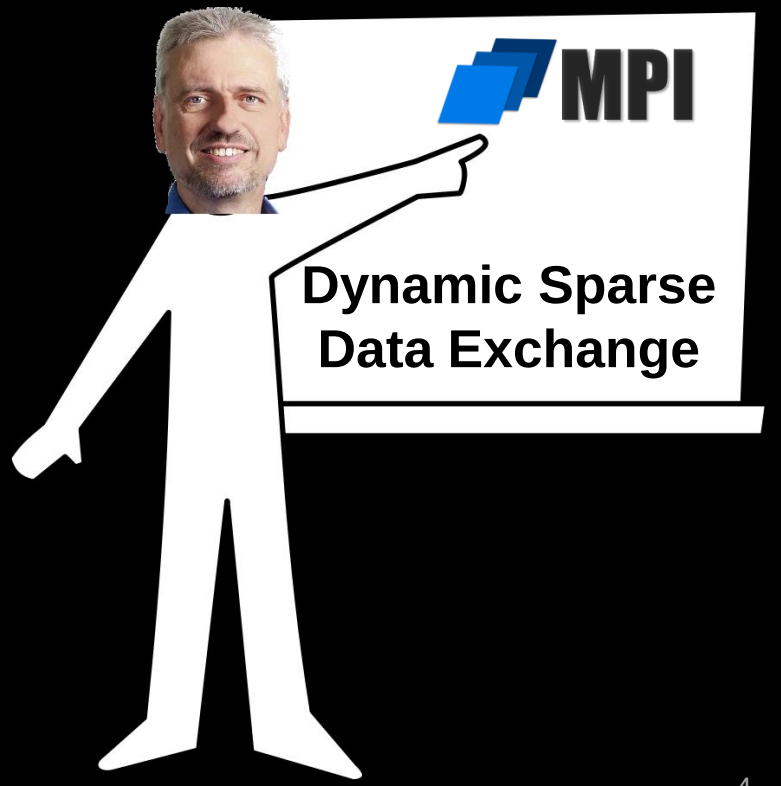


Scalable communication protocols for dynamic sparse data exchange
dl.acm.org/citation.cfm?id=1693476
by T Hoefler - 2010 - Cited by 41 - Related articles
Jan 9, 2010 - We define the dynamic sparse data-exchange (DSDE) problem and derive bounds in the well known LogGP model. While current approaches ...

MPI_Ibarrier()

+

MPI_Issend()



Neighborhood Collectives

- Just datatypes for collectives – default collectives are "contiguous", neighbor collectives are user-defined

1994 → 2004

ClusterWorld™
REDEFINING HIGH PERFORMANCE COMPUTING

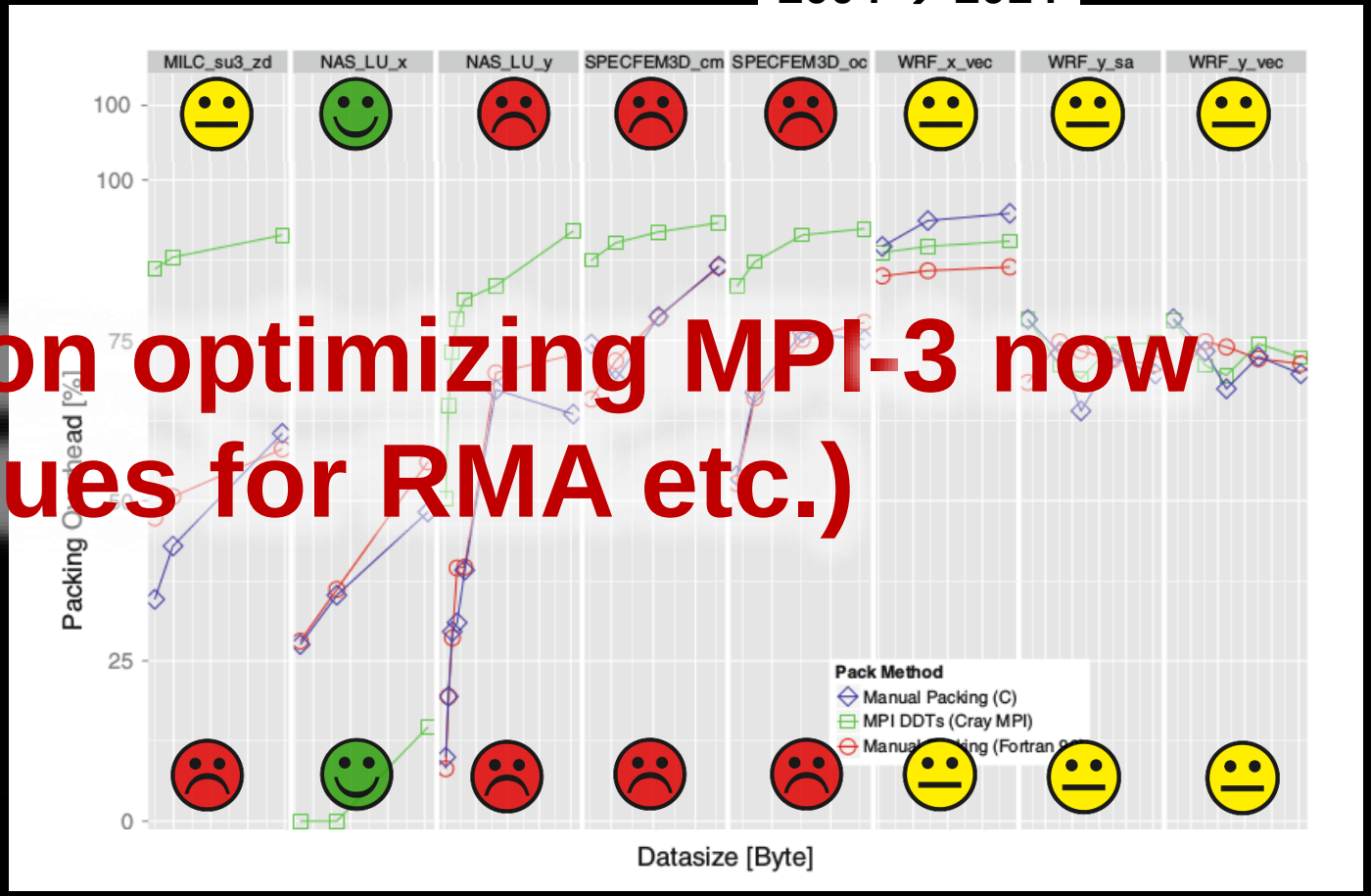
MPI Mechanic

June 2004

One common misconception with MPI datatypes is that they are slow. Early in the life of MPI, using MPI datatypes to pack messages was often slower than packing the data by hand. Datatype performance has been and continues to be an active area of research, allowing datatype implementations to achieve much higher performance. Some MPI implementations are even capable of doing scatter/gather sends and receives, completely eliminating the need to pack messages for transfer. In short, poor datatype performance is generally a thing of the past, and it's getting better every day.

Application-oriented ping-pong benchmarking - ACM Digital Library
dl.acm.org/citation.cfm?id=2597580
by T Schneider - 2014 - Cited by 6 - Related articles

2004 → 2014



We need to focus on optimizing MPI-3 now (similar issues for RMA etc.)

State of MPI today – programming has changed dramatically

until 10 years ago

today's programming

OpenMP

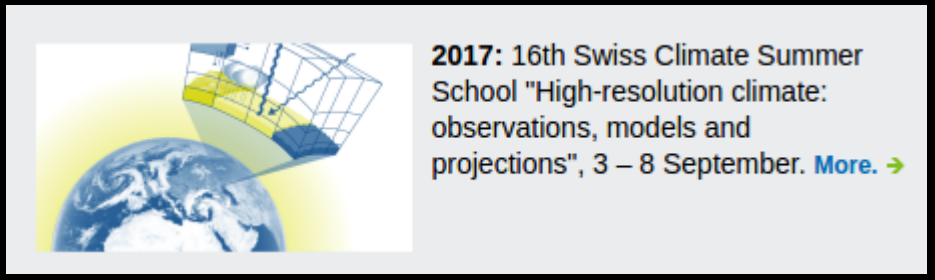
OpenACC
Directives for Accelerators

NVIDIA
CUDA

OpenCL

MPI

And the domain scientists?



HPC community codes towards the end of Moore's law (i.e., age of acceleration)

'07: Fortran + MPI

'12: Fortran + MPI + C++ (DSL) + CUDA

'13: Fortran + MPI + C++ (DSL) + CUDA + OpenACC

'??: Fortran + MPI + C++ (DSL) + CUDA + OpenACC + XXX

What is with the MPI community and how can we help?

MPI's own Innovator's Dilemma

Data-Centric Parallel Programming

DAPP

Turn MPI's principles into a language!

erc

up

Replace MPI?

- **We should have a bold <u>research</u> strategy to go forward!**

down

Rethink MPI!

MPI+X

"I want you to find a bold and innovative way to do everything exactly the same way it's been done for 25 years."

Distributed CUDA

Run MPI right on your GPU (SC'16)

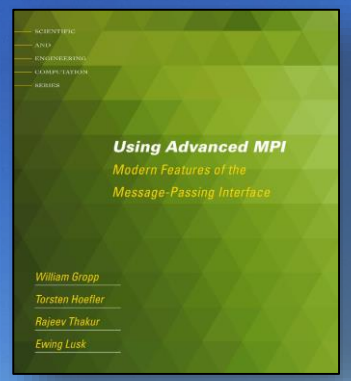
streaming Processing in Network

CUDA for Network Cards (SC'17)

MPI for Big Data

Distributed Join Algorithms on Thousands of Cores (VLDB'17)

Let's move MPI to new heights!



800m air

Torsten

https://spcl.inf.ethz.ch/Jobs/