Rusty & MPI

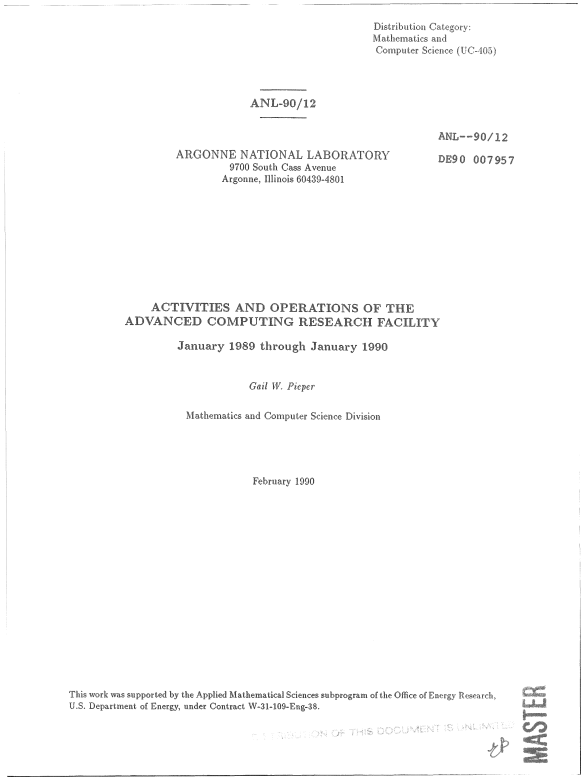
William Gropp & Marc Snir

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Computer Science

Bill

- I met Rusty when I joined ANL in 1990
- Advanced Research Computing Facility (ARCF)
  - I was deputy to Rusty
  - MCS division fielded a number of parallel systems. In 1990, when I joined ANL, they had:
    - BBN TC2000 (Butterfly II)
    - Multi-PSI (Japanese 5th-generation computer, in a workstation)
    - AMT DAP 510
    - Thinking Machines CM-2
    - Encore Multimax
    - Sequent Balance
    - Alliant FX/8
    - Ardent Titan
    - Intel iPSC/d5
    - Intel iPSC-VX/d4
  - All with different programming models, systems, and tools



Distribution Category:
Mathematics and
Computer Science (UC-405)

ANL-90/12

ANL--90/12

DE90 007957

ARGONNE NATIONAL LABORATORY
9700 South Cass Avenue
Argonne, Illinois 60439-4801

ACTIVITIES AND OPERATIONS OF THE
ADVANCED COMPUTING RESEARCH FACILITY

January 1989 through January 1990

Gail W. Pieper

Mathematics and Computer Science Division

February 1990

This work was supported by the Applied Mathematical Sciences subprogram of the Office of Energy Research, U.S. Department of Energy, under Contract W-31-109-Eng-38.

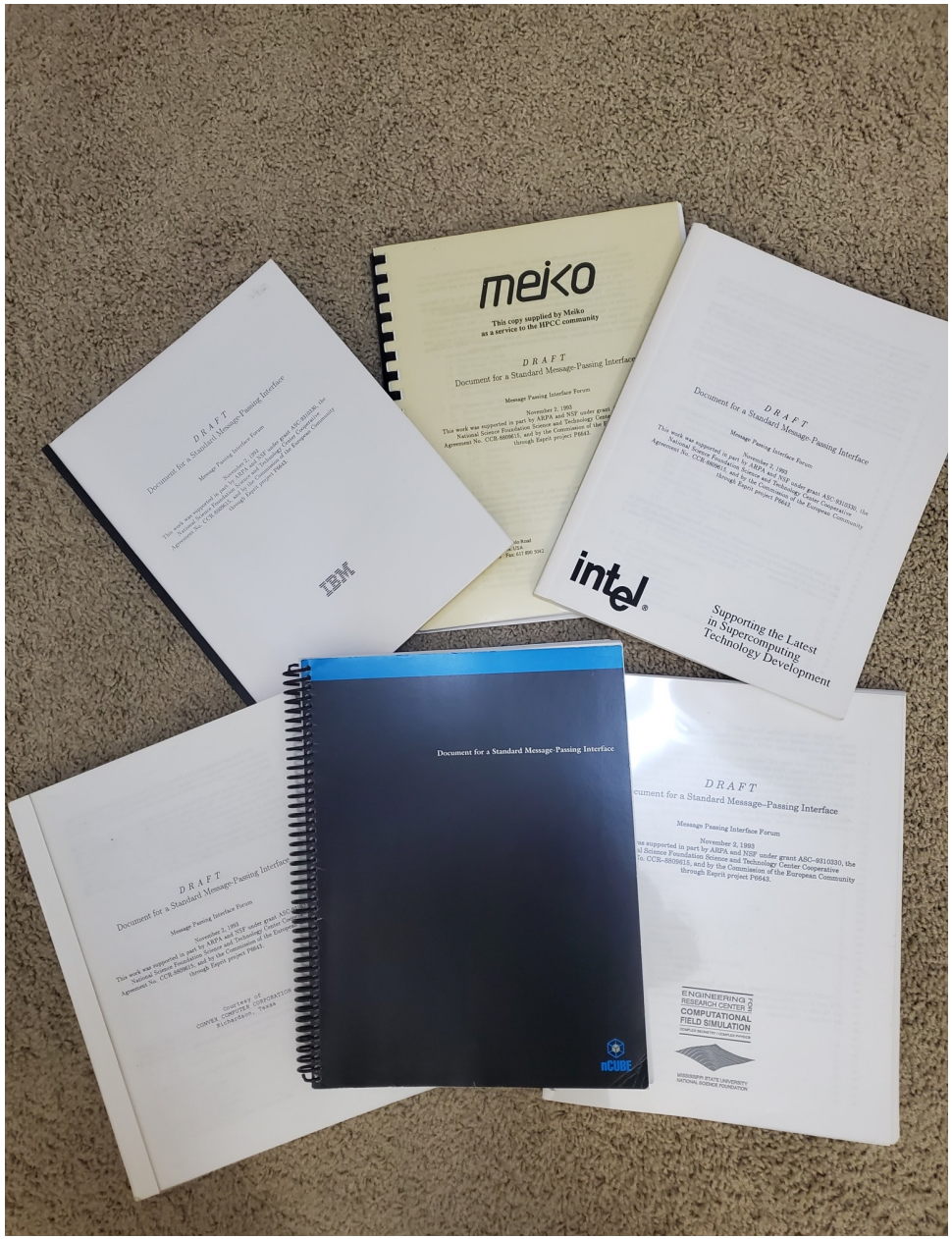
MASTER

Marc

- I actually met Rusty during his Prolog days
  - Some workshop at a forgotten place and time (Weizmann?) to discuss Parlog
  - Many people erred in their youth…

The Birth of MPI

Nov 2 1993 (SC'93, Portland, Oregon)



MPI1 presented at SC

- With the blessing of all major HPC vendors, the support of DOE labs, and many academic contributions
  - Convex, IBM, Intel, Meiko, nCUBE, Thinking Machines,..
- Developed within a year

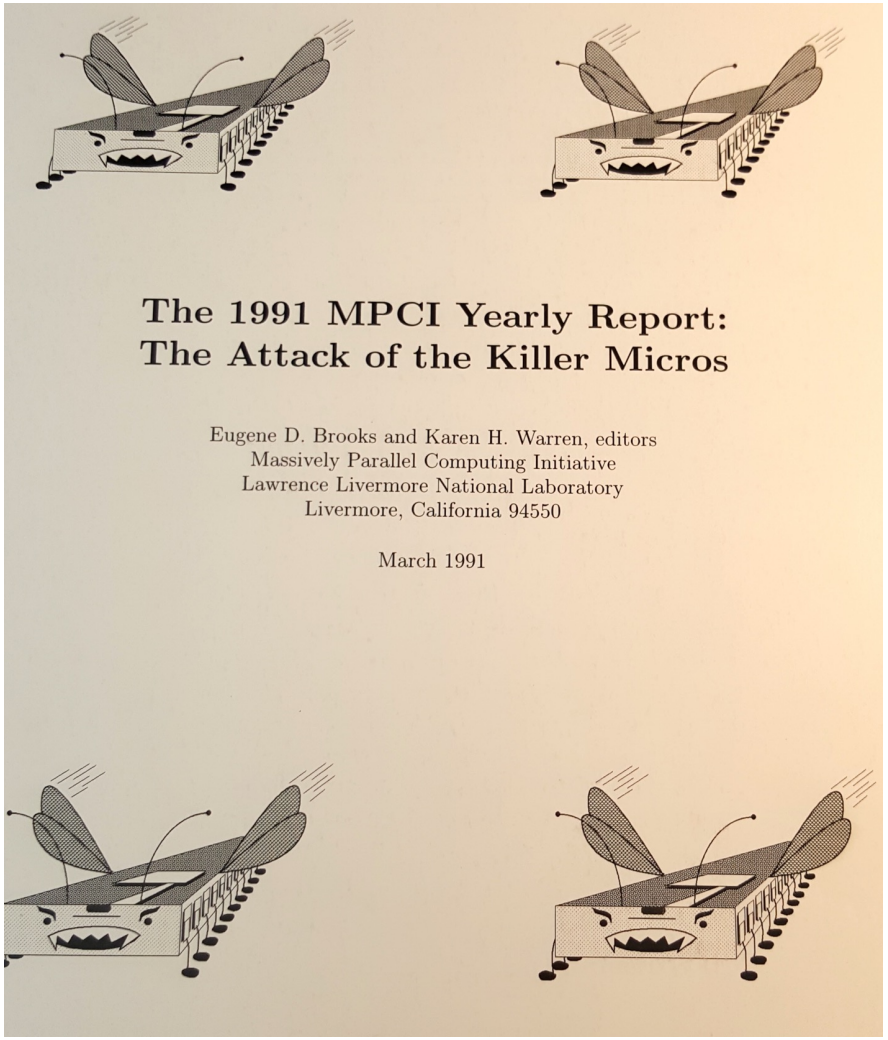
Background (1980's)

- Crazy idea: Build supercomputers by connecting large numbers of microprocessors
  - cheaper, but
- *If you were plowing a field, which would you rather use? Two strong oxen or 1024 chickens? (S. Cray)*
- Crazy idea, explored by crazy scientists (Cosmic Cube 1981->, Fox and Seitz)
- Picked the interest of Intel (iPSC 1984->), and a variety of startups (nCUBE, Kendall Square Research, Thinking Machines, Meiko,…)
- Heavily supported by ARPA and, later, DOE, as traditional vector supercomputing and fast ECL logic seemed to reach the end of its road.

Background (1990’)

- The shift to MPP’s (Massively Parallel Processing) becomes increasingly inevitable.
  - CM-5 (with 1024 processors) is on top of the first TOP500 list in June 1993
  - LLNL cancels its order for a Cray-3; no Cray-3 machine is ever sold and Cray Computer Corporation goes bankrupt in 1995.
  - DOE bets that the future of supercomputing is with MPP’s
  - Big companies enter the fray

The Attack of the Killer Micros



- 1990
  - E. Brooks talk at SC90
- 1991
- Thinking Machines CM-5 (Sparc)
- 1992
  - Intel Paragon (i860)
  - nCUBE-2 (proprietary)
- 1993 systems
  - Cray T3D (Dec Alpha)
  - IBM SP1 (IBM Power)
  - Meiko CS-2 (Sparc)

- Main weakness of these systems:
  - Much harder to program than vector machines.
  - No easy way to port vectorized codes

Possible Savior: High-Performance Fortran

- Single thread of control, data parallel execution (closer to vector model)
  - Preceded by CM-Fortran, Vienna Fortran, Fortran D
  - Standardized by a working group Nov 92 – Jan 93
- Viewed by many as the "right" long-term programming model for MPP's
  - But needing a long time to mature
    - Immature compiler technology
    - Easy to program, but hard to tune (not *transparent*)
- Meantime, need a standard, low-level, message-passing API
  - Temporary solution, until higher-level programming models mature

MPI Forum (Bill)

- Workshop organized by CRPC April 92
  - The "Williamsburg Workshop"
  - Ken Kennedy organized to discuss a common way to program the zoo of distributed memory parallel computers
  - Goal was to support HPF and compilers for other parallel languages
  - At end of the meeting
    - We were all depressed – the situation was clearly hopeless
    - Ken stood up and said something like "Clearly we can develop a common software interface" – and the sequence that led to MPI started
- MPI-1 preliminary proposal Nov 92 (Dongarra, Hempel, Hey, Walker)
- Group of interested people met at SC'92
  - Discussion of goals, constraints, …
  - Gropp presented some design principles, including no unnecessary data copies
  - Agreed to meet regularly, adopted the procedures (including the meeting hotel – the Bristol Suites in Dallas!) of the HPF Forum
- MPI Forum starts meeting every 6 weeks, using the formal procedures of the HPF forum
- Final result presented at SC93

Principles

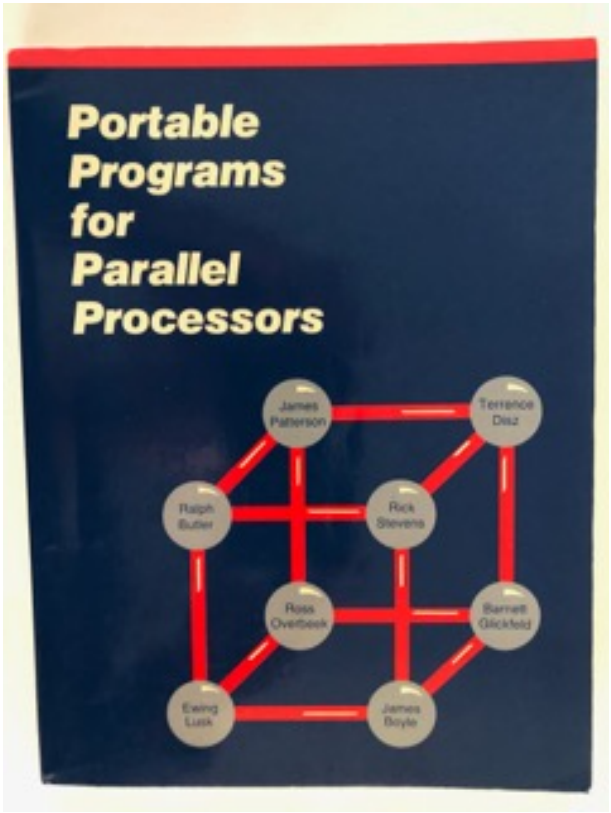
- Working together to find the golden mean
  - Rules on semantics before syntax
  - Object oriented design in C/Fortran
  - Balance between elegance, and practicality
  - Balance between short-term and long-term (e.g., scalability)
- You have to ***do*** something
  - Propose something specific – not just toss bombs
  - Differences of opinion resolved offline, e.g., at the bar
- We all had a hand in these, but Rusty was a constant supporter of processes that would get us something useful, and after MPI-1, he took over convening the MPI Forum for MPI-2

Challenges (Marc)

- Many participants – many opinions
- Some thought there is no need for a new standard; PVM, or Express, or… is what's needed
- Different existing message-passing libraries had different semantics
  - E.g., can a send complete before a receive? (CSP semantics)
- Vendors already had message-passing libraries running on their systems and wanted to preserve their investments
  - And make sure MPI matches the features of their hardware
    - E.g., how large are tags? (CM-5 issue)
- Participants had differing views on how "low-level" MPI should be (communicators vs. groups)

Why Rusty was essential to the process

- Argonne had the p4 system (Portable Programs for Parallel Processors)
- Rusty and Bill used it to prototype MPI design as it evolved
  - crucially showing that proposed constructs can be implemented efficiently
- But p4 was never pushed as "the solution"
- Rusty was an egoless promoter of the goal of defining an API that is acceptable by all and is right



Ewing L. Lusk, James M. Boyle, Ralph M. Butler, Terrence Disz, Barnett Glickfeld, Ross A. Overbeek, James H. Patterson, and Rick L. Stevens

Implementations (Bill)

- To succeed, the MPI Standard needed implementations
  - Not just any implementation! Required were
  - Performant, not just Functional
  - Transportable to a wide variety of systems
  - Able to exploit different hardware/system features
    - Remember, no standard networking, no standard CPU architecture, even byte ordering not standard
- MPICH built on Chameleon, a portability layer on top of other message passing systems
  - Chameleon designed as extremely lightweight layer
    - Used for 1992 winner of Gordon Bell Prize for Speedup
    - Direct to Intel NX, IBM EUI, CMMD, …
  - MPICH focused on performance, particularly avoiding unnecessary memory copies
- P4 was one of those message passing systems
  - Used to provide MPICH support for communication through shared memory and sockets
  - Provided portability to nearly everything

Implementations

- Rusty and I committed to "Really run everywhere"
  - There was another popular system that counted "runs on host" as the same as "runs on system".
  - Both P4 and MPICH had a strong commitment to running on (nearly) everything
  - Remember, the 1990's was an era of innovation in parallel computing systems
  - The SC Test: How long until a new machine vendor admits that their MPI is MPICH?
    - Rusty and I used to wander the SC show floor and question vendors of parallel systems to see how long it took them to admit their MPI was based on MPICH
- Actively worked with vendors on high-performance ports
  - Understand the capabilities of the hardware
  - Ensure MPI provides access to performance
  - Rusty and I spent a week in St Augustin, Germany, working with NEC on a port to the SX-4
    - Exploit vector architecture and instructions; very high memory bandwidth
    - W. Gropp and E. Lusk. A high-performance MPI implementation on a shared-memory vector supercomputer. Parallel Computing, 22(11):1513–1526, January 1997.
    - At a meeting at LANL, we showed the SX-4 performance – which literally was jaw-dropping for LANL researchers

Contributions (Marc)

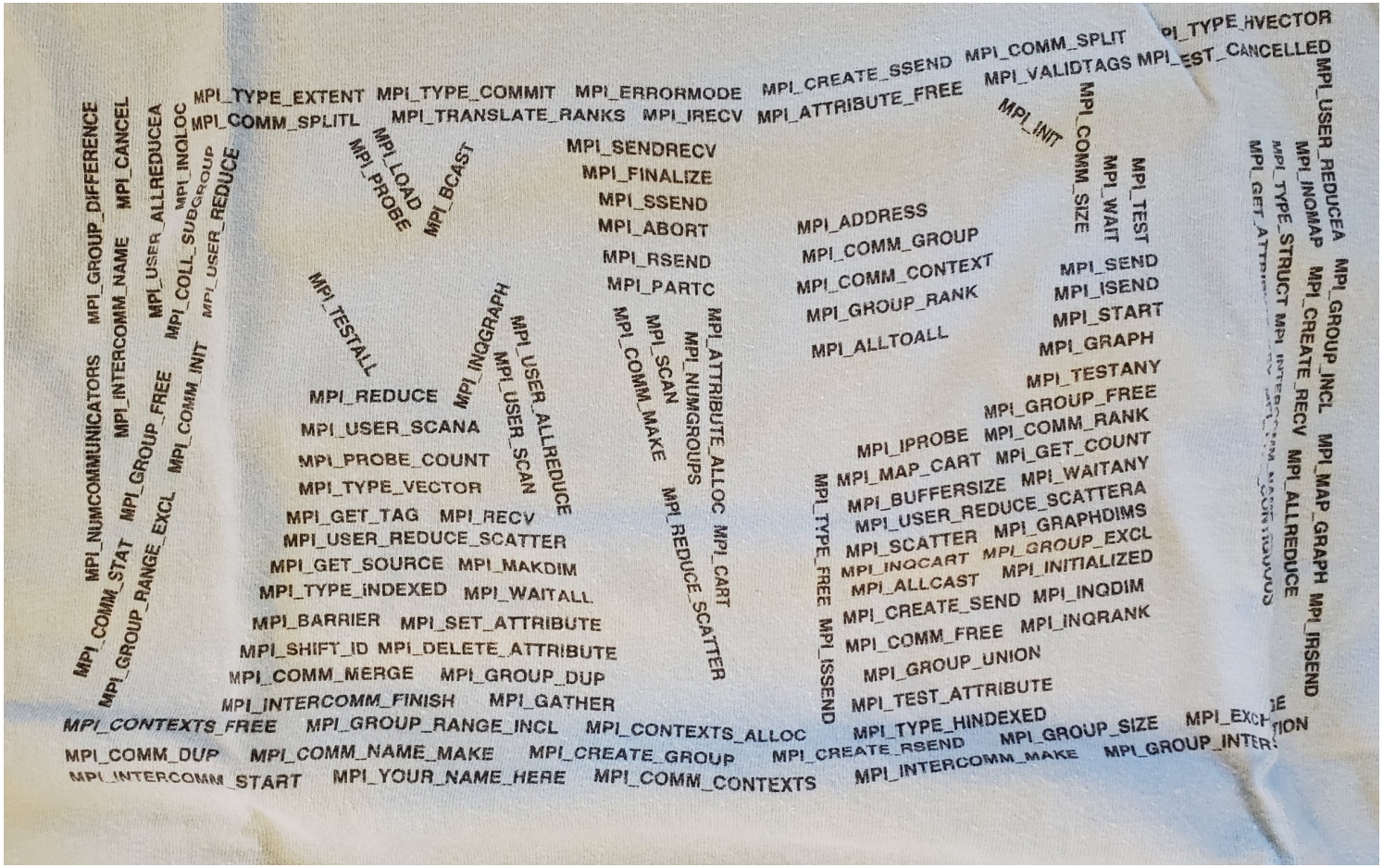
- MPI-1
  - Rusty is listed as secretary and editor of the point-to-point and language binding sections
- MPI-2
  - Rusty is listed as chair
- Main contribution
  - Getting a team of opinionated people with diverging interests to converge on one design
    - Usually, at the hotel bar
- Writing and editing the MPI documents with Rusty and Bill (at the old MCS building and the old Argonne guest house) was the most enjoyable collaboration in my career

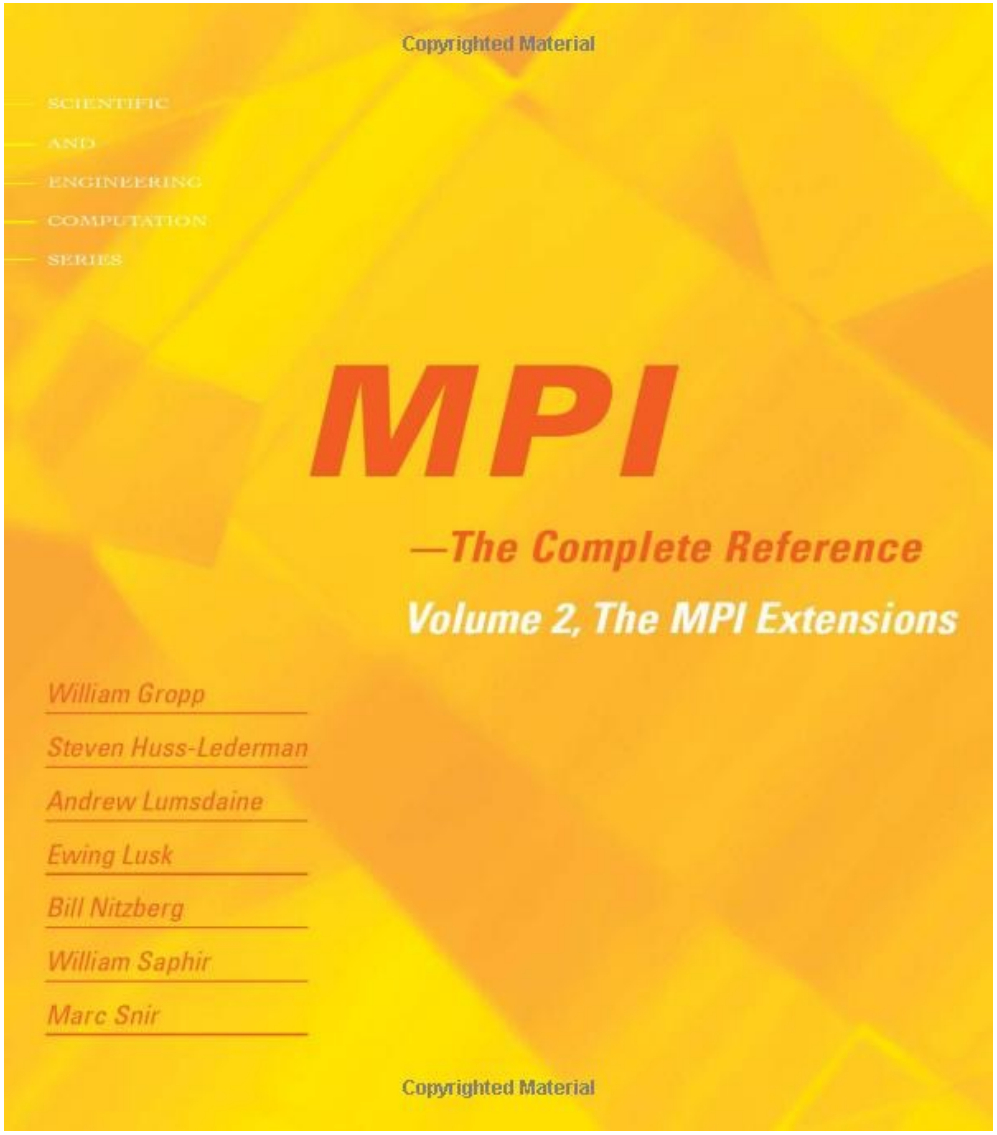
Outreach and Training (Bill)

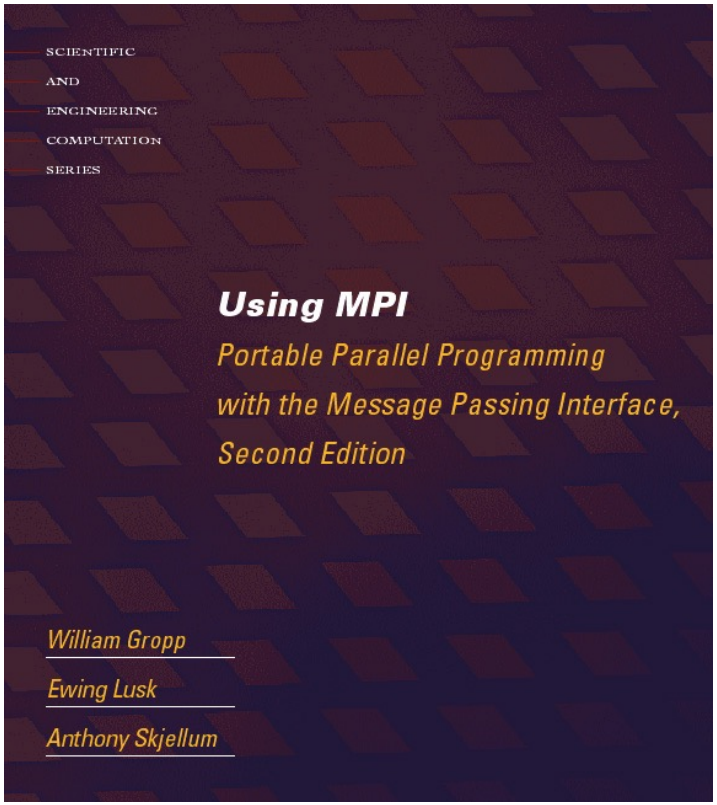
- Tutorials
  - Dialog with users. Helps identify points of misunderstanding
- Books
  - One way, but more scalable – more people able to learn about MPI by getting a book than through tutorials
- Tutorials
  - Connection with software that has to work
  - Value of feedback from tutorial attendees
    - Software bugs
    - Unclear descriptions, confusing concepts
  - Rusty's Mandelbrot program pmandel
- Books
  - "Complete Reference" (Marc, Steve Lederman as editor/LaTeX master)
    - Following the example of "The C++ Programming Language", Stroustroup
  - "Using MPI" (Bill, Rusty, and Tony)
    - There's always another typo
    - The "Bet" with MIT Press' Bob Prior

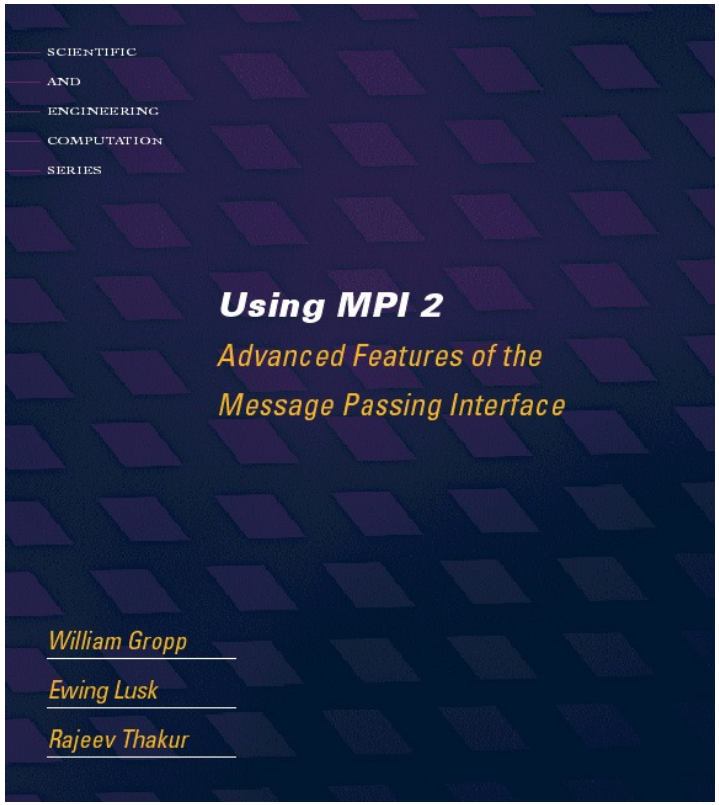
MPI-1 Documentation

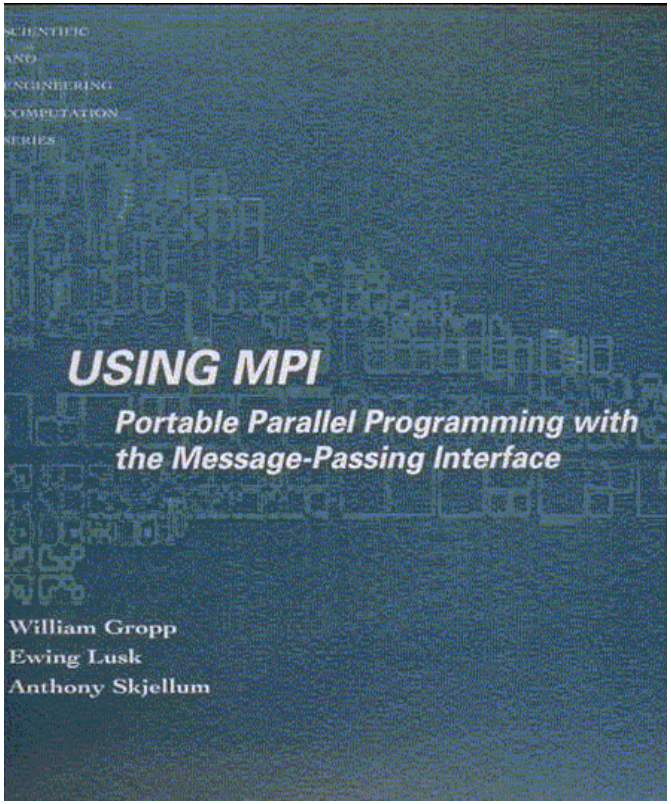
MPI-1: 128 functions

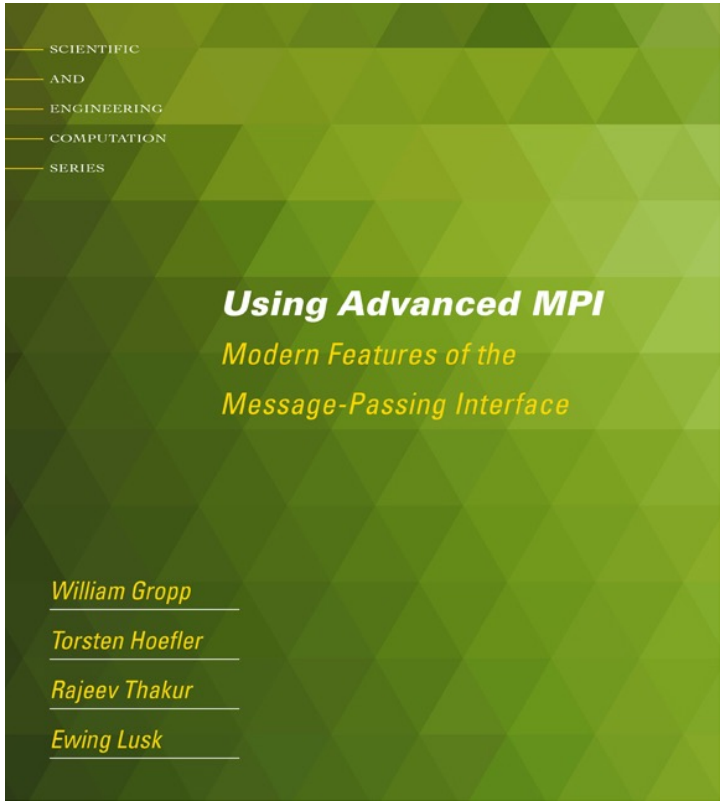



Copyrighted Material
SCIENTIFIC
AND
ENGINEERING
COMPUTATION
SERIES
MPI
—The Complete Reference
Volume 2, The MPI Extensions
William Gropp
Steven Huss-Lederman
Andrew Lumsdaine
Ewing Lusk
Bill Nitzberg
William Saphir
Marc Snir
Copyrighted Material


SCIENTIFIC
AND
ENGINEERING
COMPUTATION
SERIES
Using MPI
Portable Parallel Programming
with the Message Passing Interface,
Second Edition
William Gropp
Ewing Lusk
Anthony Skjellum


SCIENTIFIC
AND
ENGINEERING
COMPUTATION
SERIES
Using MPI 2
Advanced Features of the
Message Passing Interface
William Gropp
Ewing Lusk
Rajeev Thakur


USING MPI
Portable Parallel Programming with the Message-Passing Interface
William Gropp
Ewing Lusk
Anthony Skjellum


SCIENTIFIC
AND
ENGINEERING
COMPUTATION
SERIES
Using Advanced MPI
Modern Features of the Message-Passing Interface
William Gropp
Torsten Hoefler
Rajeev Thakur
Ewing Lusk

Lessons

- The success of MPI depended on
  - The usual – good design, attention to the needs of the community
  - The availability of good, correct, performant, transportable implementations
  - Commitment to outreach and training

Post MPI History (Marc)

- I replaced Rusty as Director of the MCS Division at Argonne in 2011
  - He was quite anxious to be replaced, but very honest about the challenges of the position – I had ~~no~~ few bad surprises after taking the job
  - Like Cincinnatus, he returned to research without looking back at the position he relinquished
    - But continued to host the department Christmas party