Grids and Clusters: Lessons for Deployment and Operation

Bill Gropp
Mathematics and Computer Science Division

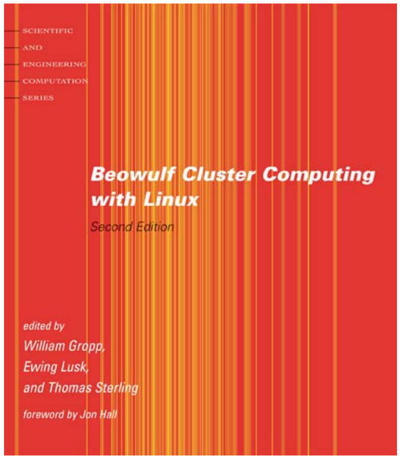
www.mcs.anl.gov/~gropp

ARGONNE NATIONAL LABORATORY • UNIVERSITY OF CHICAGO

Success of Clusters

- Engineered solutions
  - ♦ Goals characterized and resources matched to resources (online guides, common question on beowulf mailing list)
- Many studies making reproducible measurements of grid software
  - ♦ Quantitative measures of performance
  - ♦ Scientific results (reproducible simulations)
  - ♦ 53 page bibliography of papers involving MPI (mpiarticles.pdf)
    - Many are applications
- "Automatic" risk reduction
  - ♦ Multiple suppliers for software, hardware
    - OS, file systems, compilers, parallel computing approaches (MPI, Mosix, perl+ssh,…)
    - Processors, network, I/O
- Software runs everywhere
  - ♦ MPI, apps libraries run on laptops to Earth Simulator
  - ♦ Solutions *scale to users*
    - Install, operate done with configure/make or RPM or Windows Install
    - Does not require collaboration with developers
      - 8 Developers may not know users exist
    - User-oriented documentation

Practical Books on Cluster Computing



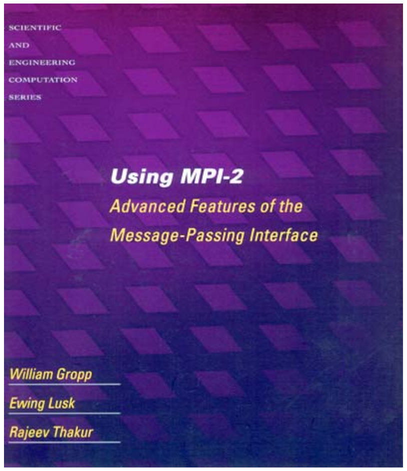
SCIENTIFIC AND ENGINEERING COMPUTATION SERIES

Beowulf Cluster Computing with Linux

Second Edition

edited by
William Gropp,
Ewing Lusk,
and Thomas Sterling

foreword by Jon Hall

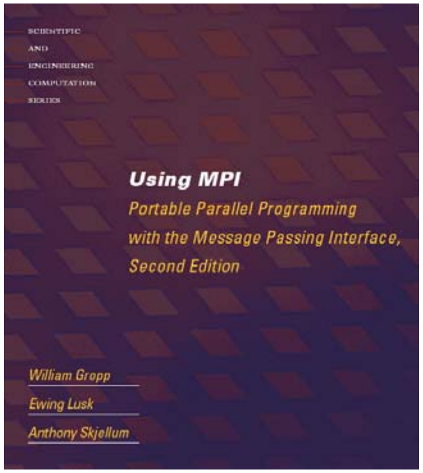


SCIENTIFIC AND ENGINEERING COMPUTATION SERIES

Using MPI-2

Advanced Features of the Message-Passing Interface

William Gropp
Ewing Lusk
Rajeev Thakur

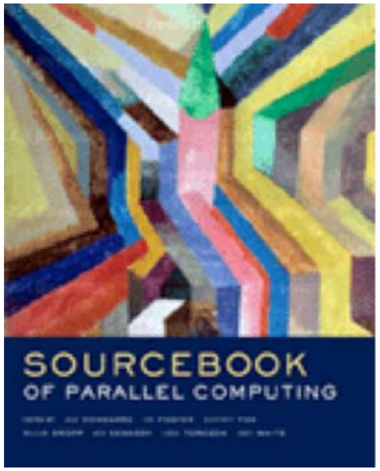


SCIENTIFIC AND ENGINEERING COMPUTATION SERIES

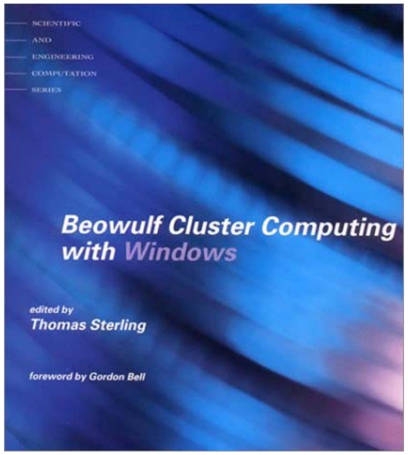
Using MPI

Portable Parallel Programming with the Message Passing Interface, Second Edition

William Gropp
Ewing Lusk
Anthony Skjellum



SOURCEBOOK
OF PARALLEL COMPUTING

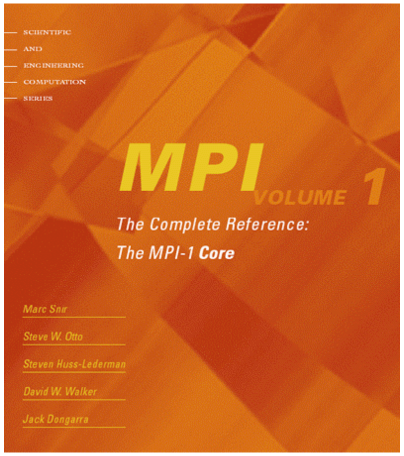


SCIENTIFIC AND ENGINEERING COMPUTATION SERIES

Beowulf Cluster Computing with Windows

edited by
Thomas Sterling

foreword by Gordon Bell

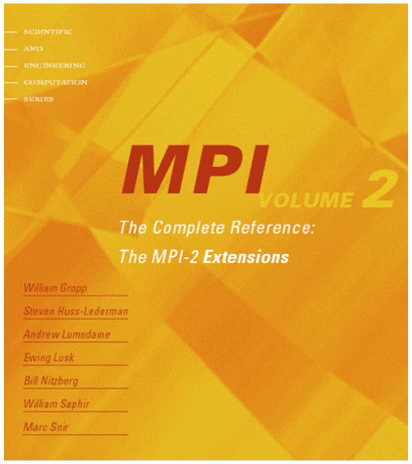


SCIENTIFIC AND ENGINEERING COMPUTATION SERIES

MPI VOLUME 1

The Complete Reference:
The MPI-1 **Core**

Marc Snir
Steve W. Otto
Steven Huss-Lederman
David W. Walker
Jack Dongarra



SCIENTIFIC AND ENGINEERING COMPUTATION SERIES

MPI VOLUME 2

The Complete Reference:
The MPI-2 **Extensions**

William Gropp
Steven Huss-Lederman
Andrew Lumsdaine
Ewing Lusk
Bill Nitzberg
William Saphir
Marc Snir

実践MPI-2

メッセージパッシング・インタフェースの
上級者向け機能

ウイリアム・グロップ/ユーイング・ラスク/ラジーブ・タークル＝著
畑崎　隆雄＝訳

Pearson Education Japan

ピアソン・エデュケーション

GridGate

- Fran’s talk gave good list of grid problems
  - ♦ “We still do not have a usable grid”
  - ♦ “There are too many partially complete, but shows proof of concept [,projects]”
  - ♦ “We blew the software by failing to plan and by not using widely accepted software engineering practices”
- Some grid successes
  - ♦ E.g., Andrew and Ed’s talks from this morning

Some Users Speak

- Jennifer Schopf and Steven Newhouse interviewed over 20 groups in the UK
- Some quotes:
  - “All users talk about is job submission and file transfer capabilities”
  - “When asked about trouble spots they also want tools to tell how jobs are progressing”
- What (These) Users Aren’t Talking About…
  - Notification – except for job progress tracking
  - Registries or resource discovery
  - Reservations, brokering, co-scheduling, other advanced scheduling techniques
  - Job migration, checkpointing
  - Accounting and pricing (these are users, not admins)
  - Data migration
  - Instruments
- Clearly, expectations are low
  - Gap between hype and reality
  - Users may not ask for what they really want (e.g., access to file contents instead of a copy of the file)

Optimism Only Goes So Far

Arthur: I command you, as King of the Britons, to stand aside!

Black Knight: I move for no man.

Arthur: So be it!

Arthur and Black Knight: Aaah!, hiyaah!, etc.

Arthur cuts off the Black Knight's left arm.

Arthur: Now stand aside, worthy adversary.
Black Knight: 'Tis but a scratch.
Arthur: A scratch? Your arm's off!
Black Knight: No, it isn't.
Arthur: Well, what's that then?
Black Knight: I've had worse.
Arthur: You liar!
Black Knight: Come on you pansy!

Arthur cuts off the Black Knight's right arm.

Moving Forward

- API design is *hard*
  - Focus on the task
    - MPI Forum took 18 months, meeting in the same ghastly hotel in the same ghastly spot every 6 weeks (just for MPI-1!)
  - Involve all stakeholders
    - Especially including the applications community
    - If you can't structure the process/discussion to deal with "hangers on", you haven't met the minimum intelligence test for creating a standard
  - Use prior art
    - Create prior art if necessary
    - Don't discourage experimentation. But avoid premature standardization
  - Start from scratch
    - Level the playing field and avoid historical errors
- Testing and evaluation
  - Establish these early
  - Helps clarify the goals and the community being served
  - "Live" test requires outside co-operation
    - Still, grid software could provide testing as a service (make testing target communicates with a machine set up solely for testing installations)

Choose the Right Solution

- Grid example—the fallacy of (single) certificates
  - ♦ Have you ever rented a car?
    - Probably not to get here :)
  - ♦ How many IDs did you show
    - Driver’s license — (admittedly weak) authentication
    - Credit card — resource authorization and guarantee
  - ♦ Each issued by a separate authority
    - Each issuer has different goals and mechanisms
    - No single-point of failure
- Multiple certificates alone are not the solution
  - ♦ No number of credit cards will compensate for the lack of a driver’s license

Challenges

- The Grid is a Fault-Rich Environment
  - ♦ Architecture must be designed for faults
  - ♦ No single points of failure
    - Idiot user
    - Single certificate
      - 8 Implies no single sign on.
- No matter what an idiot the user is, the user is always right
  - ♦ Think “canary in the mine”
  - ♦ Even the idiot users are pointing out problems:
    - Implementation — what we always claim
    - Design — where the real problem often lies
      - 8 Some problems are not fixable if the design is poor