Nizhni Novgorod State University

Optimizing Performance of MPI open-source implementations for Linux on POWER processor clusters

This work is partially supported by IBM Faculty Awards for Innovation Program

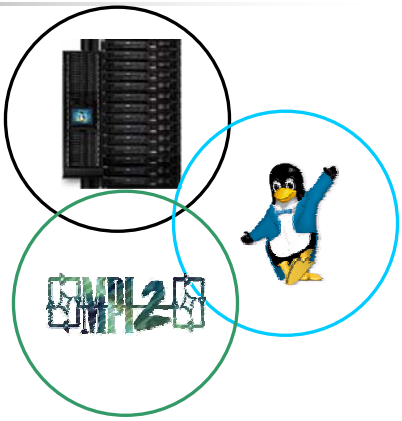
Prof. Gergel Victor,
Nizhni Novgorod, Russia
2005

Contents

- Project objectives
- Cluster architecture model
- Effective using of shared memory
- Optimizing algorithms for two-level cluster architecture
- Conclusions

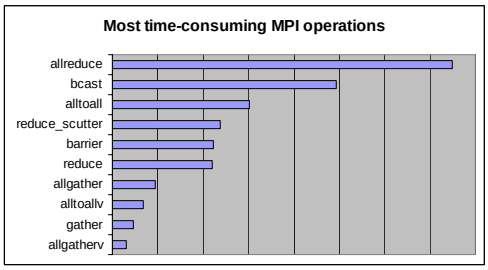
Project objective

- Increasing efficiency of parallel applications
  - run on POWER clusters,
  - under Linux,
  - developed using open-source implementations of MPI



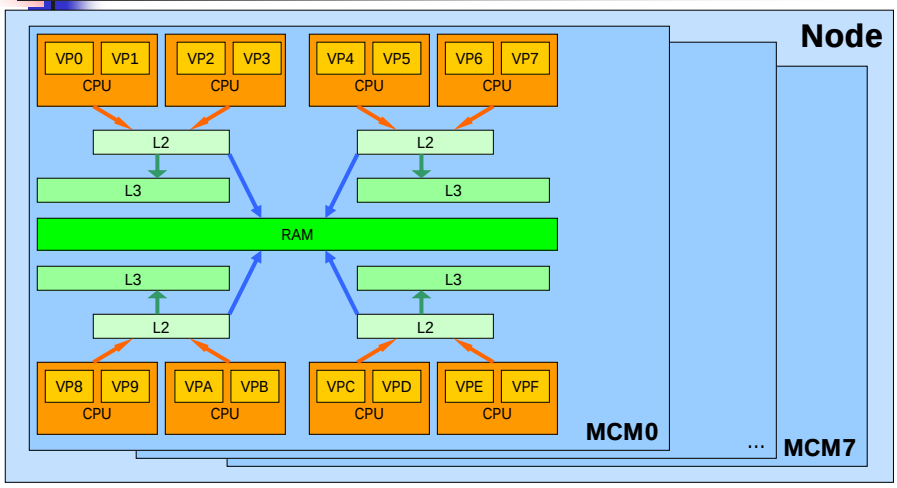
Main lines of investigation

We decide to consider collective operations because they are most time-consuming procedures in MPI

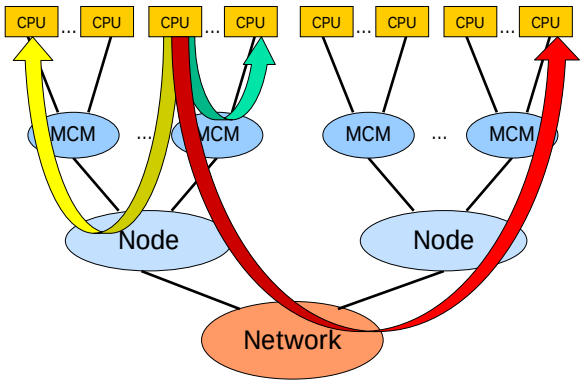


Rolf Rabenseifer. Automatic MPI Counter Profiling. 42nd GUG Conf.

Analyzing POWER5 cluster. Node architecture



Analyzing POWER5 cluster. Architecture levels

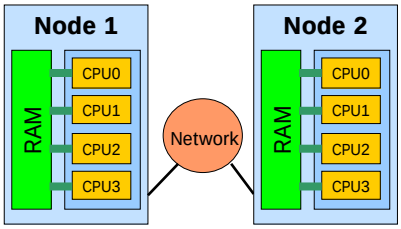


- Virtual processor
- Core
- Chip
- Multi-chip module (MCM)
- Book
- Node
- Network, Cluster

Charles Grassl "POWER5 and HPS Programming Strategies: System Architecture", IBM Corporation, 2004
Ron Kalla "IBM's POWER5 Microprocessor Design and Methodology", UT Computer Seminar, 2003

Analyzing POWER5 cluster. Cluster architecture model restriction

- Currently we restrict consideration with two levels of architecture (1st stage of project)
  - data transfer inside SMP-node over shared memory,
  - data transfer between SMP-nodes over network



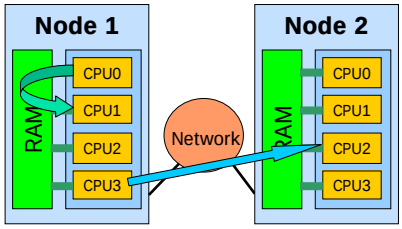
Analyzing POWER5 cluster. Applying Hockney model...

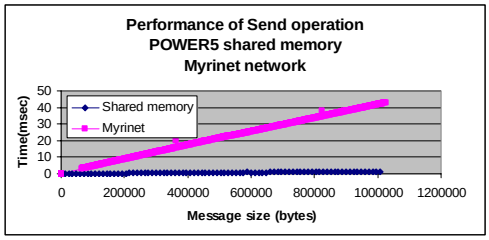
- Hockney model allows to estimate cost of message transfer using following parameters
  - α - latency (time to preparing data for transfer),
  - β - time for transferring 1 byte of data,
  - n – message size

$$T_{transfer} = \alpha + \beta * n$$

Analyzing POWER5 cluster. Applying Hockney model...

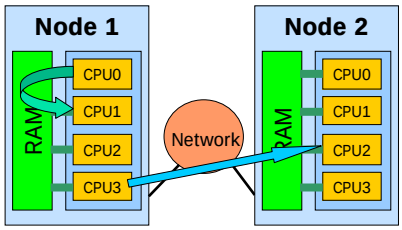
- POWER5 shared memory
  - $\alpha_{sh\_mem} \approx 1*10^{-6}$
  - $\beta_{sh\_mem} \approx 1.4*10^{-9}$
- Myrinet
  - $\alpha_{network} \approx 4*10^{-5}$
  - $\beta_{network} \approx 2.6*10^{-8}$

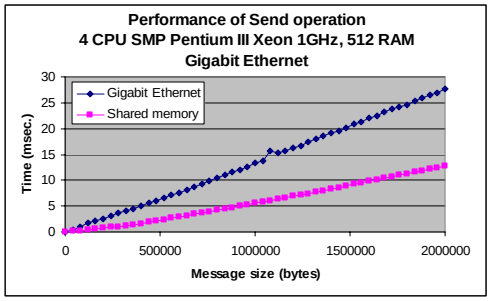




Analyzing POWER5 cluster. Applying Hockney model

- P-III shared memory
  - $\alpha_{sh\_mem} \approx 4*10^{-6}$
  - $\beta_{sh\_mem} \approx 6.6*10^{-9}$
- Gigabit Ethernet
  - $\alpha_{network} \approx 10^{-4}$
  - $\beta_{network} \approx 1.4*10^{-8}$

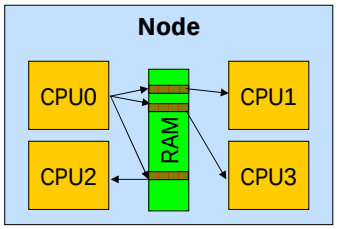




Effective using of shared memory

Using shared memory. Standard algorithms

- In case of transferring the same data from one process to several another processes data is transferred for each process with individual operation
- For transferring between each processes pair separate shared memory window is used

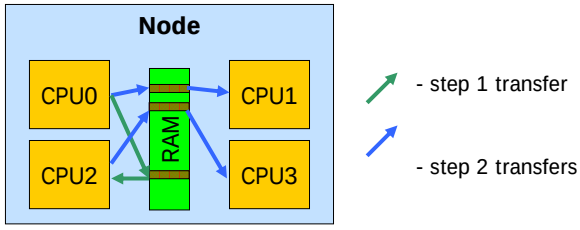


Using shared memory. Binomial tree Bcast algorithm

- Operation cost

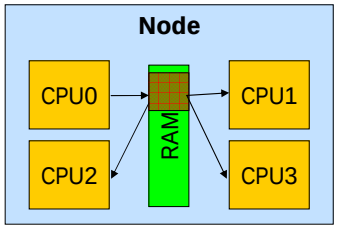
$$T_{Bcast} = (p-1) * (\alpha_{sh\_mem} + \beta_{sh\_mem} * n)$$

p – number of processes,
n – message size



Using shared memory. Optimized algorithms

- In case of transferring the same data from one process to several another processes data is transferred for every process in one operation
- Single shared memory window is used for data transfer



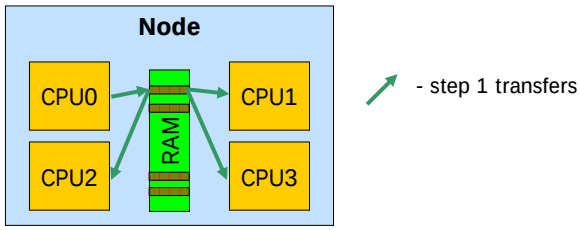
Using shared memory. Optimized Bcast algorithm

- Operation cost

  $T_{Bcast} = p/2 * (\alpha_{sh\_mem} + \beta_{sh\_mem} * n)$

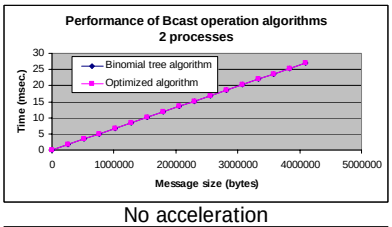
  p – number of processes,

  n – message size

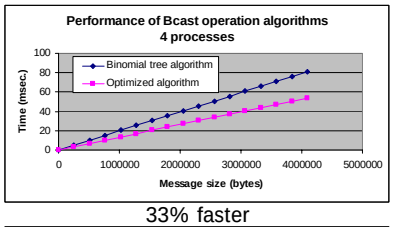


Using shared memory. Comparing Bcast algorithms...

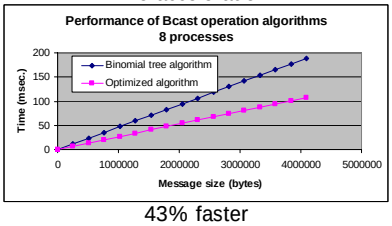
Theoretical estimation



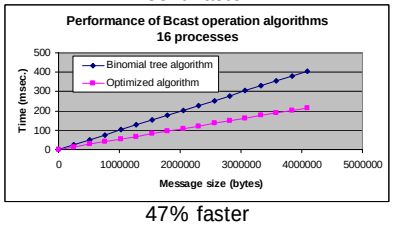
No acceleration



33% faster



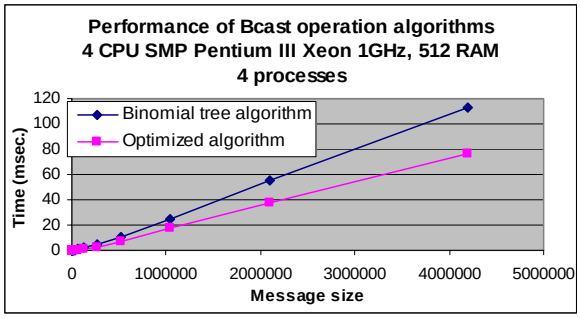
43% faster



47% faster

Using shared memory. Comparing Bcast algorithms

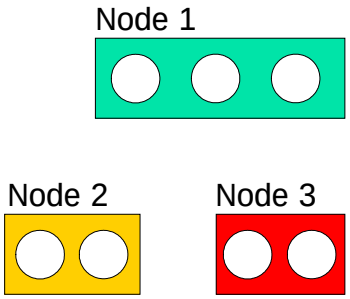
Test results



31% faster

Optimizing algorithms for two-level cluster architecture

Two-level cluster architecture...



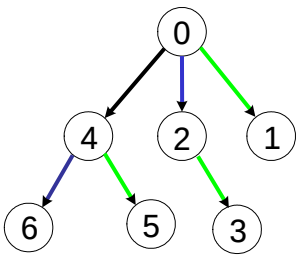
- process

• Processes within SMP-node can interact through shared memory

• Processes running on different nodes must use network for data transfer

Two-level cluster architecture. Bcast algorithm...

- Binomial tree algorithm



k - k-th process

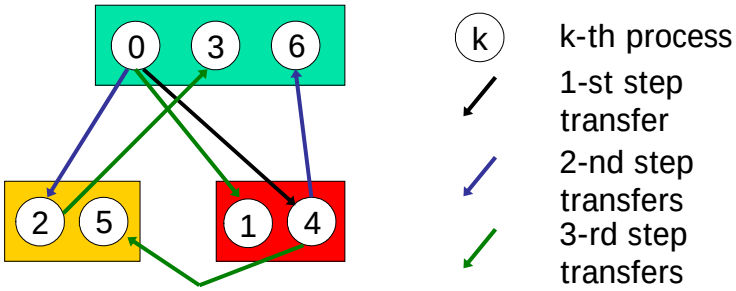
1-st step transfer

2-nd step transfers

3-rd step transfers

Two-level cluster architecture. Bcast algorithm...

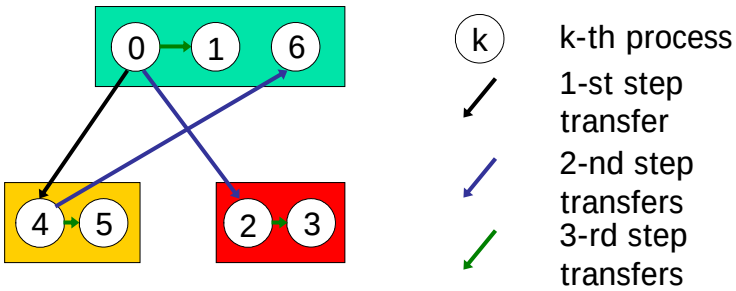
- Standard process numeration



- On each step we send data over network
- $T_{Bcast} = 3*(\alpha_{network} + \beta_{network} * n)$

Two-level cluster architecture. Bcast algorithm...

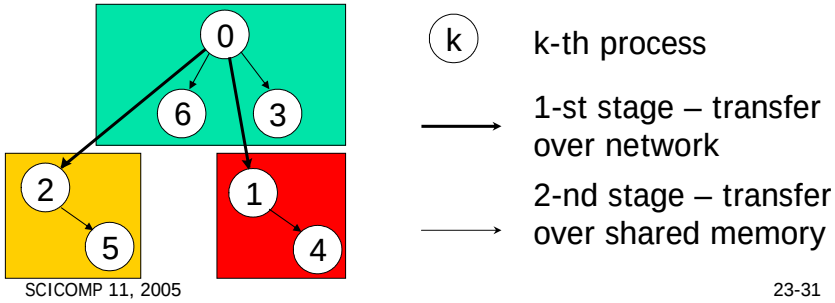
- Another process numeration



- On 3-rd step we transfer data only over shared memory
- $T_{Bcast} = 2*(\alpha_{network} + \beta_{network} * n) + (\alpha_{sh\_mem} + \beta_{sh\_mem} * n)$

Two-level cluster architecture. Bcast algorithm...

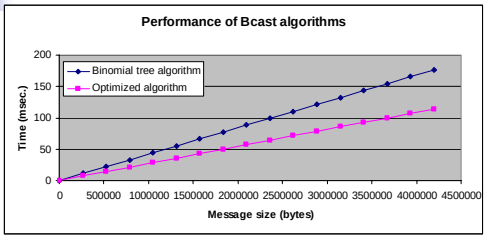
- Optimized algorithm
  - use binomial tree algorithm for transferring message to all networks nodes,
  - use binomial tree algorithm for transferring message to all processes on every SMP-node



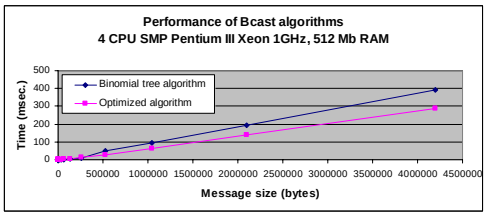
Two-level cluster architecture. Bcast algorithm...

- 8 processes running on 2 4-processor SMP nodes (4 processes on each node)
- Binomial tree algorithm
  - $T_{Bcast} = [\log_2 8]*(\alpha_{network} + \beta_{network} * n)$
    $= 3*(\alpha_{network} + \beta_{network} * n)$
- Optimized 2-stage algorithm which use binomial tree algorithm on both stages
  - $T_{Bcast} = [\log_2 2]*(\alpha_{network} + \beta_{network}*n) +$
    $+ [\log_2 4]*(\alpha_{sh\_mem} + \beta_{sh\_mem} * n)$
    $= (\alpha_{network} + \beta_{network}*n) + 2*(\alpha_{sh\_mem} + \beta_{sh\_mem} * n)$

Two-level cluster architecture. Bcast algorithm



Theoretical estimation
35% Acceleration



Test results
26% Acceleration

Conclusions

- 2-level model of POWER based clusters is proposed,
- Optimized scheme how to implement shared memory collective operations is developed,
- Effective two-stage procedures of collective operations in case of two-level (shared/ distributed) memory are worked out

Research group

- Gergel V.P., professor
- Grishagin V.A., associate professor
- Belov V.A., associate professor
- Linev A.V.
- Gergel A.V
- Grishagin A.V.
- Kurylev A.L.
- Senin A.

Contacts

- 603950, Nighny Novgorod
  Gagarina av., 23
  Nizhny Novgorod State University
  Applied Mathematics and Cybernetics faculty
- Tel: +7 (8312) 654859
- E-mail: gergel@unn.ac.ru
  vagris@unn.ac.ru
  belov@vmk.unn.ru

Nizhni Novgorod – city and region

Nizhni Novgorod has been founded in 1221 by Yuri Vselodovich.

Nizhni Novgorod is the capital of Volga river federal district (7% of territory, 22% of population of Russia).

"Third capital" of Russia.

29-31

University of Nizhni Novgorod

www.unn.ru

UNN is the **first state university**, organized in Soviet Union (1918). Nowadays, by official rating of the Ministry of Education, UNN is **among top 10 universities in Russia**

- 27 faculties (departments)
- 122 chairs (subdepartments)
- 6 research institutes
- over 1000 professors
- over 1000 PhD students
- over 26000 students

Nobel Prize winner (2004) professor **Ginzburg** worked at UNN (radiophysics faculty) for more then 20 years

SCICOMP 11, 2005

Thank you for attention

- Questions,
  - Remarks,
    - Comments