A practical Approach to the Rating of Barrier Algorithms using the LogP Model and Open MPI

Torsten Höfler, Wolfgang Rehm
TU Chemnitz, Germany

24.05.2005

CHEMNITZ UNIVERSITY
OF TECHNOLOGY

Outline

- Motivation

1. LogP Predictions
2. Implementation
3. Conclusions

CHEMNITZ UNIVERSITY
OF TECHNOLOGY

Outline

- Motivation

CHEMNITZ UNIVERSITY OF TECHNOLOGY

Motivation

- optimal solution for the barrier problem
- barrier time complexity studies
- exhaustive comparison of different algorithms
- framework for general comparison studies
- Open MPI is easily extensible
- Question: is LogP accurate enough?

Problems

- unlimited number of architectures
  - generic optimal solution = holy grail?
- definition of several constraints for a given architecure
  - Fast Ethernet, Extreme Black Diamond Switch, 512 nodes
- new architectures have to be added by hand
- several models available -> LogP should be accurate enough

Principles

- one architecture as example
- easy testing of new architectures
- framework to implement and test new algorithms

Architectural Assumptions

- full bisectional bandwidth
- full duplex operation
- unlimited switch forwarding rate
- constant latency
- overhead bigger than gap
- overhead is constant ($o_s = o_r$)

CHEMNITZ UNIVERSITY
OF TECHNOLOGY

Base Equations

$$
\begin{aligned}
several & \quad basic\ equations\ and\ variables: \\
f_r &= max\{o_r, g\} \\
f_s &= max\{o_s, g\} \\
t_r &= max\{f_r, o_s + L + o_r\} \\
&= max\{max\{g, o_r\}, o_s + L + o_r\} \\
&= max\{g, o_s + L + o_r\} \\
simplifying & \quad assumptions: \\
f_r &= f_s = o \\
t_r &= t_s = 2o + L
\end{aligned}
$$

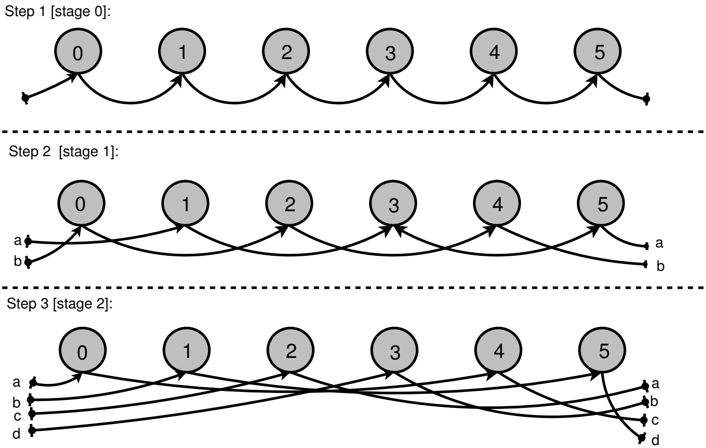
CHEMNITZ UNIVERSITY
OF TECHNOLOGY

Model Predictions

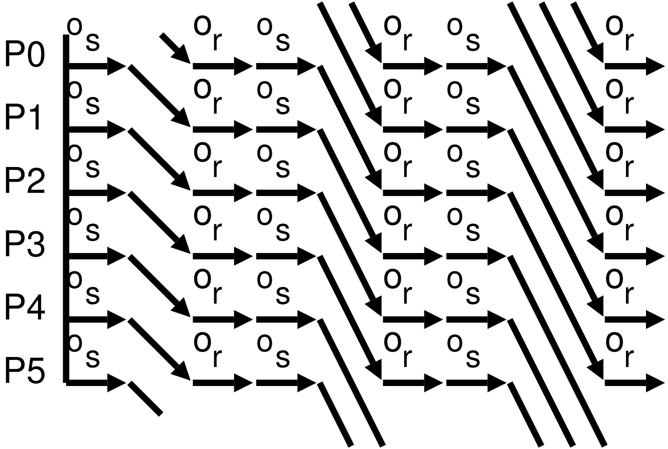
- algorithms are divided into different complexity classes
  - $O(P) \Rightarrow$ **Central Counter**
  - $O(n \cdot log_n P) \Rightarrow$ **Combinig Tree**, f-way Tournament, MCS
  - $O(log_2 P)$ + Bcast $\Rightarrow$ **Tournament**, BST
  - $O(log_2 P) \Rightarrow$ Butterfly, Pairwise Exchange, **Dissemination**
- $O(log_2 P)$ within the LogP is an optimal solution
- prove is trivial
- Assumption: Dissemination Barrier should perform best

CHEMNITZ UNIVERSITY OF TECHNOLOGY

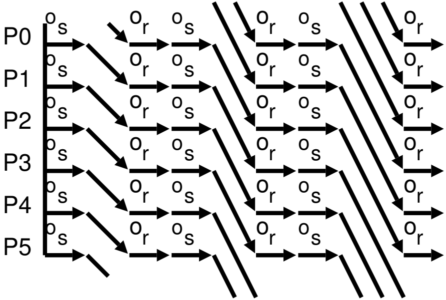
Example - Dissemination Barrier



Example - Dissemination Barrier



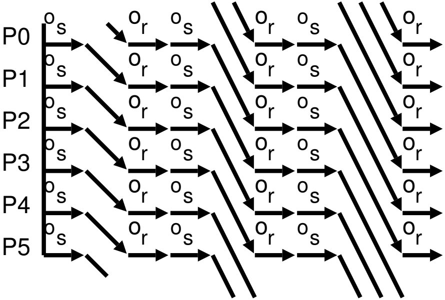
Example - Dissemination Barrier



$$
\begin{aligned}
rt &= max\{t_r, t_s\} \cdot \lceil log_2 P \rceil \\
(t_r &= max\{g, o_s + L + o_r\})
\end{aligned}
$$

CHEMNITZ UNIVERSITY OF TECHNOLOGY

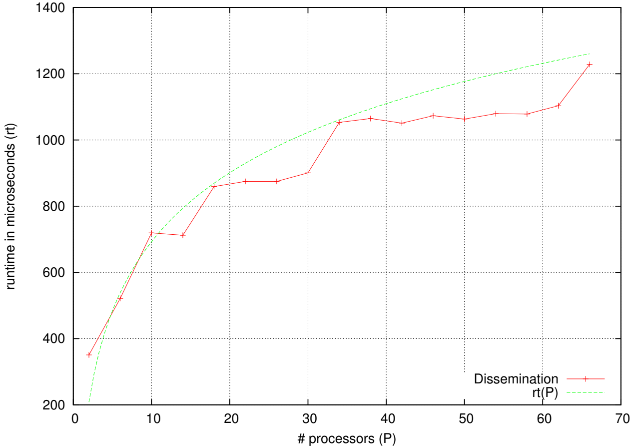
Example - Dissemination Barrier



$$assume: \quad o > g$$
$$rt = (2o + L) \cdot \lceil log_2 P \rceil$$

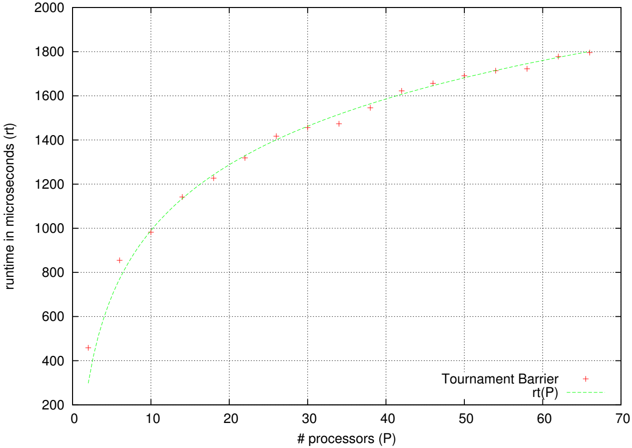
CHEMNITZ UNIVERSITY OF TECHNOLOGY

Benchmark Results



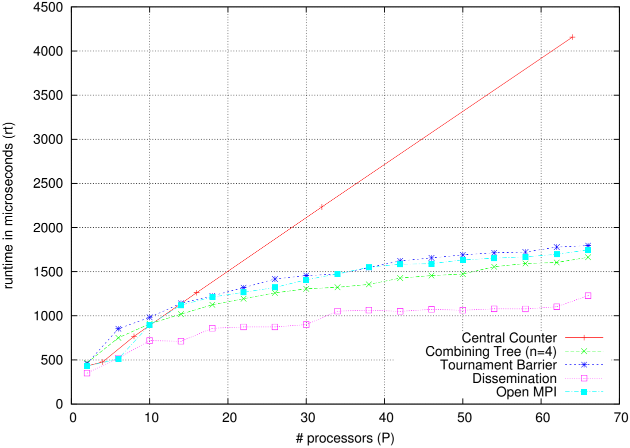
Dissemination Barrier

Benchmark Results



Tournament Barrier

Benchmark Results



Algorithm Comparison

Benchmark Results

Algorithm	**128 nodes**	**256 nodes**
Central Counter	4594.50μs	4909.67μs
Combining Tree	4009.79μs	4343.63μs
Tournament	3642.54μs	4378.77μs
Dissemination	1904.57μs	1977.12μs
Open MPI	3559.88μs	4226.88μs

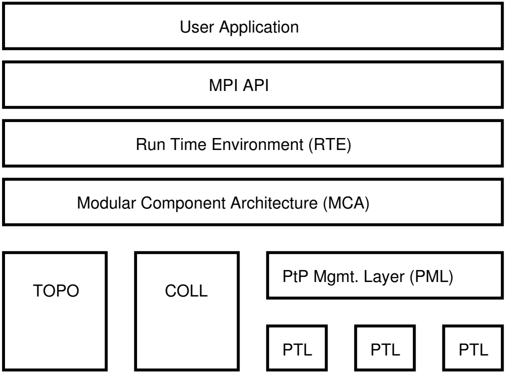
CHEMNITZ UNIVERSITY
OF TECHNOLOGY

Open MPI

- also useable for production environments
- $\Rightarrow$ Open MPI as MPI framework

Open MPI

- also useable for production environments
- $\Rightarrow$ Open MPI as MPI framework



CHEMNITZ UNIVERSITY OF TECHNOLOGY

Component Implementation

- initialization returns user-defined priority
- algorithm selection:
  - 0: automatic benchmark
  - 1: Central Counter
  - 2: Combining Tree
  - 3: Tournament
  - 4: Dissemination
  - 5: Binomial Tree
  - 6: n-way Dissemination
- Checkpoint/Restart is handled by lower layers

Conclusions

- taken assumptions are valid
- LogP model is accurate
- Dissemination is optimal for given scenario
- different networks exhibit different behavior
- derivation of new algorithms for different hardware (e.g. offloading based HW) could require detailed models
- $\Rightarrow$ general methodology for developing optimal barrier algorithms has been shown

Future Work

- new model for small messages for offloading based NICs (LoP)
- new barrier algorithms to support hardware parallelism
- simplification of the LoP model (non linear, >6 parameters)

CHEMNITZ UNIVERSITY OF TECHNOLOGY