A Case for Application-Oblivious Energy-Efficient MPI Runtime

Paper Reading Group

Akshay Venkatesh
Abhinav Vishnu
Khaled Hamidouche
Nathan Tallent
Dhabaleswar (DK) Panda
Darren Kerbyson
Adolfy Hoisie
Presents: Maksym Planeta

12.11.2015

Table of Contents

Table of Contents

Where do the joules go?

- Computations (we want it)
- Communication (we can't avoid this)
- MPI library (here is the target)

MPI energy consumption

What is slack?

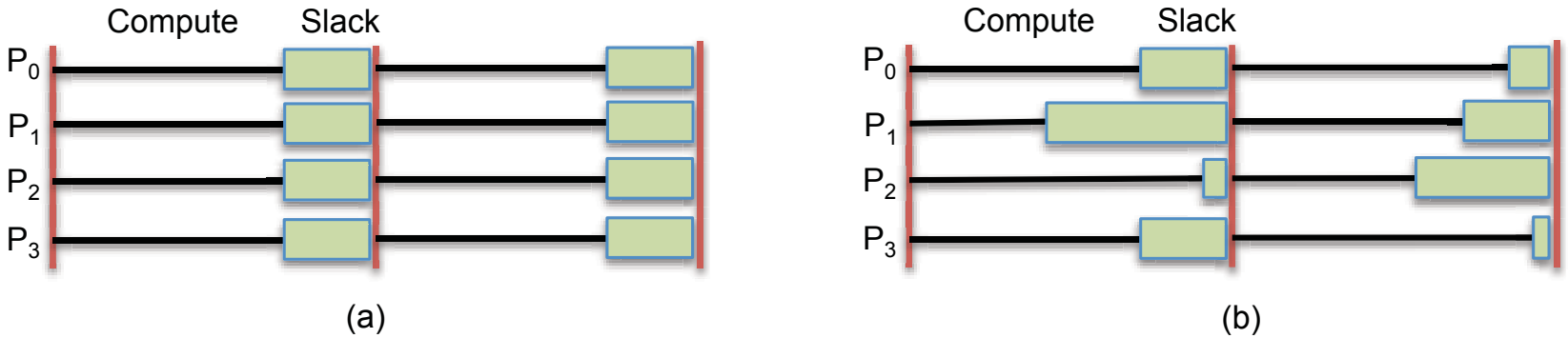


Figure 2: (a) Iterative/Temporal Pattern and (b) Iterative/Non-temporal Pattern

We define slack to be the actual time spent by an MPI process in a single MPI call...

Idea

- User specifies accepted overhead ρ
- Set of power levers: $L = (\delta, \gamma, \psi)$
- Overhead of a lever: γ
- Time threshold for a lever: $\delta = \frac{\gamma}{\rho}$
- Power improvement: ψ

Lever types

1. Polling ($\psi = 0, \delta = 0$)
2. Blocking
3. DVFS (not evaluated)

Table of Contents

LogP

How to model communication

- L an upper bound on the *latency* between messages
- o the *overhead*, CPU time required to process a message of each message
- g is a *gap*, interval between messages
- P number of *processors*

LogP broadcast

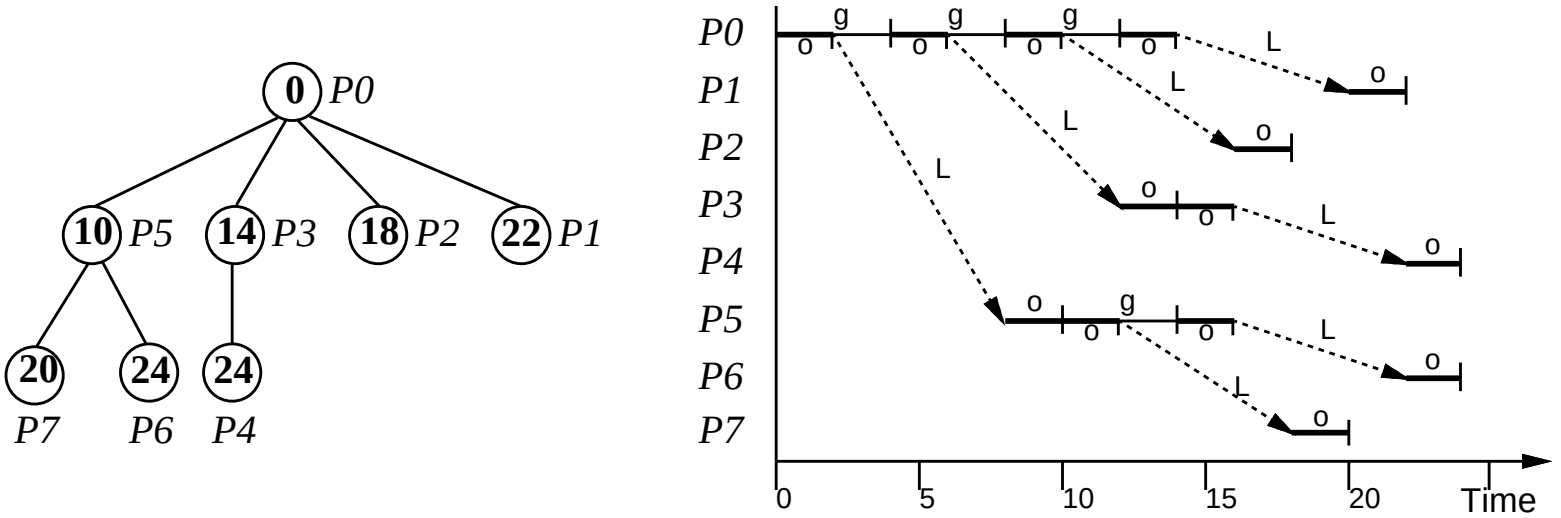


Figure 3: *Optimal broadcast tree for* $P = 8, L = 6, g = 4, o = 2$ *(left) and the activity of each processor over time (right). The number shown for each node is the time at which it has received the datum and can begin sending it on. The last value is received at time 24.*

LogGP

Messages can be big or small

- g is a *gap* between small messages
- G the *Gap* per byte for long messages,

A lot of them...

- LogGOP
- LogGPS
- MLogP
- others

Table of Contents

Challange

Decide to use a method before the slack is known

Be communication aware

Message size

- Eager
- Randezvous

Synchonization

- Blocking
- Non-blocking

Participants

- Point-to-point
- Collective

Lever example

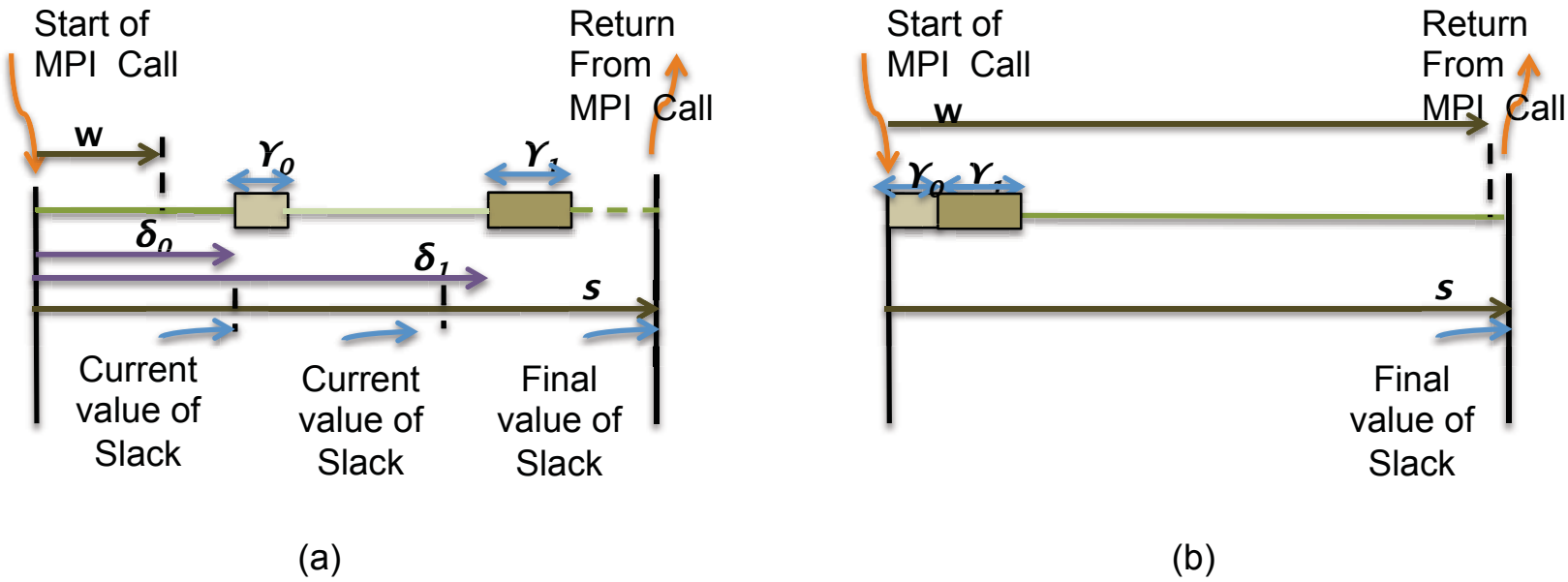


Figure 3: An example of using two power levers in EAM. Left: expected communication time is much lesser than slack, levers are applied as their thresholds are crossed. Right: Expected communication time exceeds the thresholds for each lever. The power levers are applied at the start of the MPI call, maximizing the energy efficiency

Expected comunication time

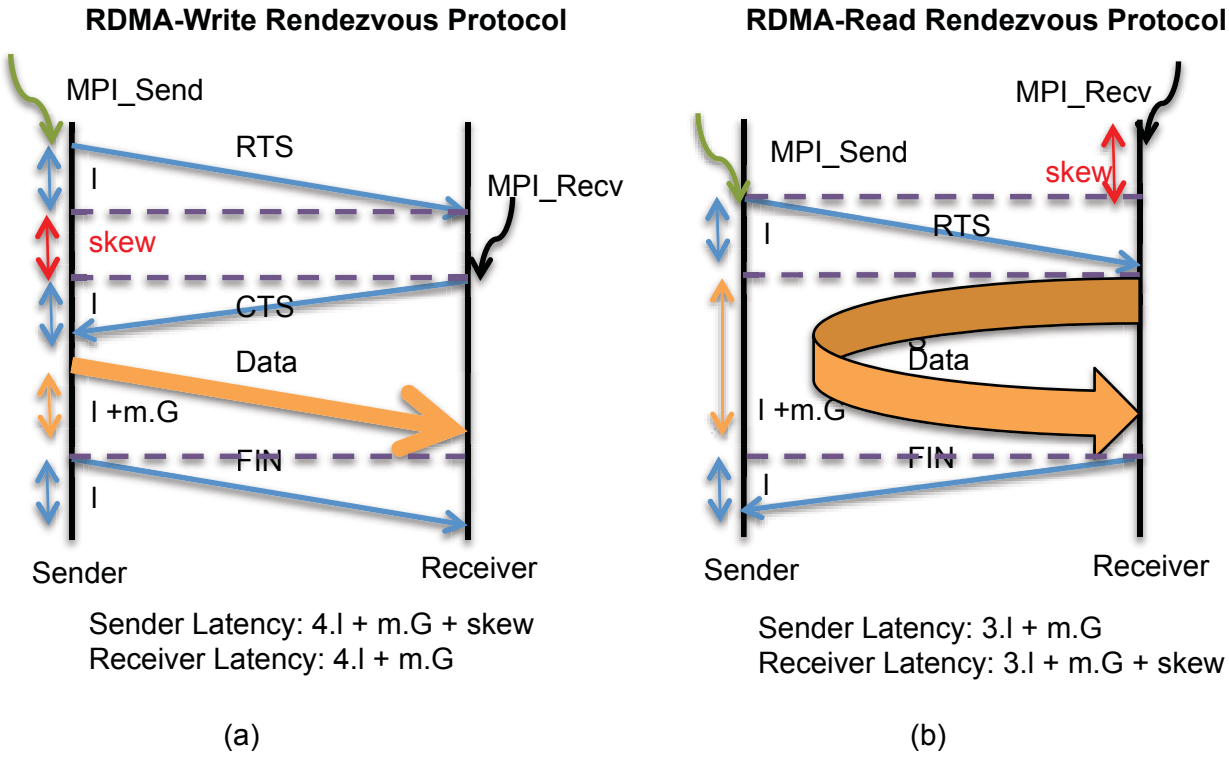
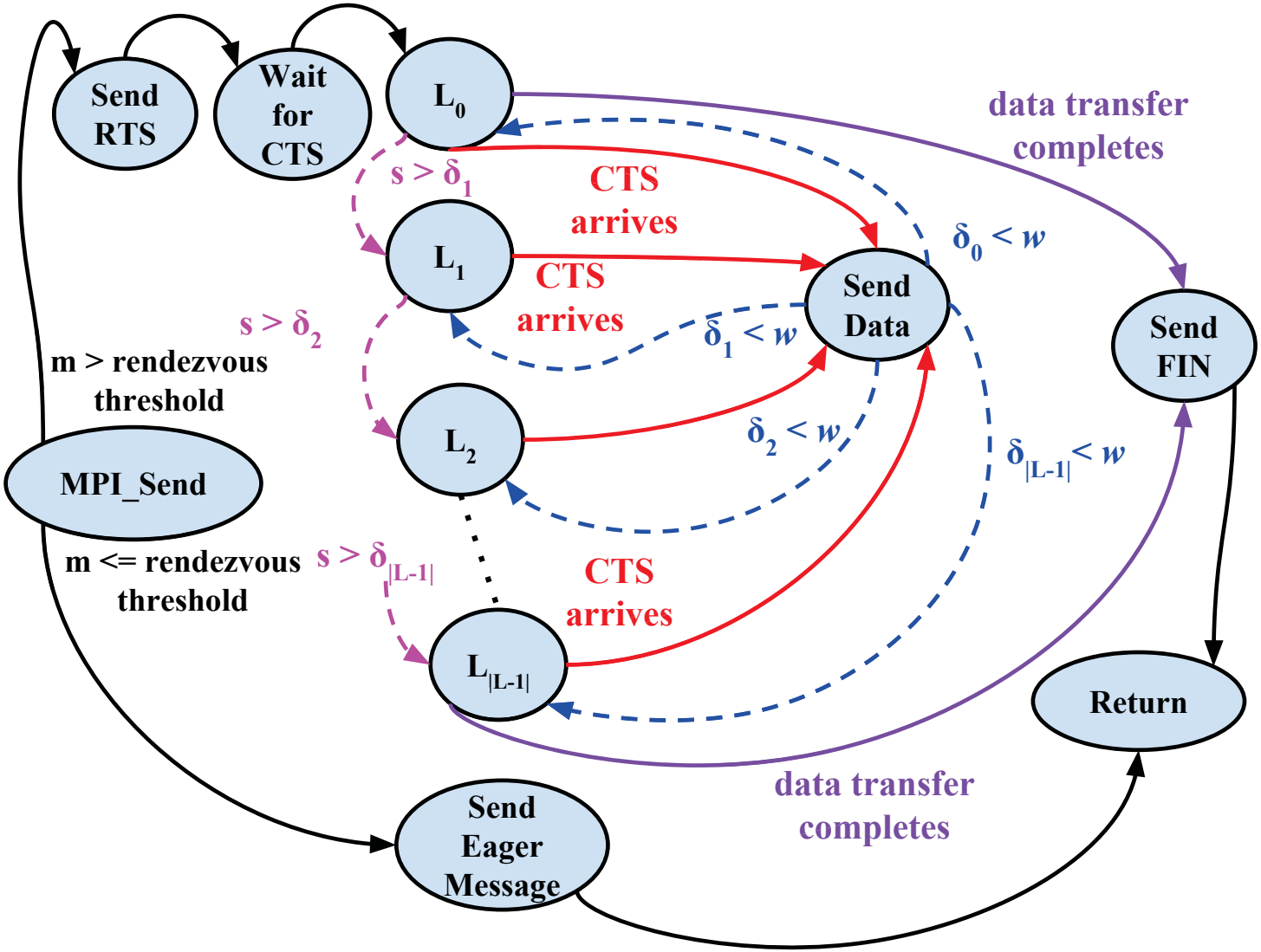


Figure 4: RDMA-Write and RDMA-Read based rendezvous protocols. Figure (a) shows delayed receiver, and Figure (b) shows delayed sender. The delay is referred as skew.

$$\underbrace{l+o+m_{RTS}\cdot G}_{\text{RTS}}+\underbrace{l+o+m_{CTS}\cdot G}_{\text{CTS}}+\underbrace{l+o+m\cdot G}_{\text{payload}}+\underbrace{l+o+m_{FIN}\cdot G}_{\text{FIN}}.$$ Since control messages are small, $w = (4\cdot l+m\cdot G)$ $(o \ll l)$. However, this time is a lower bound for

State transitions



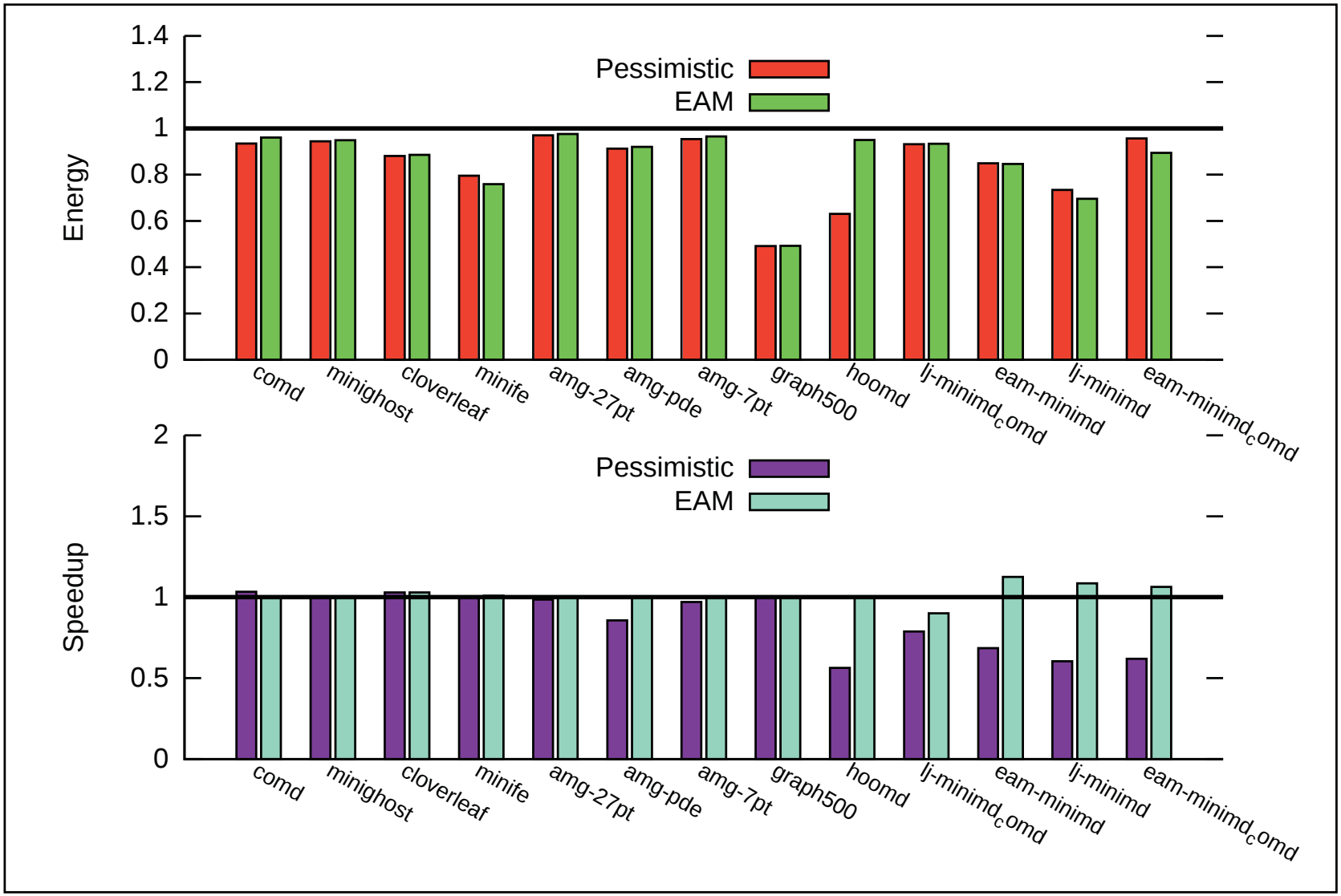
(a) RDMA-Write: Sender's Eager and Rendezvous transition rules

Table of Contents

Types of setup

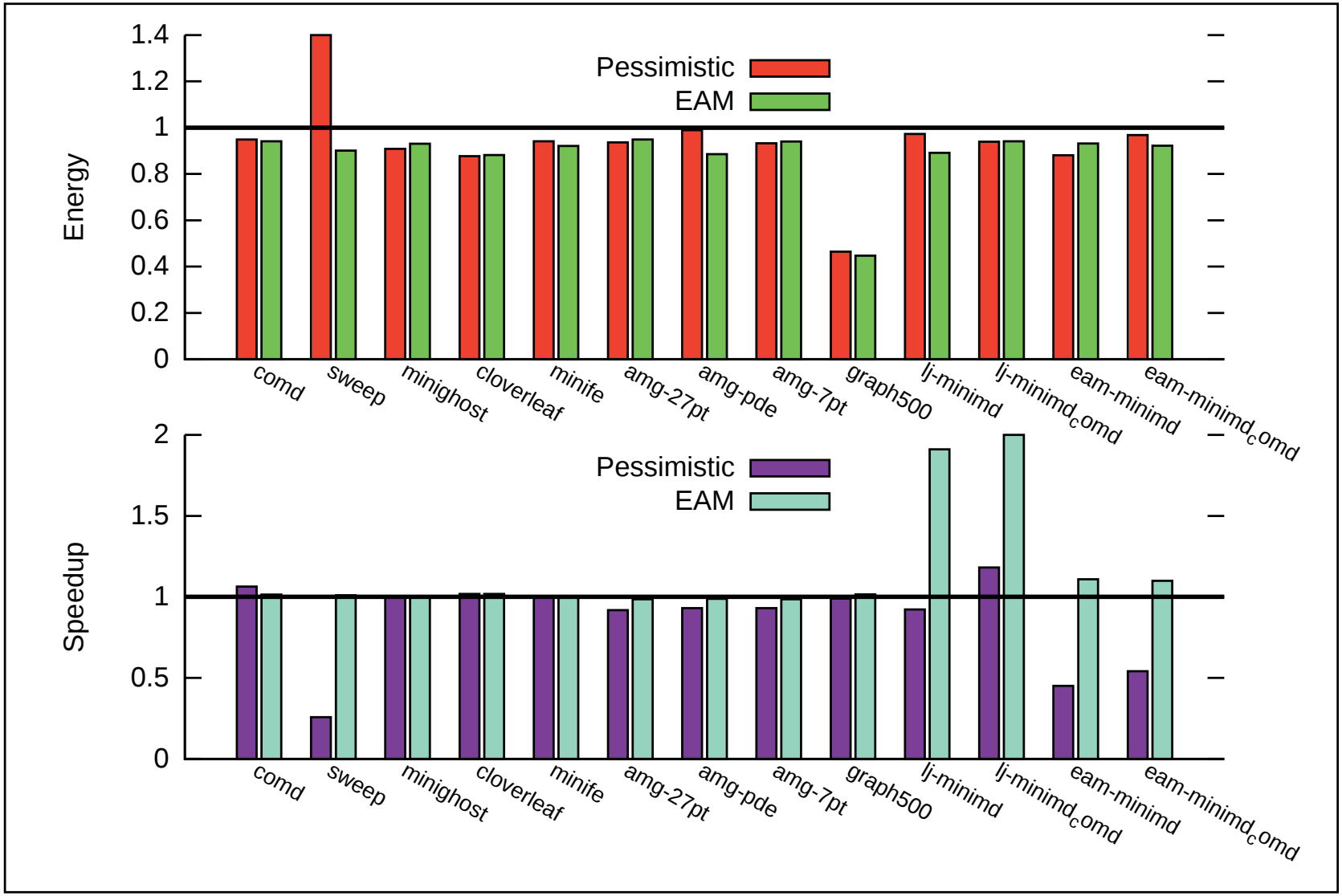
1. Pessimistic
2. Optimistic
3. EAM

Small



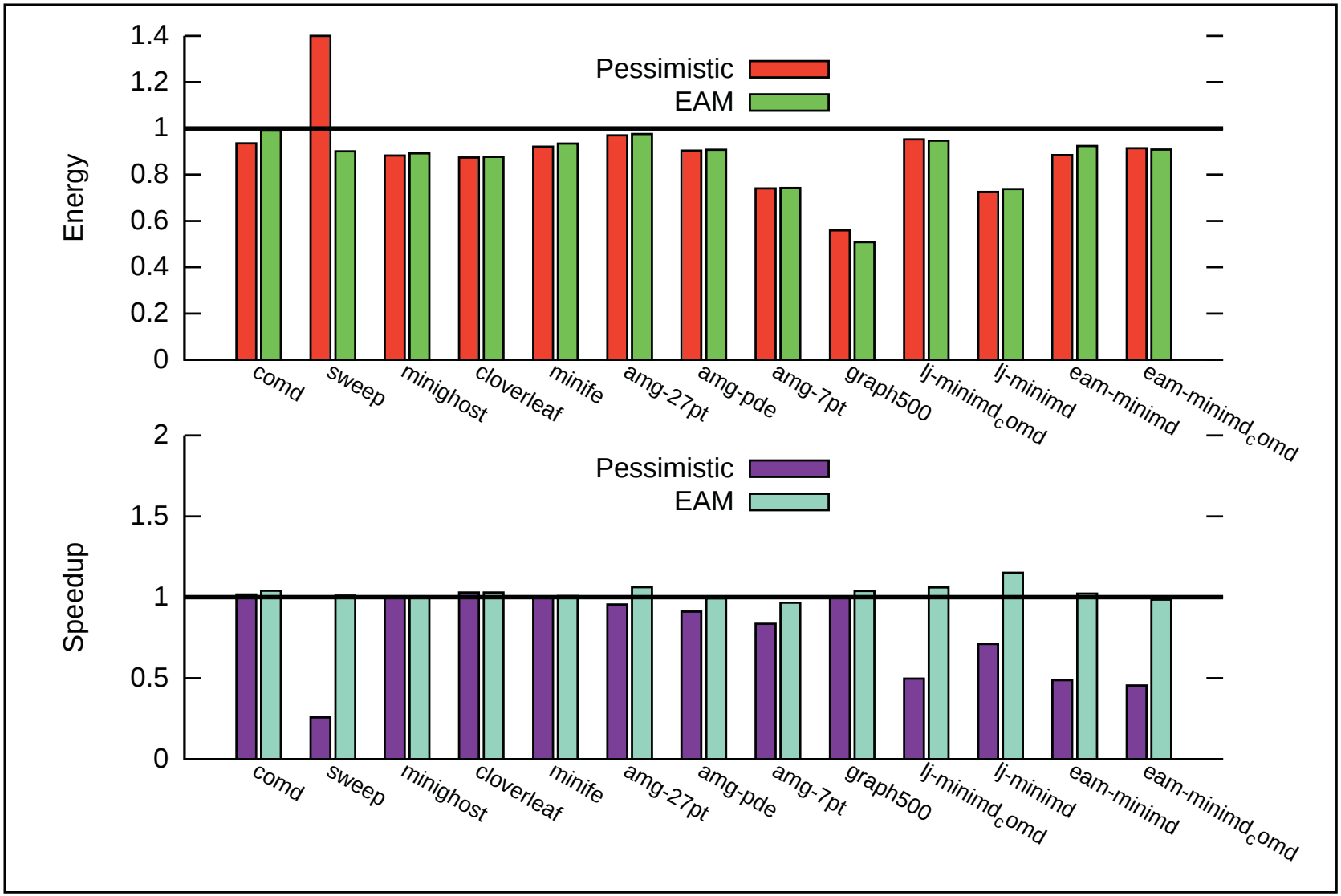
(a) 512 Processes

Medium



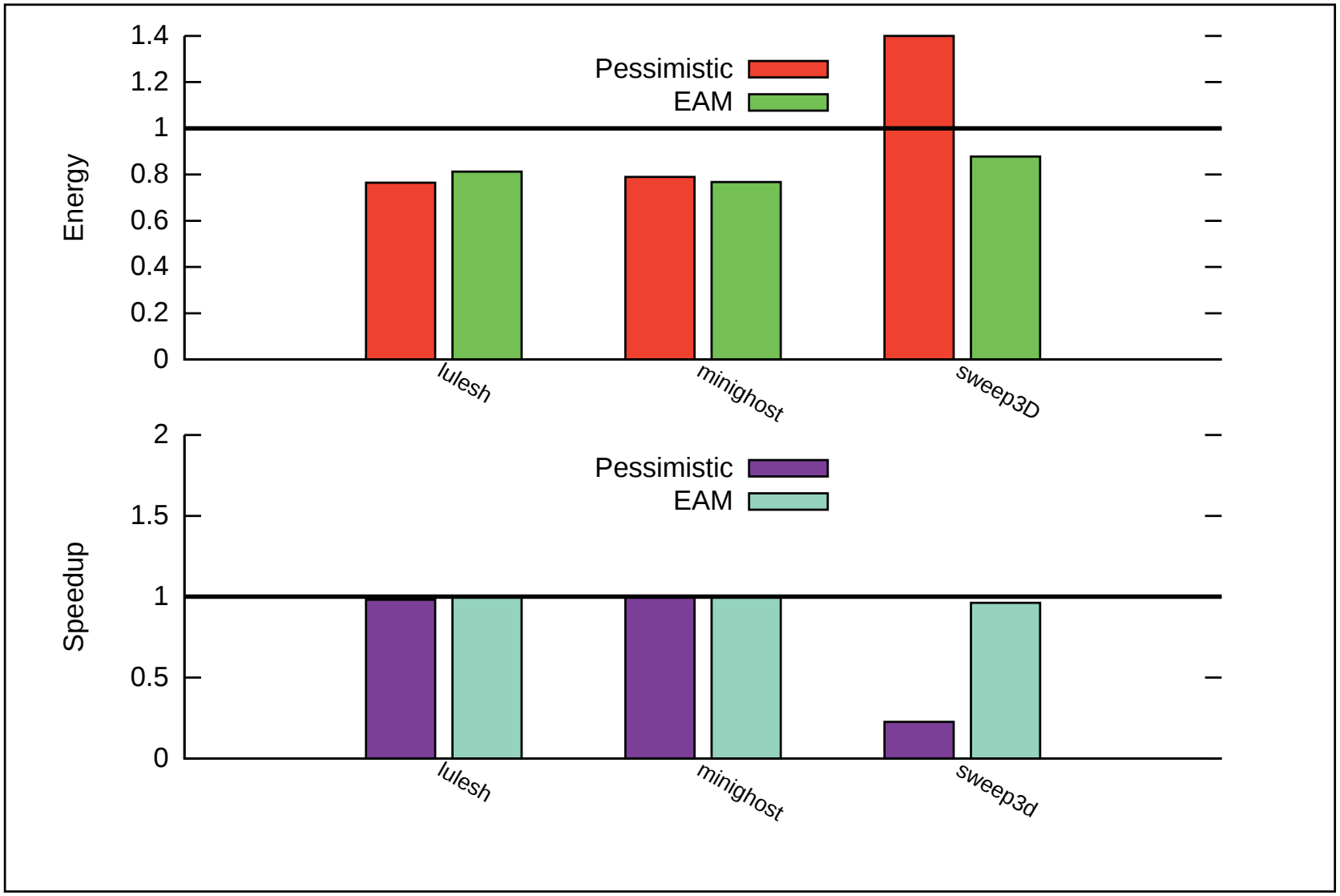
(b) 1,024 Processes

Large



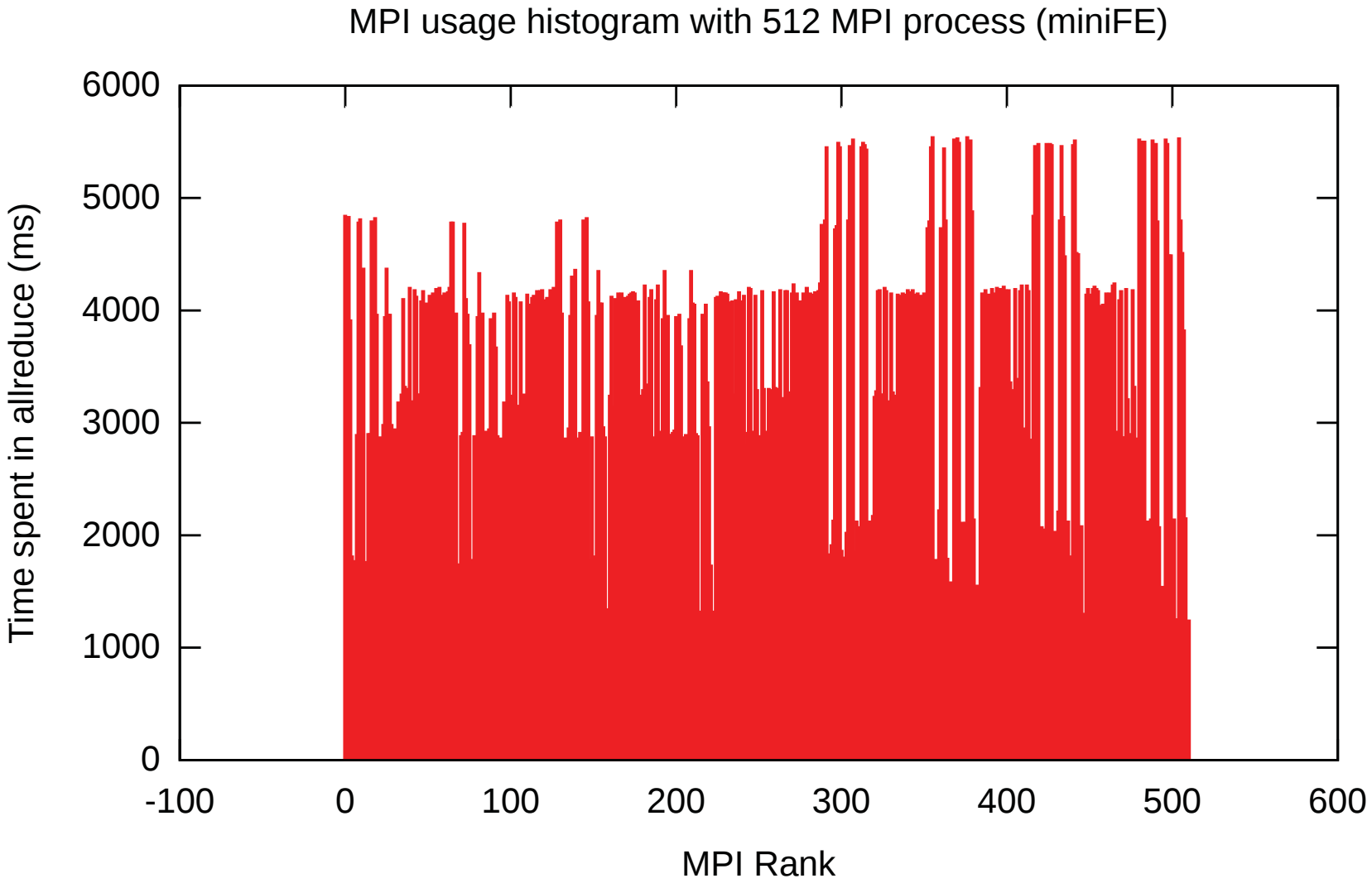
(c) 2,048 Processes

Extra Large



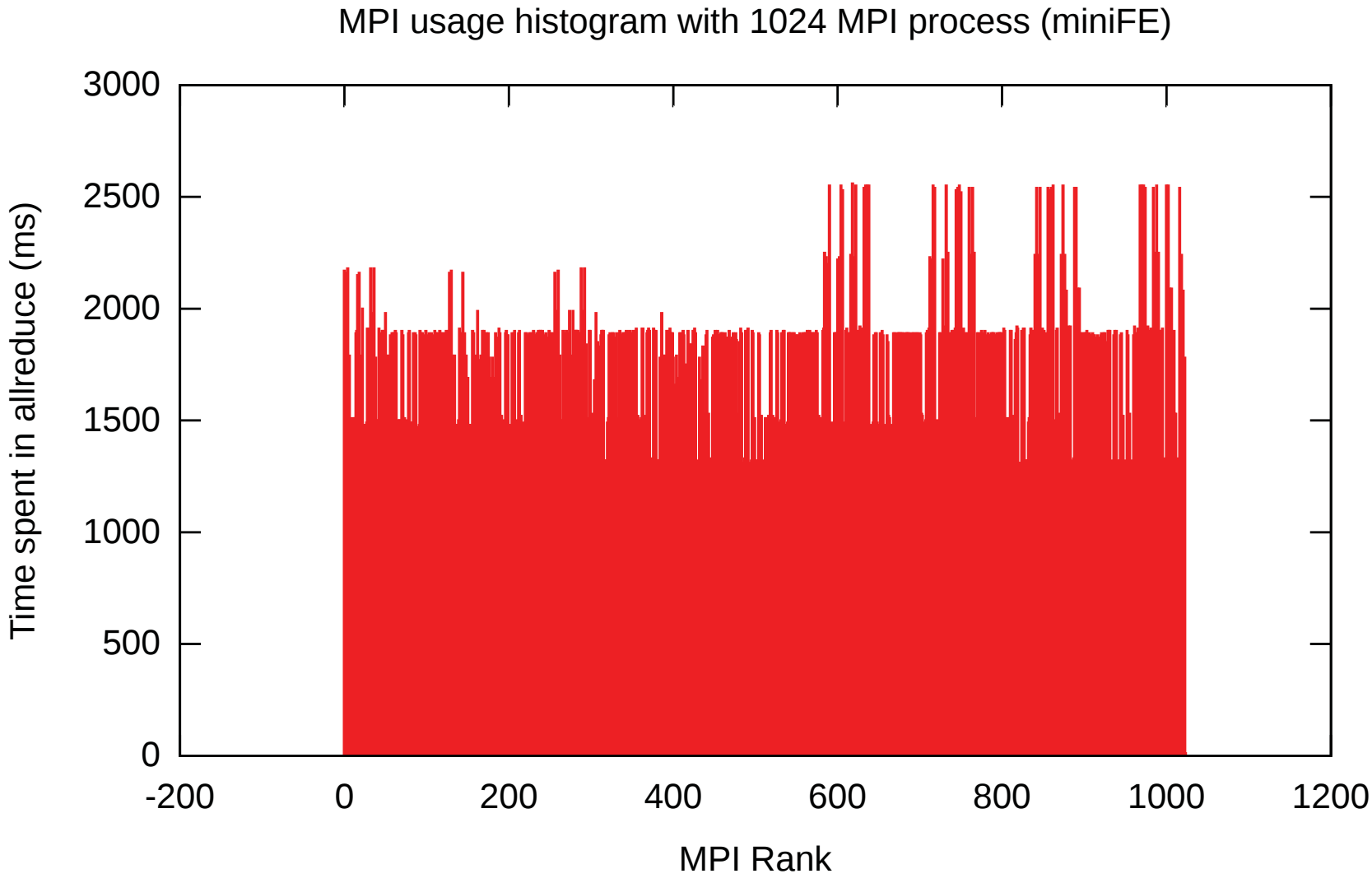
(d) 4,096 Processes

von Parteimueller would fit here



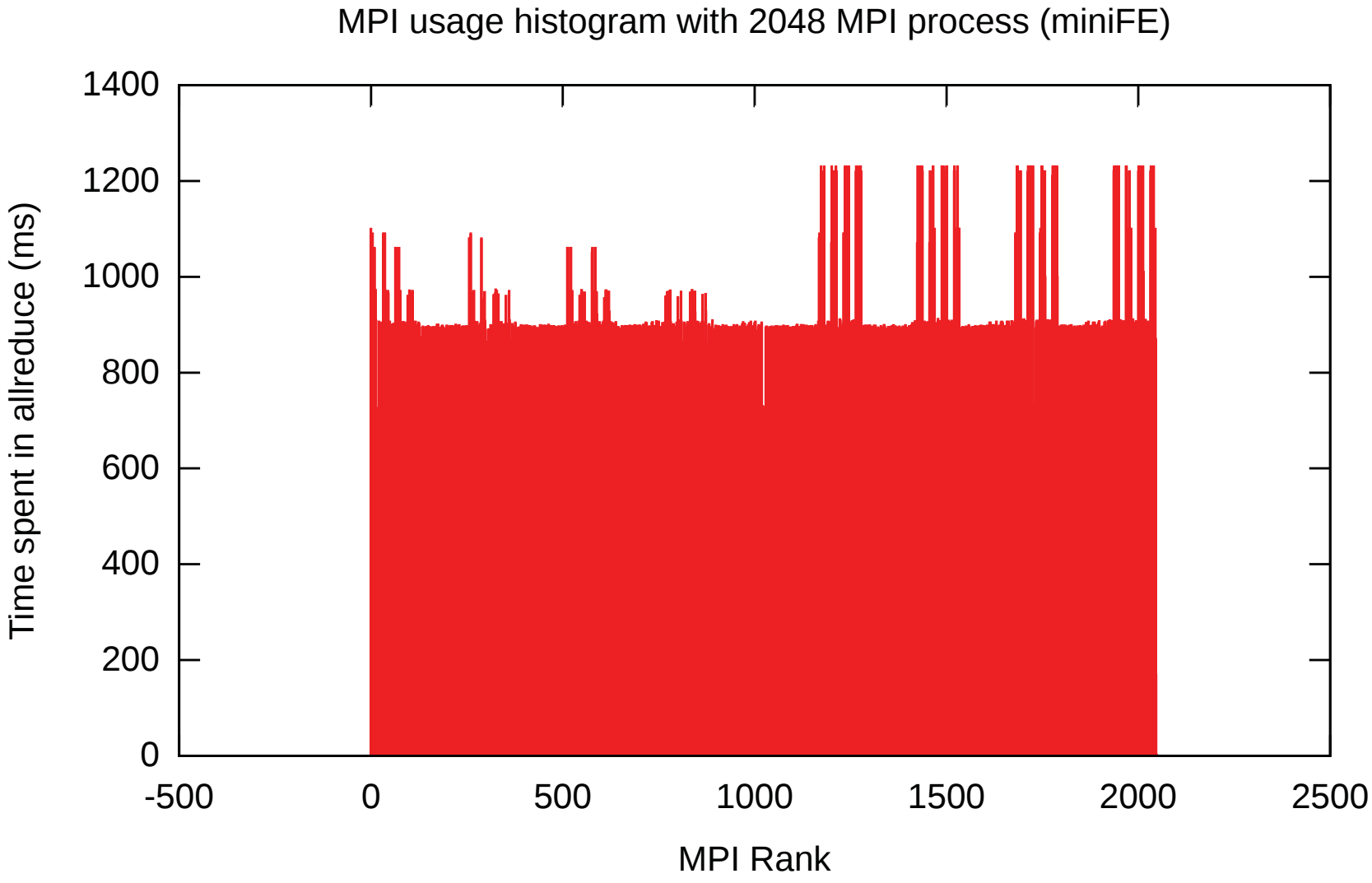
(a) 512 Processes

but they didn't know about it



(b) 1,024 Processes

although even boxplot would be nicer



(c) 2,048 Processes

Table of Contents

Conclusion

- Obvious goal
- Simple idea
- Good formalization
- Good results