High performance libraries for Windows 2000 : from a developer standpoint

Laurent Lefèvre, Roland Westrelin

UCBL

RESAM
Réseaux haut-débit et Supports d'Applications Multimédia

ECOLE NORMALE SUPERIEURE DE LYON

INRIA
RHÔNE-ALPES

Supported by: Microsoft

Basic Interface for Parallelism:BIP

- Developed by:
  - Patrick Geoffray, Loïc Prylli, Bernard Tourancheau and Roland Westrelin
  - at ENS Lyon and University of Lyon (France)
- Goals:
  - Maximization of the application level performance (close to the hardware maximum)
  - Providing legacy programming interfaces: MPI (MPI-BIP based on MPICH) and IP
- Initial targets: Myrinet, Linux
- Composed of: libraries (BIP, MPI-BIP), an OS module, a firmware (architecture and OS independent), a basic runtime

From Linux to Win2000: the OS module

- A driver
- An IP driver
- Provide direct access to the Network Interface and Memory management

➡ The idea is to use the GM OS module (available on several architecture, in 2 layers)

- Easier
- Incremental
- Trick: use winNT driver (not win2k)

From Linux to Win2000: the libraries

- Rely on the cygwin porting layer: emulate UNIX libc calls with WIN32 calls
- Makes maintenance easier
- gcc!
- Open-source, widely used, quickly improving
- No syscall in the critical path of communication in BIP (memory registration is cached)
- Still have to rewrite part of the library in native code (driver calls, synchronization)
- MPI: *`./configure -arch=LINUX; make`*

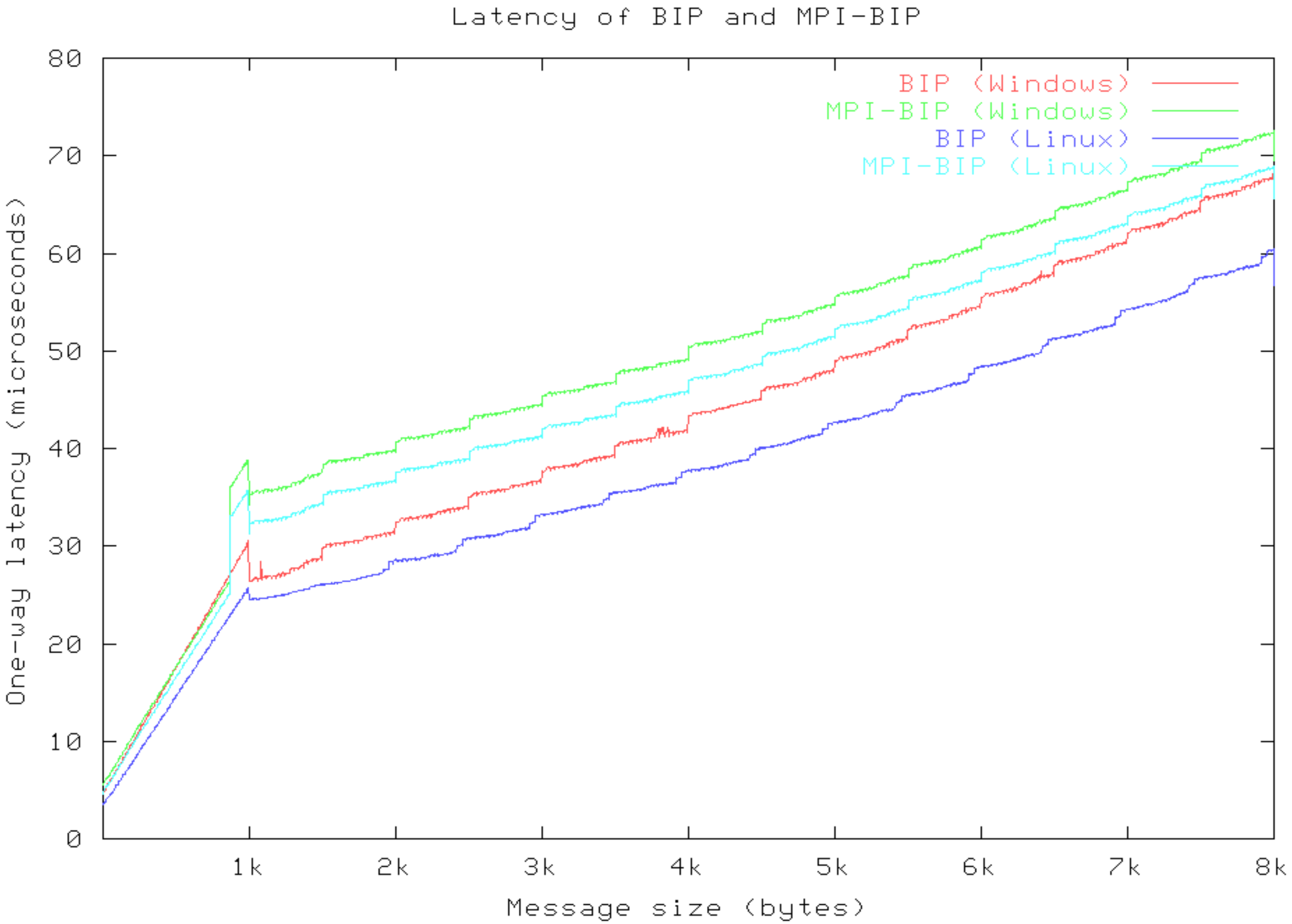
From Linux to Win2000: the runtime environment

- Perl scripts + ssh
- Route discovery, setting configuration, launching jobs
- Cygwin provides perl and ssh!
- Private key authentication (no password) only works with local accounts, NOT with domain controller
- Disk area shared with a linux samba server (no password authentication)

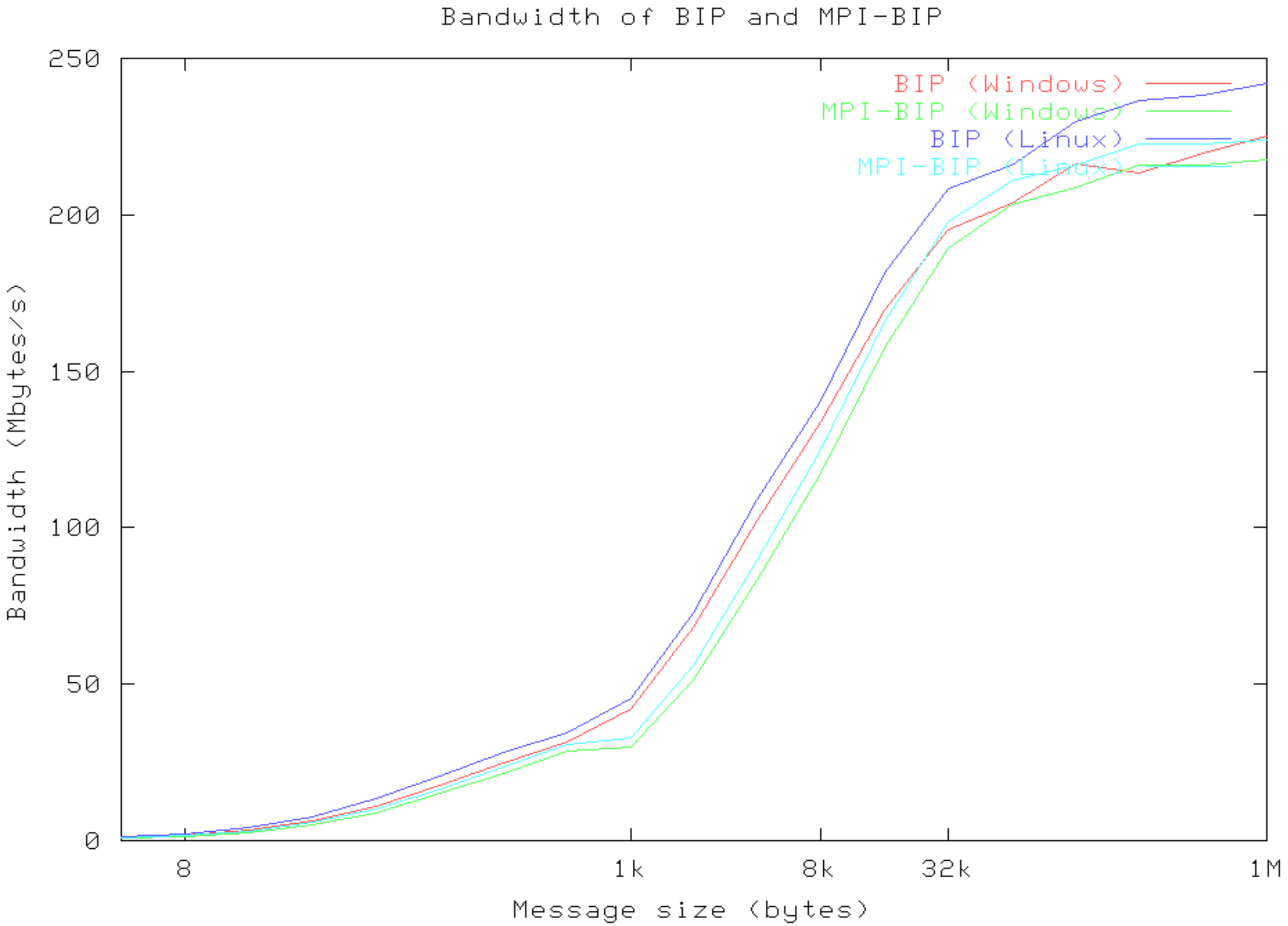
Managing the cluster

- We chose the brute-force method: replicate raw partition over the network (account management, software installation)
- SCSI disk performance: 30-50MB/s; use IP/Myrinet
- Broadcast the partition in a ring fashion (pipeline)
- Replicating the 5GB partition on 8 machines takes 2mn30s (~a local copy)
- Scale quite well up to 100 machines (theoretically)
- Performed under linux
- IP address managed by DHCP
- Hostname: edit the windows registry under linux

Perfomance: latency



Performance: bandwidth



Performance: NAS application benchmarks

	Sequential		4 processes		8 processes		16 processes	
	Win.	Linux	Win.	Linux	Win.	Linux	Win.	Linux
IS (class A)	9,47	9,46	2,52	2,52	1.53	1.46	1.31	1.27
			(3.6)	(3.8)	(6.2)	(6.5)	(7.2)	(7.4)
IS (class B)	38	38	10.70	10.31	6.05	5.94	5.34	5.22
			(3.6)	(3.7)	(6.3)	(6.2)	(7.1)	(7.3)
LU (class A)	1597	1230	398	309	201	156	196	138
			(4)	(4)	(7.9)	(7.9)	(8.1)	(8.9)
LU (class B)		5646	1647	1419	862	674	536	479

Conclusions

- Same level of performance under Windows and Linux,
- Using cygwin provides a UNIX-like environment for the runtime,
- Replicating the disk efficiently allows an easy management of the cluster,
- Clustering under windows is quite feasible even though Linux remains easier.