Validation of Dimemas communication model for MPI collective operations

Sergi Girona, Jesús Labarta, Rosa M. Badia

European Center for Parallelism of Barcelona

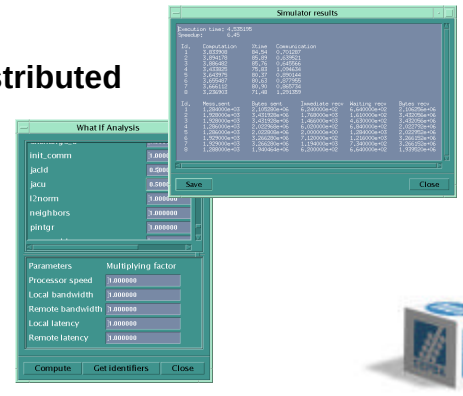
Departament d´Arquitectura de Computadors

Technical University of Catalonia

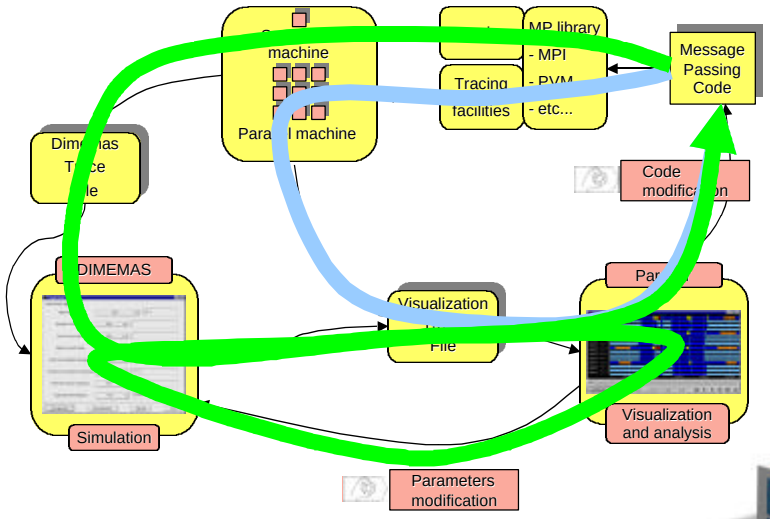
Barcelona, Spain

Dimemas

- Application performance analysis tool for message passing programs
- In development since 1992
- On a workstation
- Dimemas currently distributed by CEPBA

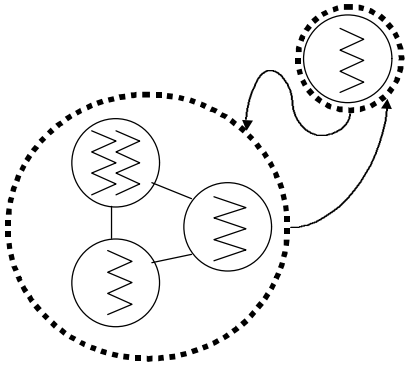


Tuning Methodology



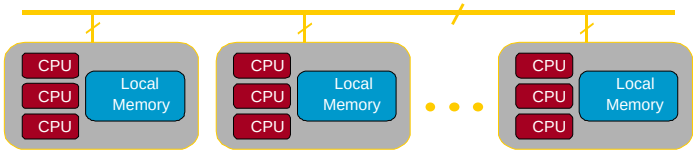
Tracefile

- **Characterizes application**
  - Sequence of resource demands for each task
  - Sequence of events: communication
- **Application model**



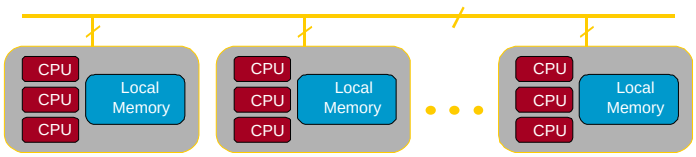
Simulated Architecture

- **"Abstract" architecture**
  - Simple/General
    - ✓Network of SMPs
  - Fast simulation
  - Key factors influencing performance
  - Abstract interconnect
    - ✓Local/remote latency/BW
    - ✓Injection mechanism (#links, half/full duplex)
    - ✓Bisection BW, contention



System

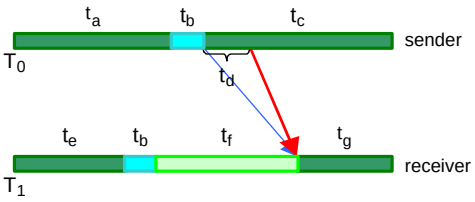
- **Process to processor mapping**
- **Multiprogramming**
  - Tasks sharing node
  - Different applications



Point to Point Communication

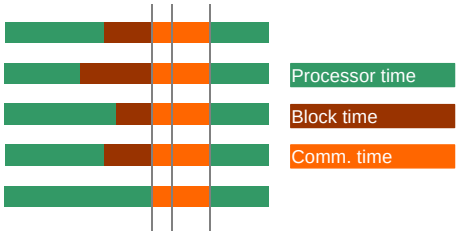
$$T = \text{Latency} + \frac{\text{Size}}{\text{Bandwidth}}$$

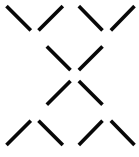
- **Latency**
- **Bandwidth**
- **Resource contention**



Collective Communication Model

- **Barrier**
- **Fan-in/fan-out phases**
  - Size of message
  - Null/Const/Lin/Log





Collective Communication Model

- **Communication time**

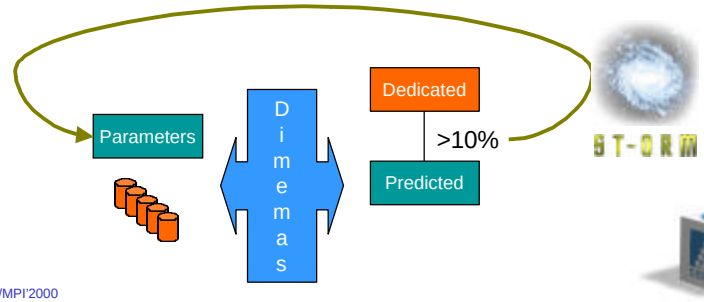
$$\text{Time} = \left(\text{Latency} + \frac{\text{Size}}{\text{Bandwidth}}\right) * \text{MODEL\_FACTOR}$$

- **Model factor**

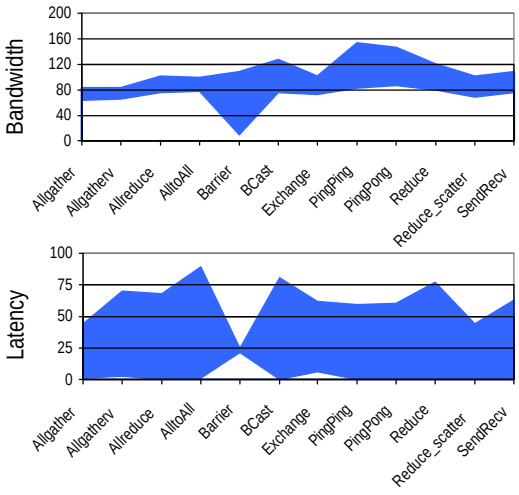
Model	Factor
Null	0
Constant	1
Linear	P
Logarithmic	$Nsteps = \sum_{i=1}^{\lceil \log_2 P \rceil} steps_i, steps_i = \left\lceil \frac{C}{B} \right\rceil$

Parameters Acquisition

- **Execution of PBM on SGI Origin**
  - Dedicated: execution time
  - Shared: traces for Dimemas
- **Compute latency, bandwidth, links, buses, phases, ...**
  - ST-ORM **http://www.cepba.upc.es/ST-ORM**
  - Objective: Predicted time with less than 10% error



System Characterization



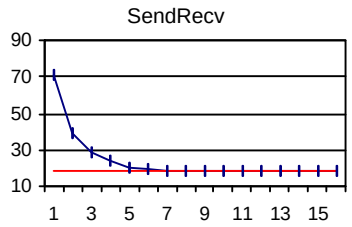
< 10% error regions

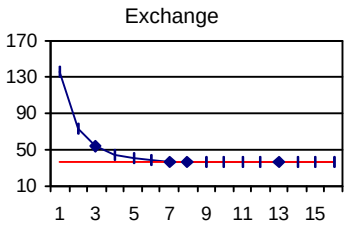
System Characterization

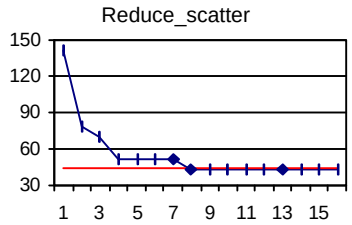
	IN		OUT	
Operation	**Model**	**Size**	**Model**	**Size**
Barrier	LIN	MAX	LIN	MAX
Bcast	LOG	MAX	NULL	
Gather	LOG	MEAN	NULL	
Gatherv	LOG	MEAN	NULL	
Scatter	NULL		LOG	MEAN
Scatterv	NULL		LOG	MEAN
Allgather	LOG	MEAN	LOG	MEAN
Allgatherv	LOG	MEAN	LOG	MEAN
Alltoall	LOG	MEAN	LOG	MAX
Alltoallv	LOG	MEAN	LOG	MAX
Reduce	LOG	2MAX	NULL	
Allreduce	LOG	2MAX	LOG	MAX
Reduce_Scatter	LOG	2MAX	LOG	MIN
Scan	LOG	MAX	LOG	MAX

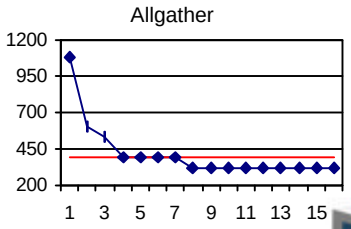
- **Latency = 25 μseconds**
- **Bandwidth = 87.5 MB/s**
- **1 HD link per node**

Influence of Buses



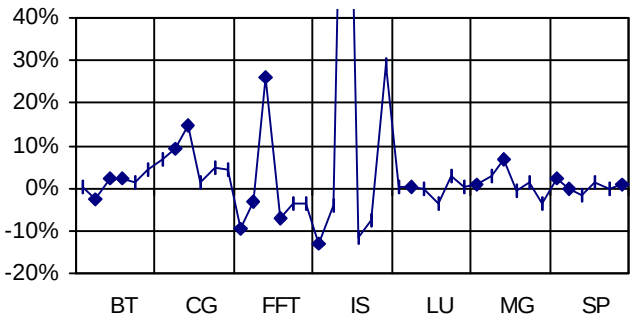






Validation

- **NAS benchmarks**
- **Classes W, A**
- **Size: 8/9, 16, 25/32**



Conclusions

- Simple but accurate formulation for collective communication
- Methodology for model validation
- Dimemas is a valid tool for performance analysis of message passing programs, parallel machines and message passing libraries
- Future: RMA and I/O operations pending for validation