Incremental Monitoring on Programmable Network Interface Cards

Laurent Lefèvre

INRIA / LIP (UMR CNRS, INRIA, ENS, UCB)

Laurent.lefevre@inria.fr

Dieter Kranzlmüller, Martin Maurer

GUP - Joh. Kepler Univ. Linz

Kranzlmueller@gup.jku.at

PDPTA04 - Las Vegas - June 2004

This research is partly supported by French “Programme d’Actions Intégrées Amadeus” funded by the French Ministery of Foreign Affairs and the Austrian Exchange Service (OAD), WTZ Program Amadeus under contract no. 13/2002

Monitoring …

… influences the observed program in

- Time
  - Events are delayed due to monitoring overhead
  - Ordering of events is perturbed
- Space
  - Storing monitoring data requires memory space

Monitoring optimizations

- Minimization of monitor overhead through minimal invasive instrumentation
- Minimization of monitor overhead through exploitation of additional hardware

Programmable network cards

- Myrinet NIC
  - Processor on board (Lanai)
  - Memory
  - Communications between host CPU and NIC:
    - Programmed Input/Output (PIO)
    - Direct memory access (DMA)
  - GM software
    - Software library
    - Kernel module
    - Myricom Control Program (MCP)

Monitoring on Myrinet NIC

- Architecture based on 3 steps :
  1. Preparation and instrumentation
  2. Initial record step
  3. Repeated replay phases
     - To increase amount of observation data
     - To perform program analysis

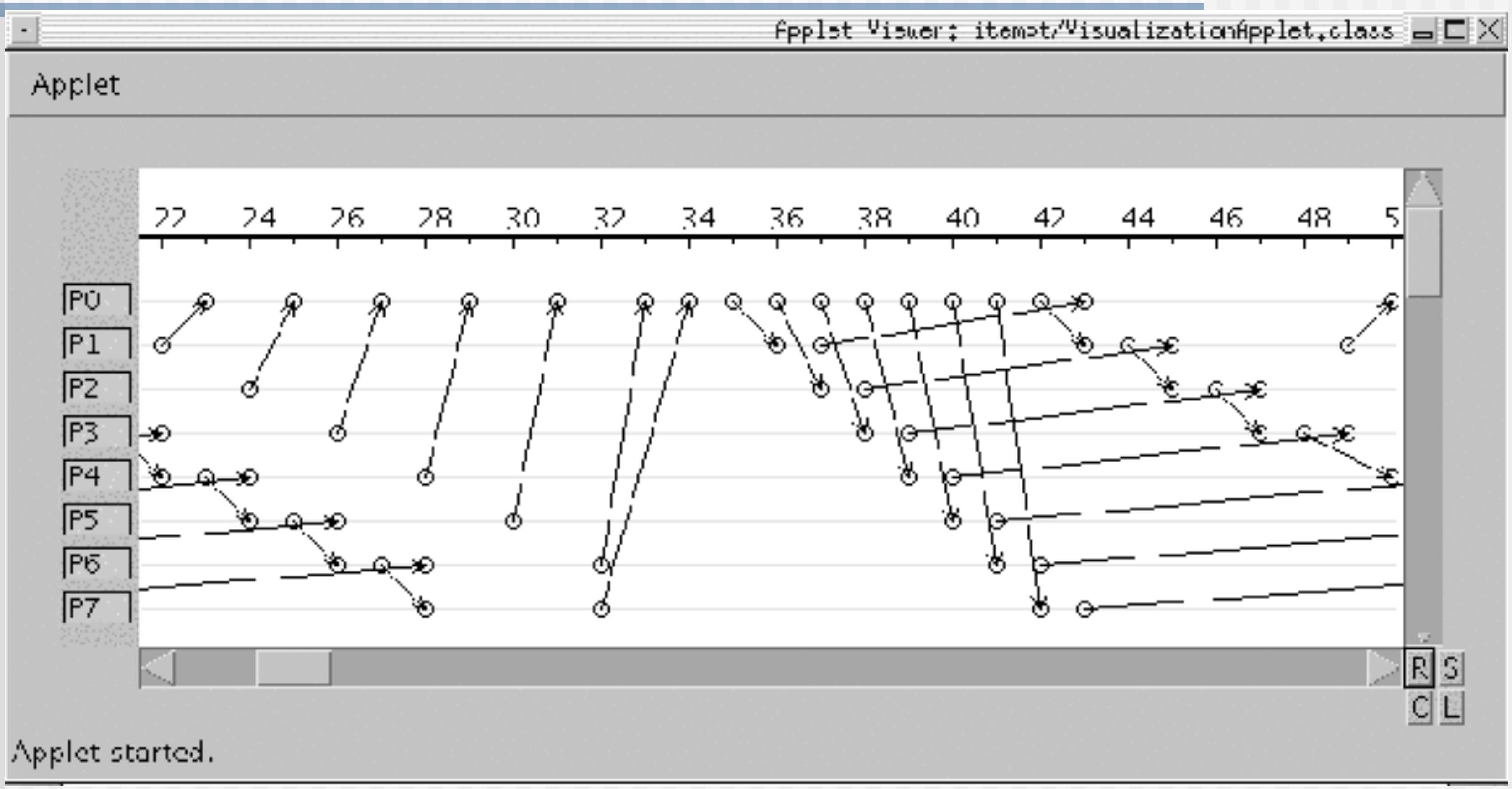
Preparation and instrumentation

- Loading modified MCP onto NIC
- Instrumentation of MPI program by including modified MPI header file
- Compiling application with modified MPICH library

Optimizing tasks on MCP

- Upon initialization of MPI program : memory reservation on NIC to store order of incoming messages
- If buffer full : transfer asynchronously to host memory during execution
- After execution : file generation of monitoring information extracted from NIC

Event graph



Applet started.

Debugging tool DeWiz screenshot with events collected on programmable card

Conclusion and current work

- Advantages:
  - Minimal intrusion of during initial record phase
  - Generation of arbitrary analysis data during follow-up replay phases
- Monitoring without user knowledge
- Adding QoS functionality on the NIC to filter monitoring actions