The new paradigm of Cluster Computing

Ferdinand Geier
ParTec AG
Company background

• 1995 Presentation of the ParaStation Hardware.
  – Development of the ParaStation Software.
• 1998 Presentation of ParaStation2 on top of MyriNet.
• 1999 Founding of ParTec AG.
  – Spin-Off from the Karlsruhe University.
• 2001 Presentation of the ParaStation3 software.
• 2003 April: Presentation of ParaStation 4.
ParaStation: Product overview

- **ParaStation 3**
  - based on Myrinet
  - User Level communication

- **ParaStation FE**
  - based on TCP/IP

- **New: ParaStation 4**
  - enhanced functionality
  - „hybrid communication“

- **Expertise: Cluster Competence Center**
  - Interconnects
  - Software
  - Benchmarks
ParaStation: Overview

ParaStation Graphical Management Interface

ParaStation Application Management

ParaStation Parallel Filesystem

ParaStation Batch System

ParaStation Console Management

ParaStation Server Provisioning

ParaStation Communications Stack

ParaStation Application

Kernel

Eth

HA (Myrinet)

Highspeed Application Data Switch

Mgmt Network Switch

ParaStation: Overview

ParaStation Graphical Management Interface

ParaStation Application Management

ParaStation Parallel Filesystem

ParaStation Batch System

ParaStation Console Management

ParaStation Server Provisioning

ParaStation Communications Stack

ParaStation Application

Kernel

Eth

HA (Myrinet)

Highspeed Application Data Switch

Mgmt Network Switch

ParaStation: Overview

ParaStation Graphical Management Interface

ParaStation Application Management

ParaStation Parallel Filesystem

ParaStation Batch System

ParaStation Console Management

ParaStation Server Provisioning

ParaStation Communications Stack

ParaStation Application

Kernel

Eth

HA (Myrinet)

Highspeed Application Data Switch

Mgmt Network Switch

ParaStation: Overview

ParaStation Graphical Management Interface

ParaStation Application Management

ParaStation Parallel Filesystem

ParaStation Batch System

ParaStation Console Management

ParaStation Server Provisioning

ParaStation Communications Stack

ParaStation Application

Kernel

Eth

HA (Myrinet)

Highspeed Application Data Switch

Mgmt Network Switch
ParaStation: Management

- Job launch:
  - can be launched on every node
  - efficient startup, no ssh/rsh
- Monitoring:
  - process / job
  - node
ParaStation: Management

- Jobs will be canceled if
  - a hardware problem (node crashed)
  - a software problem (process crashed)
  - process terminates

- All processes belonging to a job will be terminated!

- No orphaned processes left eating up CPU cycles!
ParaStation: Management

- Concept of virtual nodes
  - pool of physical nodes
  - mapped at startup
  - no node list
  - unavailable nodes are ignored
• Load balancing:
  – automatic
  – user defined (host list)

• Criterias for process placement:
  – current load
  – number of processes running
  – node list / host list
  – exclusive
  – other resources (disk, memory)
  – farming
ParaStation: Management

- Forwarding of input/output
- Forwarding of environment variables
- "implicit" user accounts: account not required on compute node, only on frontend
- Launching of jobs can be restricted to dedicated users
ParaStation: Graphical User IF

ParaStation Graphical Management Interface

- Monitoring of:
  - Activities
  - Availability
  - System parameters
- Based on SNMP
- Web-Frontend
ParaStation: Lowlevel Monitor

- Low level management with or without OS:
  - Hardware parameters
  - Console access
  - Based on IPMI
  - No additional KVM switch, terminal server, ...
ParaStation: Batchsystem

- Batchsystem:
  - Easy integration:
  - LSF, LSF Parallel
  - OpenPBS
  - PBS-Pro
  - Grid Engine
ParaStation: Installation

- (Semi) automated node:
  - Installation
  - Configuration
  - Update
ParaStation: Parallel Filesystem

- Easy installation
- Optimized performance
- ROM-I/O
- Based on PVFS
PVFS - Write Performance

Throughput [MBytes/sec] vs Number of nodes

- 4 ION / 8MB
- 16 ION / 32MB
- 32 ION / 64MB
PVFS - Read Performance

Throughput [MBytes/sec]

Number of nodes

4 ION / 8MB
16 ION / 32MB
32 ION / 64MB
- Supported networks
  - Ethernet (17µsec, 200MB/sec)
  - Myrinet (9µsec, 320MB/sec)
  - Infiniband (8µsec, 800MB/sec)
  - Shared memory (<1µsec, 2GB/sec)
  - (TCP/IP) (27µsec, 140MB/sec)
Kernel Communication

- Transparent to the application
- Multi hardware
- Effective local communication
- Protocol overhead
- Expensive switch from Userspace to Kernelspace
Userspace Communication

• Slim Protocol
• Direct hardware access
• Fast communication
• Multi hardware support harder to implement
• Security problem possible
ParaStation: Communication

- Communication Benchmark:
  - Pallas MPI Benchmark PMB2.2

- Test System:
  - Dual XEON System  2.6 GHz
  - 2 GB Memory
  - SuperMicro P4DPE-G2 (E7500)
  - Intel E1000 (82540) on board
  - Broadcom NetXtrem BCM5701
  - All numbers without switch
ParaStation: Communication

Vergleich Latenzzeit ParaStation 4/MPIch mit CH_p4

![Chart showing latency comparison between CH_p4 and ParaStation 4 for different packet sizes in bytes. The x-axis represents packet size in bytes starting from 0 to 1024, and the y-axis represents latency in micro seconds (µsec) ranging from 0 to 50. CH_p4 is shown in blue, and ParaStation 4 is shown in red.](chart.png)
ParaStation: Communication

Vergleich Durchsatz ParaStation 4/MPIch mit CH_p4

Durchsatz [MBytes/sec]

Paketgröße [Bytes]

CH_p4
ParaStation 4
ParaStation: Summary

- Modular, high speed, robust, easy to use compute cluster environment
- Linux
  - all major distributions
  - all kernel versions
- IA32, (IA64 soon)
- Fully supported (support@par-tec.com)
Contact

http://www.par-tec.com

info@par-tec.com