SuSE Linux Cluster Solutions: Experience with the Asgard Cluster at ETH Zürich

Matthias Troyer
Department of Physics
ETH Zürich

Overview

- Motivation: why a Beowulf cluster?
- Experiences with WTO procurement
- The "Asgard" Beowulf cluster at ETH Zürich
  - Hardware: 192 Dual-CPU Pentium III CPUs, 1 Gigabyte RAM each
  - Operating system and software: SuSE Linux
- User projects
- User experiences

Motivation / situation in fall 1998

- Increased need for simulations in physics research
  - Strongly correlated electron systems in condensed matter physics
  - Monte Carlo simulations in particle physics
  - PDE solvers in all areas of physics
- Downscaling of centralized computing resources at ETHZ
  - Intel Paragon decommissioned, no replacement planned
  - Upgrade of vector supercomputers uncertain
  - No plans for new centralized supercomputing resources
- Solution: Beowulf cluster for Department of Physics

Why a Beowulf cluster?

- Existing simulation codes perfectly parallelizable
  - Monte Carlo simulations
  - Series expansions
  - Other simulations that need to be repeated many times for different parameters
- Education in "Computational Sciences and Engineering"
- Development of new parallel simulation codes
- We need a machine with
  - Large computing power
  - Small to moderate communication needs
- The ideal solution is a Beowulf cluster

History

- Winter 1998/1999: Feasibility study for a Beowulf cluster for the Department of Physics at ETH Zürich
- March 1999: The Department of Physics submits a proposal to the ETHZ executive board to start a competition for the procurement of a Beowulf cluster.
- May 12, 1999: Start of the WTO government procurement procedures for a Beowulf cluster.
- July 8, 1999: Deadline for submitting a tender
- Summer 1999: Evaluation period. The offered systems vary considerably as was to be expected.
- Summer 1999: The Department of Mathematics joins the project

History (continued)

- Sept. 13-17, 1999: Negotiations with all suppliers, with fixed technical specifications.
- October 14, 1999: Decision to buy the system offered by DALCO electronics is published
- November 8, 1999: Contract with DALCO electronics.
- Dec. 20-23, 1999: Delivery of the complete system to ETHZ
- January - March, 2000: Acceptance tests.
- March 21, 2000: All accpetance tests are fulfilled.

WTO procurement procedure

- Necessary for all purchases of more than about 250'000 CHF
- Experiences
  - Requirement to have service and support in Switzerland eliminated all overseas companies from the competition
    - 23 companies requested documentation
    - Only four tenders submitted
  - Call for tender
    - Memory, network and file server capacity was specified
    - CPU type and number of nodes not specified, but approximately 50 Gigaflop sustained performance requested
    - Benchmarks with user programs to be reported in tender. Greatly varying tenders required additional negotiations for a system with fixed specifications
  - Complex rules: legal advice essential
  - Competition allowed to get lower prices
  - Added a 6-month delay
    - Still possible to get up-to-date system
    - Delays led to sharp rise in RAM prices and finally 35% higher costs

Alpha vs. Pentium

- Benchmarks with user programs relevant for decision
  - Loopsim: Quantum Monte Carlo simulation in condensed matter physics
    - Mostly integer arithmetic
  - Transfer: Transfer matrix simulation in statistical physics
    - Double precision floating point and integer operations
  - Trajek: Monte Carlo simulation in particle physics
    - Single precision floating point and integer operations

	wallclock time [sec]			Price/performance normalized to Pentium systems			geom.
	loopsim	transfer	trajek	loopsim	transfer	trajek	mean
667 MHz Alpha EV6	1791	36	1569	2.43	1.31	1.83	1.80
500 MHz Pentium III	**2898**	**108**	**3363**	**1.00**	**1.00**	**1.00**	**1.00**
333 MHz UltraSparc	3618	170	6318				

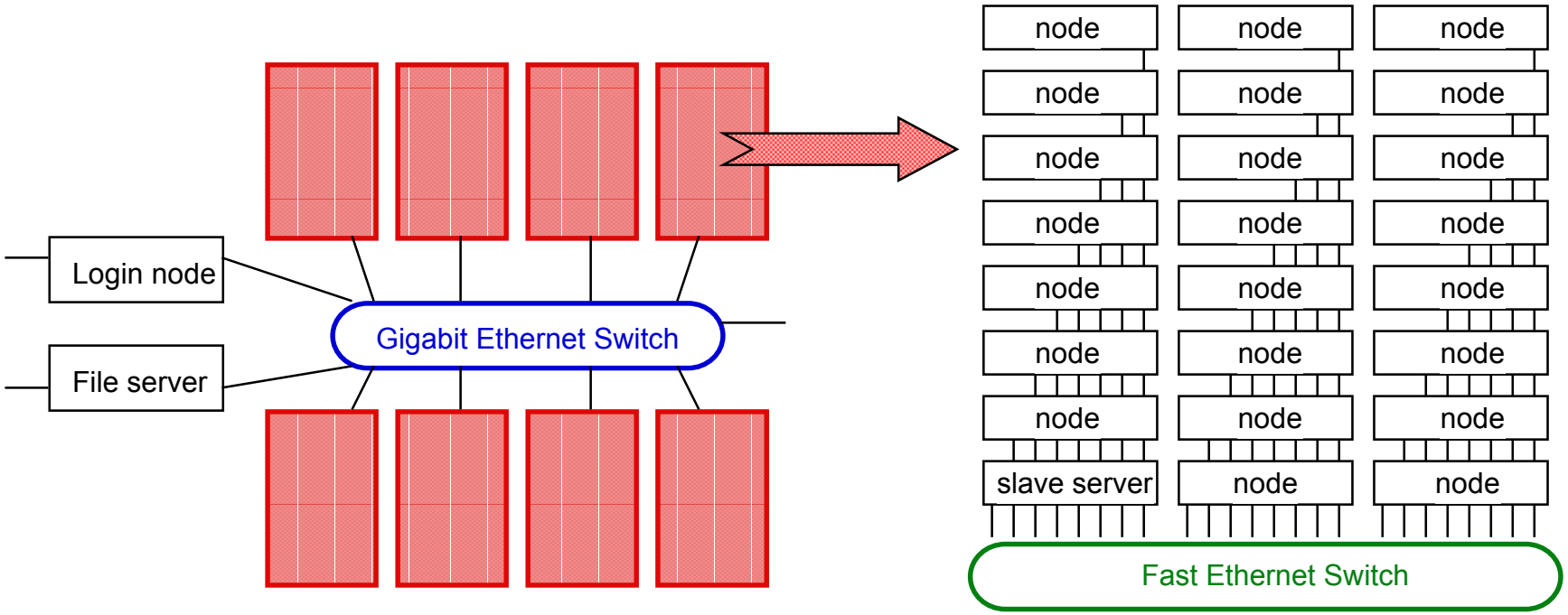
- Conclusions
  - Alpha systems consistently faster
  - Pentium system offers much better performance at same price
  - Linpack-performance irrelevant for our applications

The Asgard cluster

- 192 compute nodes
  - Dual-CPU 500 MHz Pentium III
  - Intel N440BX motherboards
  - 1 Gigabyte RAM
  - 6 Gigabyte IDE hard disk
- Network
  - 24 nodes connected by Fast Ethernet to a switch
  - 8 Fast Ethernet switches connected by Gigabit Ethernet to Gigabit switch
- Software
  - SuSE Linux 6.3 + cluster management software by SuSE
  - Complete software installation done by Anas Nashif of SuSE

Schematic layout

- Gigabit Ethernat backbone with
  - Eight "frames" with compute nodes
  - Login node
  - Connection to "meta-cluster"
  - File server with
    - User home directories
    - User software
- Compute frames
  - 24 dual-CPU nodes
  - Connected to Fast Ethernet switch
  - One node is "slave server"
    - /usr and root directories
    - Boot server

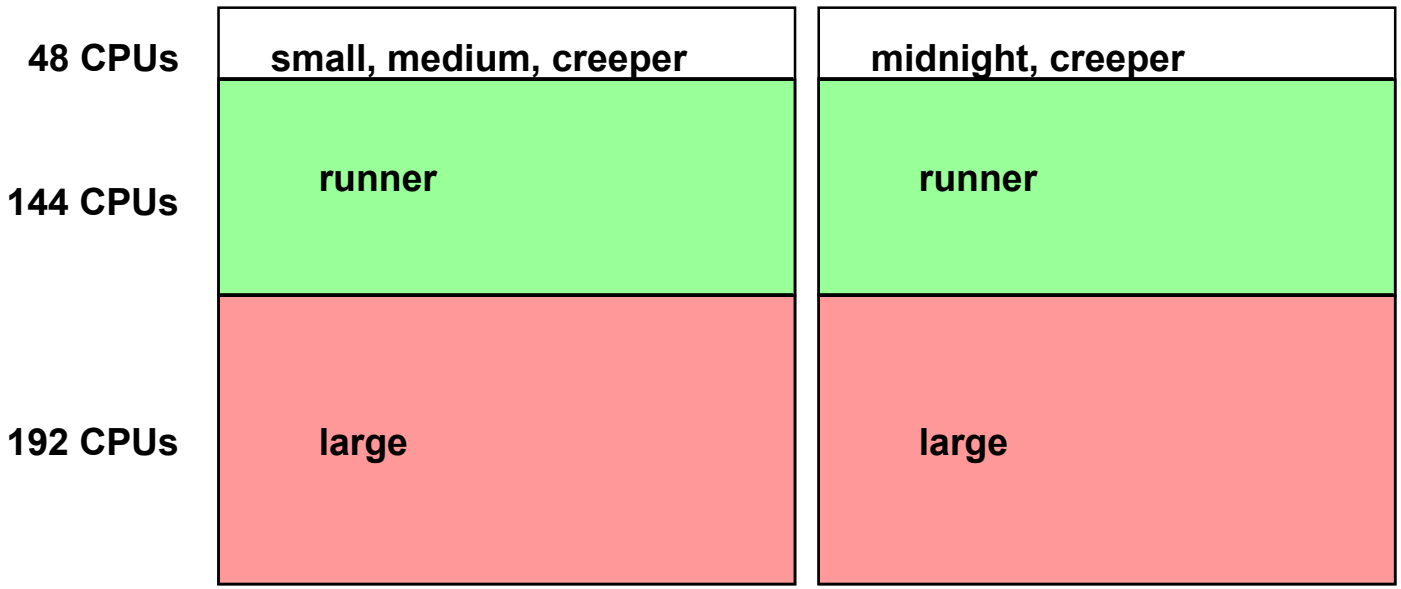


Hardware installation by Dalco electronics

The complete system

The queue structure

- **Small:** up to 8 nodes (16 CPUs), 30 minutes for testing and debugging
- **Medium:** up to 24 nodes (48 CPUs), 60 minutes for testing and debugging
- **Large:** up to 96 nodes (192 CPUs), 8 hours for parallel jobs
- **Runner:** up to 1 node (2 CPUs), 24 hours for single-node jobs
- Midnight: up to 24 nodes (48 CPUs), 4 hours for short jobs at night
- **Creeper:** up to 1 node (2 CPUs), 99 hours low priority jobs using idle time



Weekends: possible to use whole machine: 192 nodes, 384 CPUs

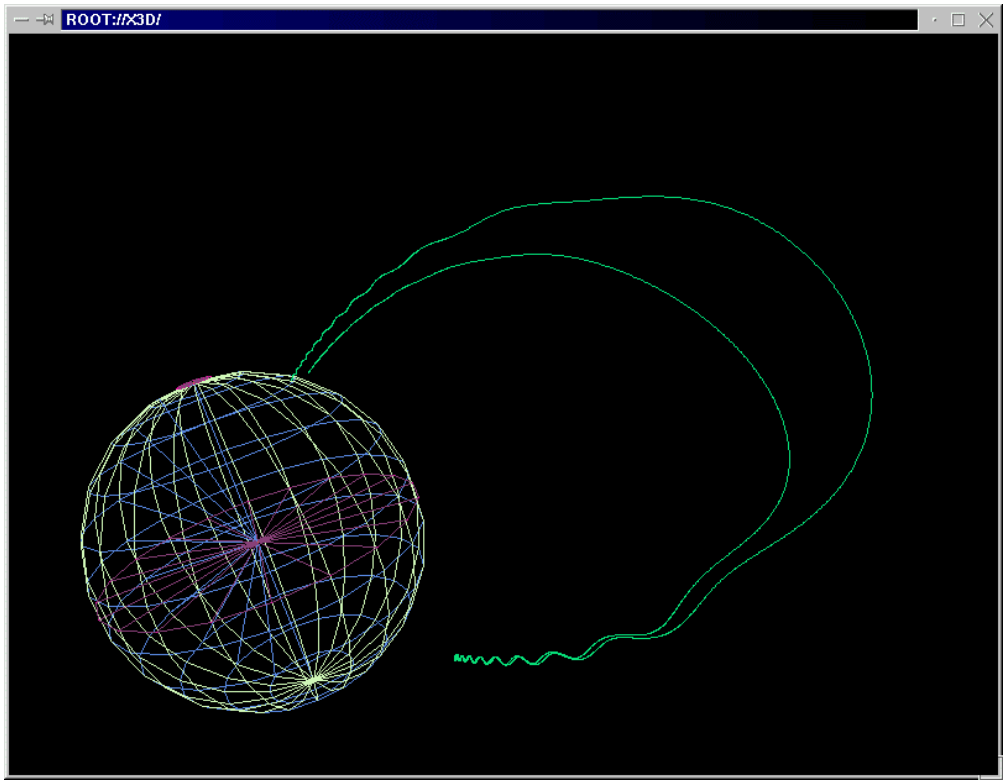
User projects

- Theoretical physics
  - Quantum phase transitions
  - Nonlinear Schrödinger equations
  - Vortex matter
- Particle physics
  - Detector simulations for AMS and CMS large scale experiments
  - Simulation of high intensity particle beams
- Astrophysics
  - Simulation of relativistic plasma plasma waves
- Education
  - Course in "parallel computing"
  - Semester, diploma and doctoral theses in physics, mathematics and computational sciences

Simulations for the AMS experiment

- "Alpha magnetic spectrometer" measures charged cosmic radiation
- Will be installed on the International Space Station
- Prototype was flown as main-experiment on board of space-shuttle 'Discovery' on mission STS91 (June 1998)
- Simulations of tracks of charged particles in earth magnetic field

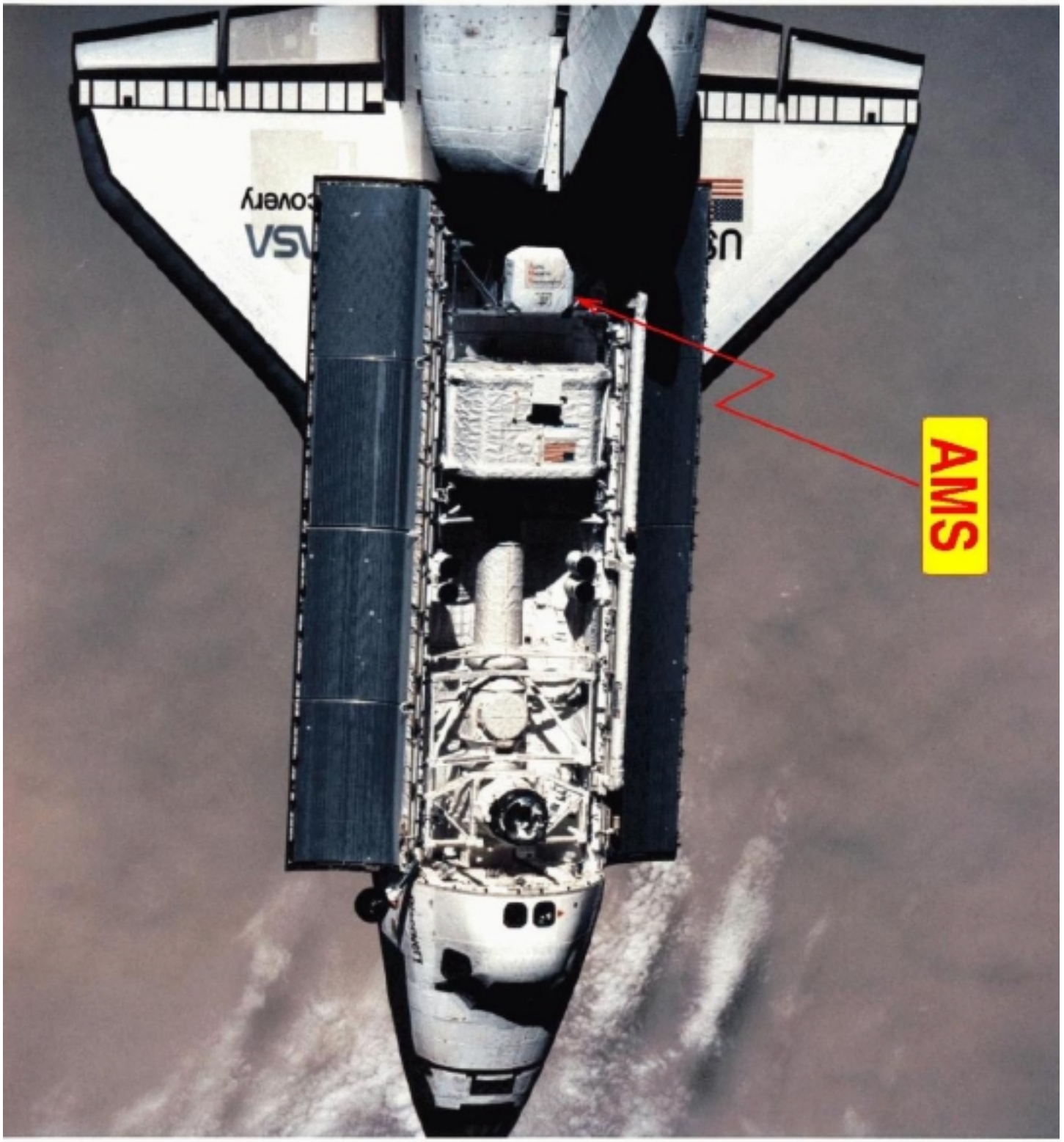
  essential for evaluation of measurements



National Aeronautics and
Space Administration

NASA7-726-068

Lyndon B. Johnson Space Center
Houston, Texas 77058

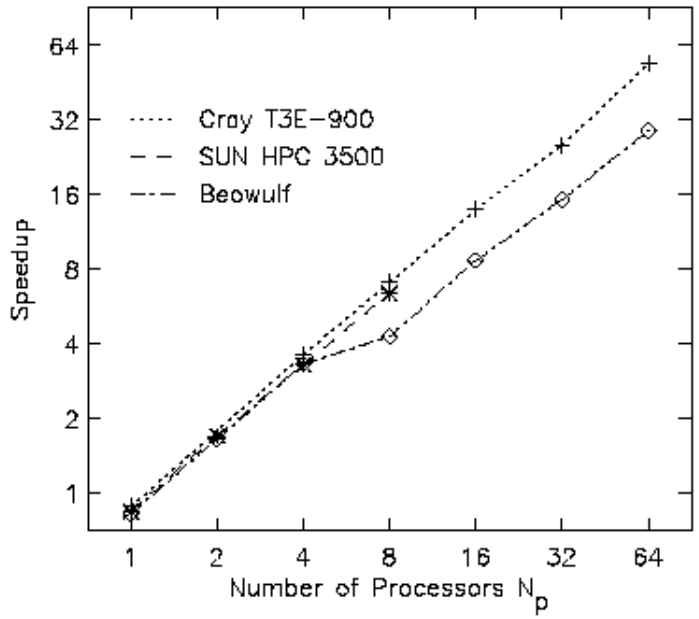


y9823

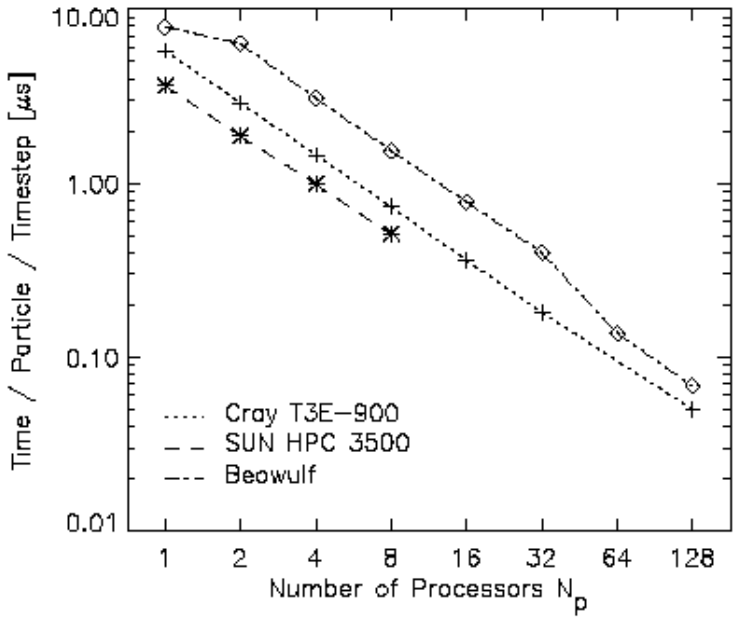
Benchmark of relativistic PIC code

- par-T: A Parallel Relativistic Fully 3D Electromagnetic Particle-in-Cell Code
- P. Messmer, Institute of Astronomy

- Fixed problem size



- Problem size scales with N_p



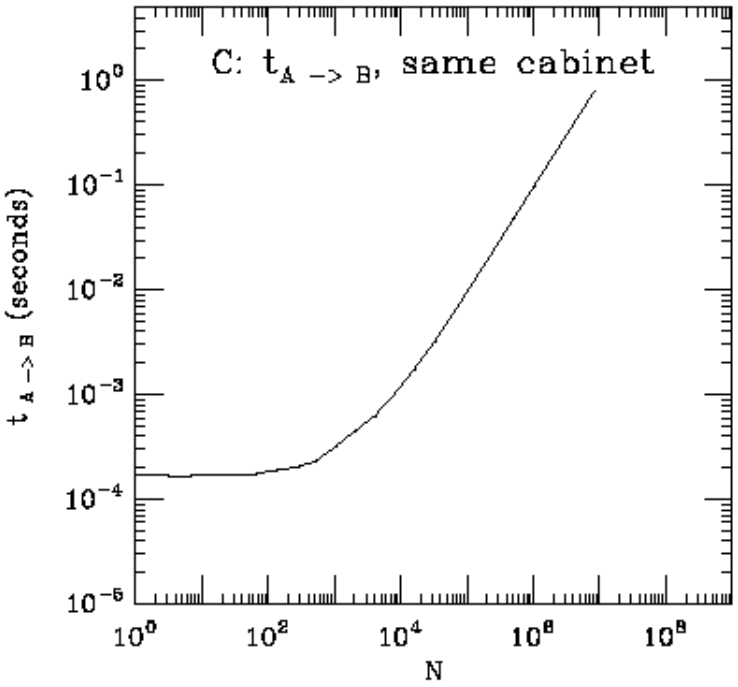
- Scales well on Beowulf architecture

Tests of communication bandwidth and latency

- W.P. Petersen and A. Friedli
- Test performed in the course "parallel programming"

- Bandwidth and latency estimated from timing of ping-pong test for various message sizes using MPICH

Language	Latency	Bandwidth
Fortran	500μS	9 MB/s
C	208μS	1025 MBs/



Experiences in the first four months

- Hardware problems
  - Replacements of 3 defective nodes in the first six weeks
  - 26 unexpected reboots of single nodes necessary in first four months; reasons unknown.

- Very positive user experience
  - Performance and reliability exceed expectations
  - Fast Ethernet completely sufficient for our current applications
  - Ideal system for development and optimization of parallel programs
  - Ideal for teaching - no problems despite 25 students on the system

- Bottleneck is disk-I/O at file server
  - Extremely slow when many nodes read/write at the same time
  - Separate file servers for user home directories and simulation data preferable
  - can a parallel file system help?

Summary

- Beowulf architecture exceeded our expectations
  - Pentium CPUs had better price/performance for our applications
  - 100 Mbit Ethernet sufficient for more problems than expected
  - Ideal system for our simulations
- No problems with acceptance tests
- Good experiences with software installation by SuSE
- Disk-I/O at central file server is main bottleneck