

# Experiences with HPC on Windows

Christian Terboven

terboven@rz.rwth-aachen.de

Center for Computing and Communication  
RWTH Aachen University

# Agenda

- High Performance Computing (HPC)
  - OpenMP & MPI & Hybrid Programming
- Windows-Cluster @ Aachen
  - Hardware & Software
  - Deployment & Configuration
- Top500 submission
- Case Studies
  - Dynamic Optimization: AVT
  - Bevel Gears: WZL
  - Application & Benchmark Codes
- Summary

2

# High Performance Computing (HPC)

- HPC begins ... when performance (runtime) matters!
  - HPC is about reducing latency - the „Grid“ increases latency
  - Dominating programming languages: Fortran (77 + 95), C++, C
- Focus in Aachen is on Computational Engineering Science!
- We do HPC on Unix for many years - why try Windows?
  - Attract new HPC users:
    - Third party cooperations sometimes depend on Windows
    - Some users look out for the Desktop-like HPC experience
    - Windows is a great development platform
  - We rely on tools: Combine the best of both worlds!
  - Top500 on Windows: Why not 😊?

3

Center for

Computing and

Communication

HPC  
OverviewWindows  
Cluster

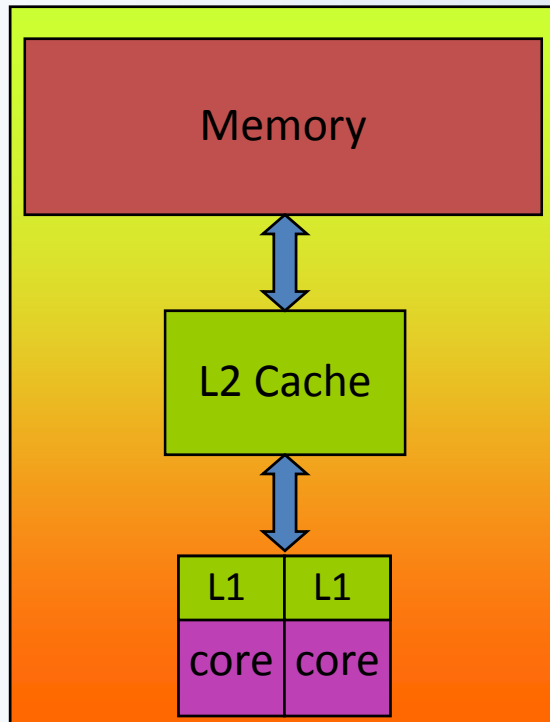
Top500

Case Studies

Summary

# Parallel Computer Architectures: Shared-Memory

- Dual-Core processors are common now
  - A laptop/desktop is a Shared-Memory parallel computer!



- Multiple processors have access to the same main memory.
- Different types:
  - Uniform memory access (UMA)
  - Non-uniform memory access (ccNUMA), still cache-coherent
- Trend towards ccNUMA

*Intel Woodcrest-based system*

*I am ignoring instruction caches, address caches (TLB), write buffers, prefetch buffers, ... as data caches are most important for HPC applications.*

4

Center for

Computing and

Communication

HPC  
Overview

Windows  
Cluster

Top500

Case Studies

Summary

## Programming for Shared-Memory: OpenMP

The logo for OpenMP, featuring the text "OpenMP" in a large, bold, teal font with a horizontal line above and below it. A small "TM" trademark symbol is located to the right of the "P".[www.openmp.org](http://www.openmp.org)

```
const int N = 100000;
double a[N], b[N], c[N];
[...]
#pragma omp parallel for
for (int i = 0; i < N; i++) {
    c[i] = a[i] + b[i];
}
[...]
```

Master Thread

Serial Part

Parallel Region

Slave  
Threads

Serial Part

5

○ ... and other threading-based paradigms ...

enter for

Computing and

Communication

HPC  
OverviewWindows  
Cluster

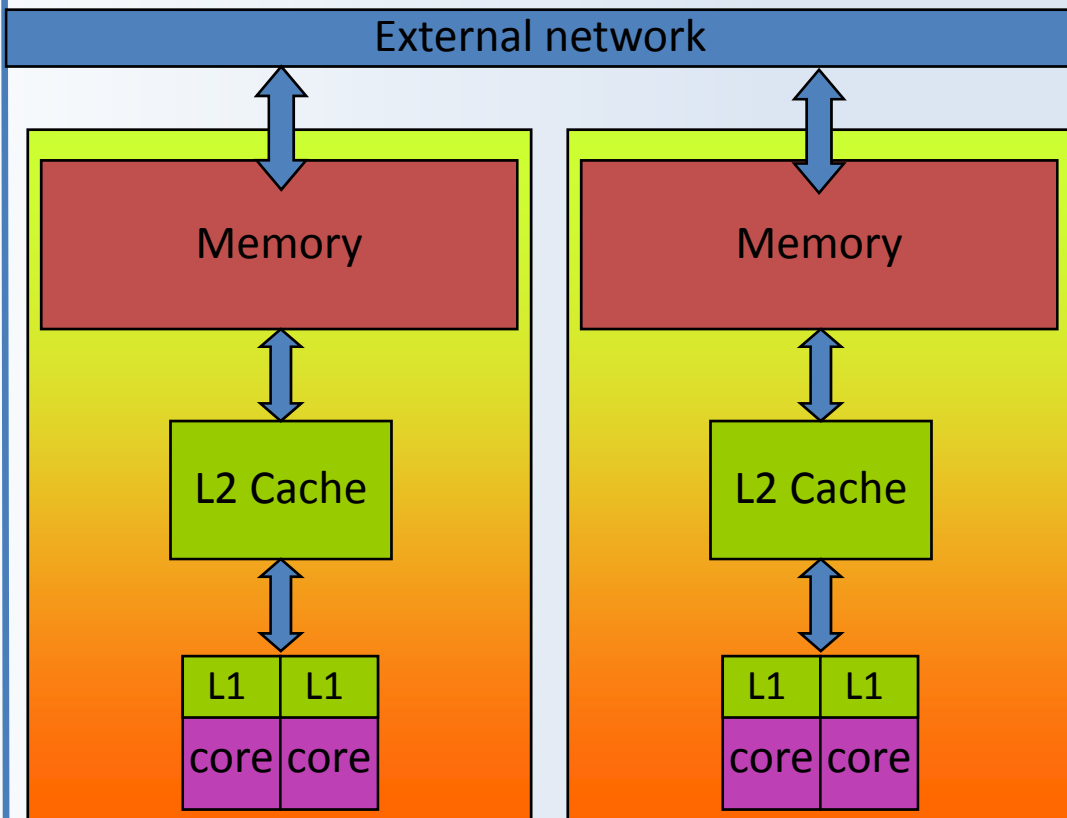
Top500

Case Studies

Summary

## Parallel Computer Architectures: Distributed-Memory

- Distributed-Memory: Each processor has only access to it's own main memory
- Programs have to use external network for communication



- Cooperation is done via message exchange  
→ Cluster

6

Center for

Computing and  
CommunicationHPC  
OverviewWindows  
Cluster

Top500

Case Studies

Summary

# Programming for Distributed-Memory: MPI



[www.mpi-forum.org](http://www.mpi-forum.org)

```
int rank, size, value = 23;
MPI_Init(...);
MPI_Comm_size(MPI_COMM_WORLD, &size);
MPI_Comm_rank(MPI_COMM_WORLD, &rank);
if (rank == 0) {
    MPI_Send(&value, 1, MPI_INT, rank+1, ...)
}
else {
    MPI_Recv(&value, 1, MPI_INT, rank-1, ...)
}
MPI_Finalize(...);
```

MPI Process 1

MPI Process 2

rank: 0

rank: 1

MPI\_Send

Msg: 1 int

MPI\_Recv

7

enter for

Computing and  
Communication

HPC  
Overview

Windows  
Cluster

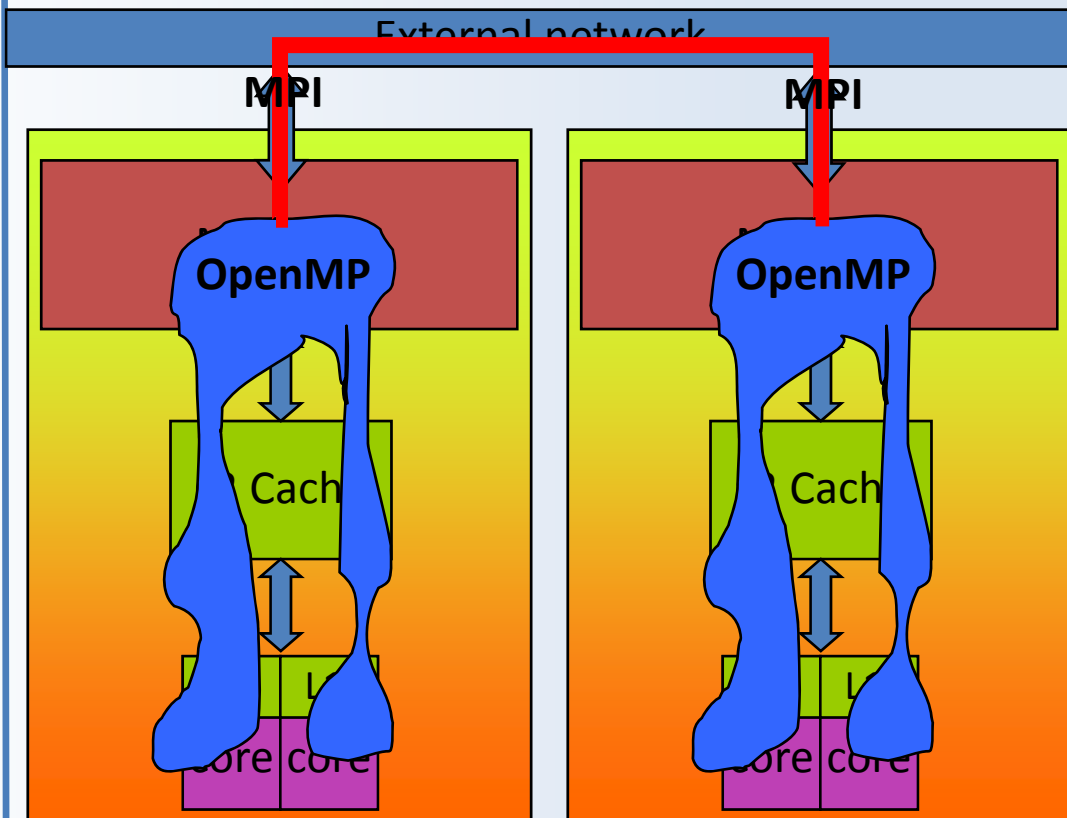
Top500

Case Studies

Summary

# Programming Clusters of SMP Nodes: Hybrid

- Multiple levels of parallelism can improve scalability!
  - One or more MPI tasks per node
  - One or more OpenMP threads per MPI task



- Many of our applications are hybrid: Best fit for clusters of SMP nodes.
- The SMP node will grow in the future!

8

Center for

Computing and  
CommunicationHPC  
OverviewWindows  
Cluster

Top500

Case Studies

Summary

## Speedup and Efficiency

- The *Speedup* denotes how much a parallel program is faster than the serial program:

$$S_p = \frac{T_1}{T_p}$$

$p$  = number of threads / processes  
 $T_1$  = execution time of the serial program  
 $T_p$  = execution time of the parallel program

- The *Efficiency* of a parallel program is defined as:

$$E_p = \frac{S_p}{p}$$

- According to Amdahl's law the Speedup is limited to  $p$ .

*Note: Other (similar) definitions for other needs available in the literature.*

# Agenda

- High Performance Computing (HPC)
  - OpenMP & MPI & Hybrid Programming
- Windows-Cluster @ Aachen
  - Hardware & Software
  - Deployment & Configuration
- Top500 submission
- Case Studies
  - Dynamic Optimization: AVT
  - Bevel Gears: WZL
  - Application & Benchmark Codes
- Summary

10

Center for

Computing and

Communication

HPC  
OverviewWindows  
Cluster

Top500

Case Studies

Summary

# Harpertown-based InfiniBand Cluster

- Recently installed Cluster:
  - Fujitsu-Siemens Primergy RS 200 S4 servers
    - 2x Intel Xeon 5450 (quad-core, 3.0 GHz)
    - 16 / 32 GB memory per node
    - 4x DDR InfiniBand (latency: 2.7 us, bandwidth: 1250 MB/s)
  - In total: About 25 TFLOP/s peak performance (260 nodes)
- Cluster Frontend
  - Windows Server 2008
  - Better performance when machine gets loaded
  - Faster file access: Lower access time, higher bandwidth
  - Automatic load balancing: Three machines are clustered to form „one“ interactive machine:  
`cluster-win.rz.rwth-aachen.de`

11

Center for

Computing and

Communication

HPC  
OverviewWindows  
Cluster

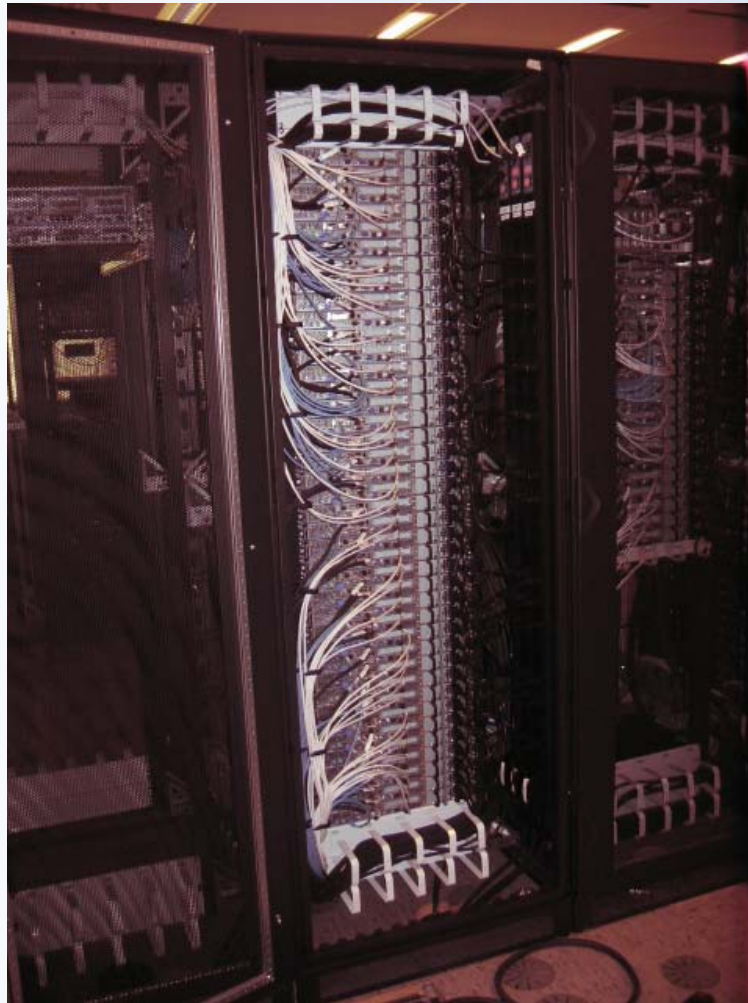
Top500

Case Studies

Summary

# Harpertown-based InfiniBand Cluster

- Quite a lot of Pizza boxes... 4 cables per box:



- InfiniBand: MPI
- Gigabit Ethernet: I/O
- Management
- Power



*Pictures taken during Cluster installation*

12

Center for

Computing and  
Communication

HPC  
Overview

Windows  
Cluster

Top500

Case Studies

Summary

# Software Environment

- Complete Development Environment
  - Visual Studio 2005 Pro + Microsoft Compute Cluster Pack
  - Subversion Client, X-Win32, Microsoft SDKs, ...
  - Intel Tool Suite:
    - C/C++ Compiler, Fortran Compiler
    - VTune Performance Analyzer, Threading Tools
    - MKL library, Threading Building Blocks
    - Intel MPI + Intel Tracing Tools
  - Java, Eclipse, ...
- Growing list of ISV software
  - ANSYS, HyperWorks, Fluent, Maple, Mathematica, Matlab, MS Office 2003, MSC.Marc, MSC.Adams, ...
  - User-licensed software hosting, e.g. GTM-X

13

# ISV codes in the batch system

- Several requirements
  - No usage of graphical elements allowed
  - No user input allowed (except redirected standard input)
  - Execution of compiled set of commands, Termination

- Exemplary usage instructions for Matlab

Software share

Command line:

```
\\cifs\cluster\Software\MATLAB\bin\win64\matlab.exe  
/minimize /nosplash  
/logfile log.txt  
/r "cd('\\cifs\cluster\home\YOUR_USERID\YOUR_PATH' ) ,  
YOUR_M_FILE,"
```

Disable GUI elements

Save output

Change dir  
& Execute

- The .M file should contain „quit;“ as last statement

14

enter for

Computing and  
Communication

HPC  
Overview

Windows  
Cluster

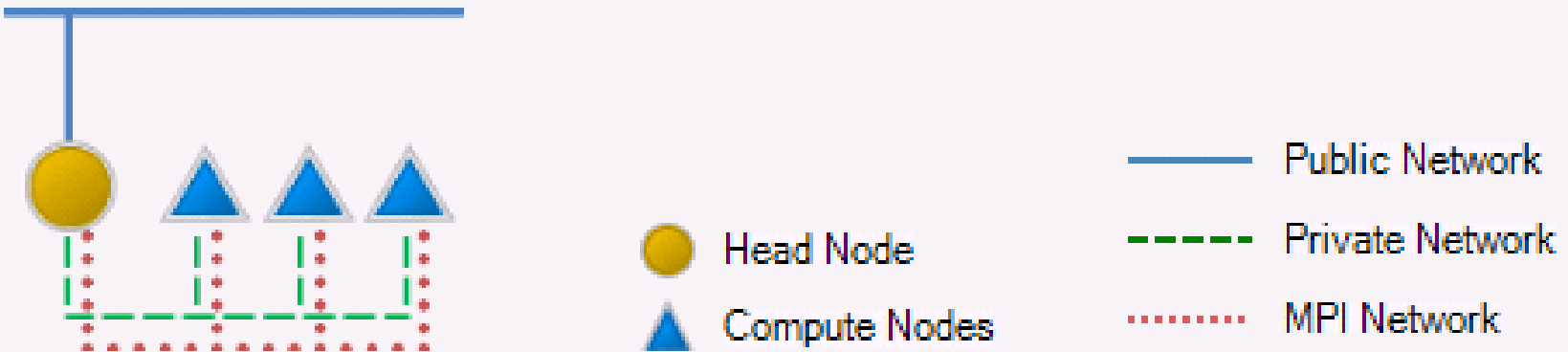
Top500

Case Studies

Summary

# Deploying a 256 node cluster

## Topology No. 3



- Setting up the Head Node: approx. 2 hours
  - Installing and configuring Microsoft SQL Server 2005
  - Installing and configuring Microsoft HPC Pack
- Installing the compute nodes:
  - Installing 1 compute node: 50 minutes
  - Installing n compute nodes: 50 minutes? (Multicasting!)
    - We installed 42 nodes at once in 50 minutes

15

Center for

Computing and

Communication

HPC  
OverviewWindows  
Cluster

Top500

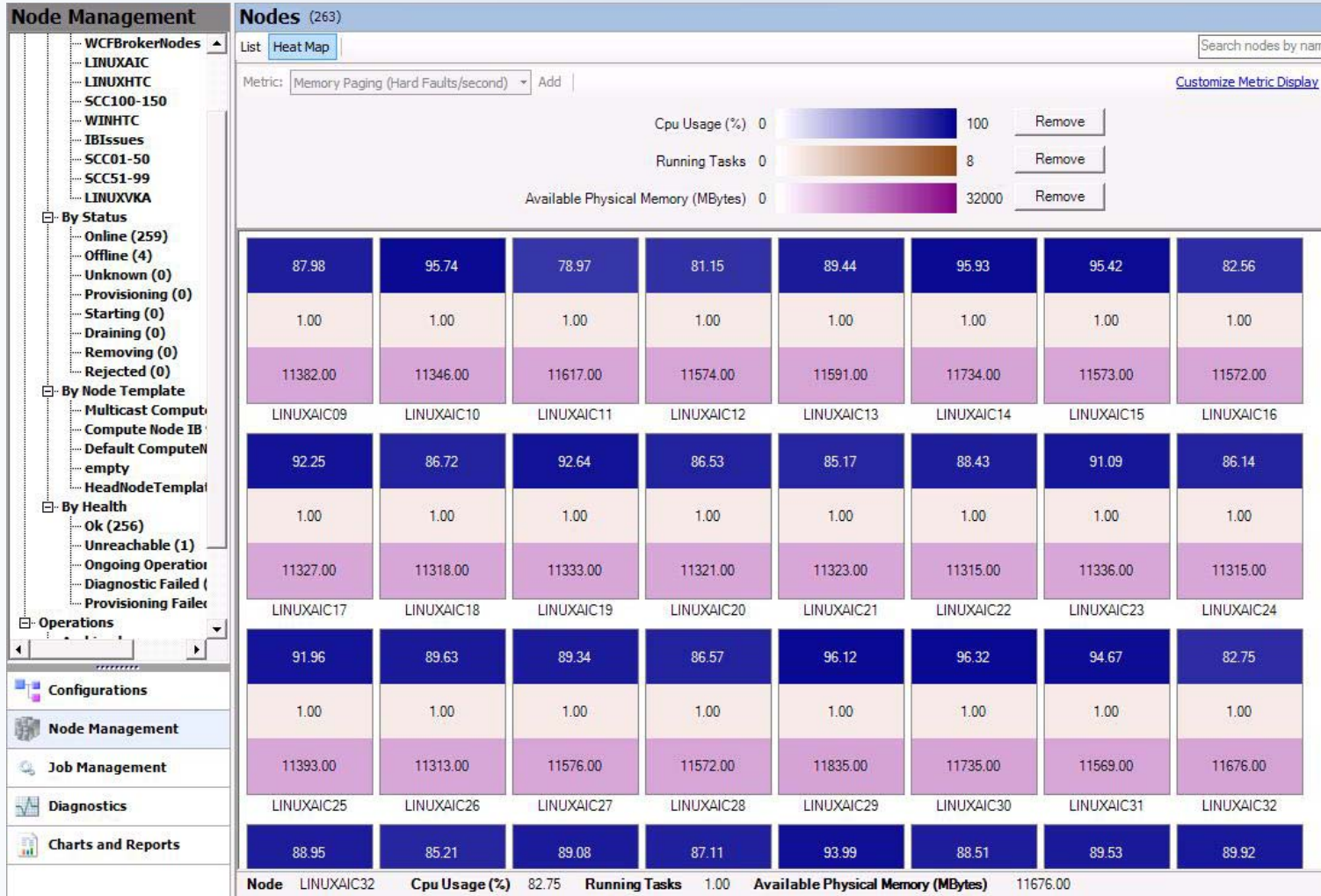
Case Studies

Summary



# Configuring a 256 node cluster

- Updates? Deploy an updated image to the compute nodes!



17

# Agenda

- High Performance Computing (HPC)
  - OpenMP & MPI & Hybrid Programming
- Windows-Cluster @ Aachen
  - Hardware & Software
  - Deployment & Configuration
- Top500 submission
- Case Studies
  - Dynamic Optimization: AVT
  - Bevel Gears: WZL
  - Application & Benchmark Codes
- Summary

18

Center for

Computing and

Communication

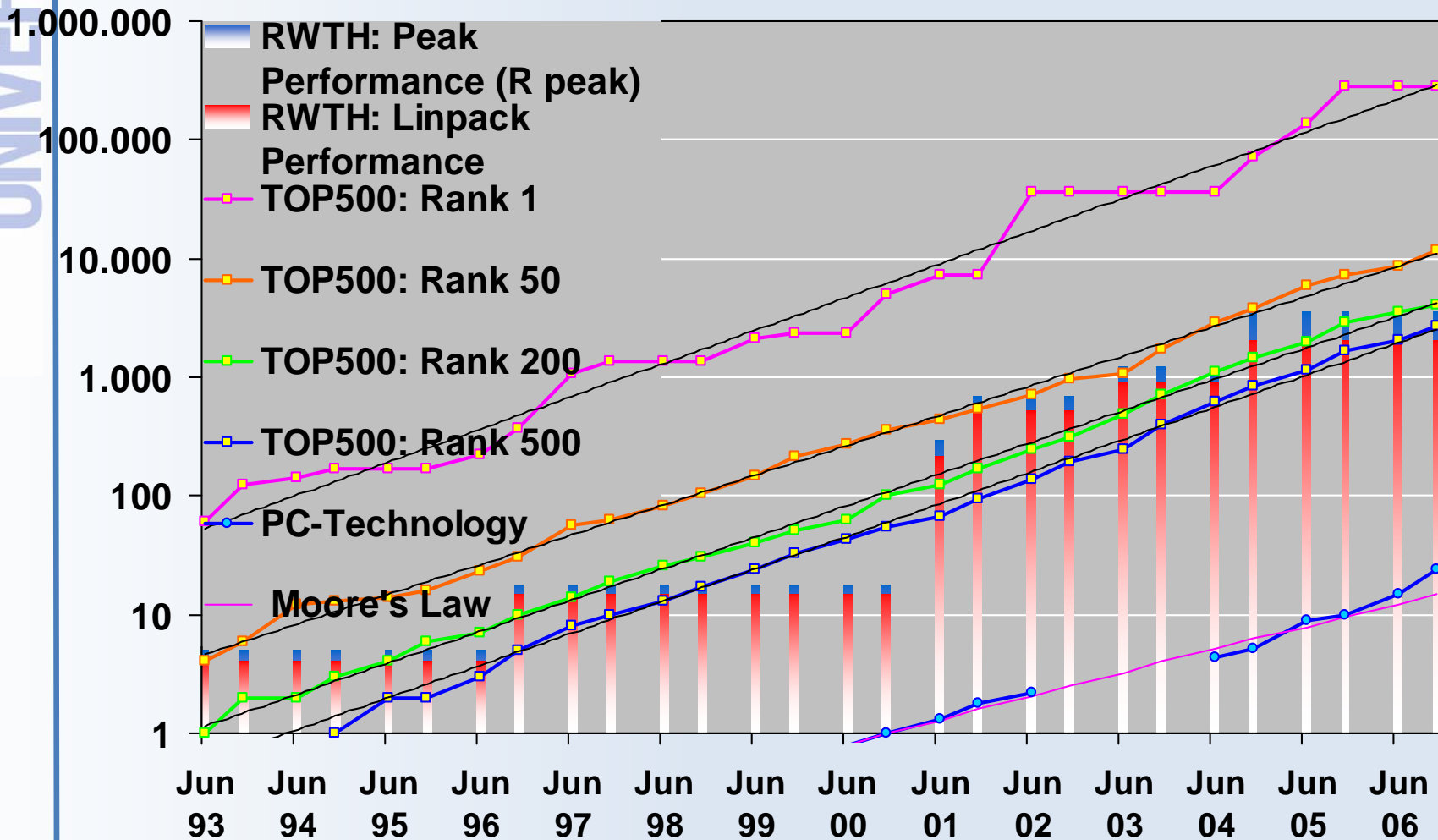
HPC  
OverviewWindows  
Cluster

Top500

Case Studies

Summary

# The Top500 List ([www.top500.org](http://www.top500.org))



19

# The Top500 List: Windows clusters

## Operating system Family share for 11/2007

In addition to the table below, you can view the visual charts using the [TOP500 charts page](#). A direct link to the charts is also [available](#).

Operating system Family	Count	Share %	Rmax Sum (GF)	Rpeak Sum (GF)	Processor Sum
Linux	426	85.20 %	4897046	7956758	970790
Windows	6	1.20 %	47495	86797	12112
Unix	30	6.00 %	408378	519178	73532
BSD Based	2	0.40 %	44783	50176	5696
Mixed	34	6.80 %	1540037	1900361	580693
Mac OS	2	0.40 %	28430	44816	5272
<b>Totals</b>	<b>500</b>	<b>100%</b>	<b>6966169.82</b>	<b>10558086.75</b>	<b>1648095</b>

- Why add just another Linux cluster? → Top500 on Windows!
- Watch out for the Windows/Unix ratio in the upcoming Top500 list in June...

20

Center for

Computing and  
CommunicationHPC  
OverviewWindows  
Cluster

Top500

Case Studies

Summary

# The LINPACK benchmark

- Ranking is determined by running a single benchmark: HPL
  - Solve a dense system of linear equations
  - Gaussian elimination-type of algorithm:  $\frac{2}{3}n^3 + O(n^2)$
  - Very regular problem → high performance / efficiency
  - [www.netlib.org/hpl](http://www.netlib.org/hpl)
- Result is not of high significance for our applications ...
- ... but it's a pretty good sanity check for the system!
  - Low performance → Something is wrong!
- Steps to success:
  - Cluster setup & sanity check
  - Linpack parameter tuning

21

# Results of initial Cluster Sanity Check

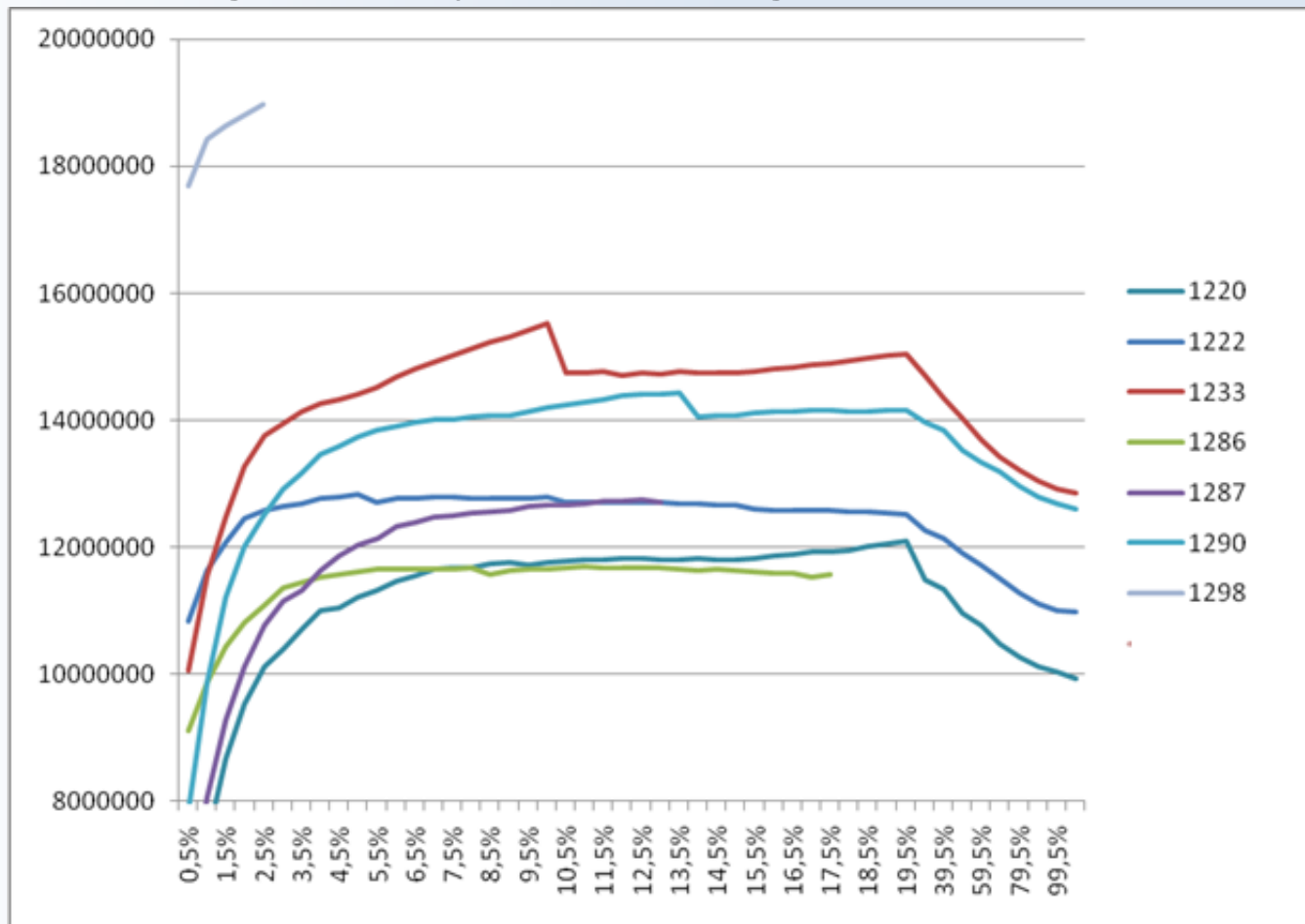
- The cluster was in normal operation since end of January
- Bandwidth test: Variations from 180 MB/s to 1180 MB/s

		Nodes 1 to 9								
Nodes 1 to 21		714	719,2	719,3	719,7	719,4	649,2	719,3	719,1	649
		184,6	185,1	185	185,1	185,1	185	185	185	185,1
		1162,6	1177,3	1180,8	1175,4	1173,4	722,4	1176,3	1179,5	723
		1159,6	1178,4	1180,6	1180	1176,5	722	1174,7	1181,6	722,5
		714,5	718,5	720,1	719,8	719,7	649	719,8	719,8	648,6
		1163	1178,2	1176,1	1178,4	1173,7	722,8	1177,7	1180,1	721,8
		1162,5	1175,4	1031,5	1125,4	1174,6	722,4	1179,3	1179,6	723,2
		1164,3	1177,9	1181,9	1178,2	1175,8	721,7	1176,4	1180,4	722,8
		1160,2	1178,2	1180	1179,4	1177,2	722,7	1177,8	1179,2	723,3
		710,8	716,2	717,1	716,4	717,8	649	717,5	718	649
		186,9	187,3	187,3	187,3	187,3	187,3	187,3	187,3	187,3
		184,6	185	185	185	185	185,1	185	185	185
		1160,1	1177,5	1179,1	1177,8	1175	722,4	1174,4	1181,5	721,8
		1161,5	1174,7	1179,6	1177,7	1175,5	721,9	1176,5	1176,6	723
		1165,8	1179	1176,1	1180,1	1176,7	722,4	1175	1180,8	722,7
		1164,7	1178,5	1178,3	1174,7	1175,7	722,3	1177,3	1179,5	722,7
		1165,7	1177	1179	1178,6	1177,8	721,2	1172,9	1179,2	721,9
		1162,9	1178,9	1177,1	1180,6	1177,7	722	1176,2	1178,3	721,2
		713,5	720,3	718,2	720,2	718,4	649,4	719,6	719,8	648,8
		1161,5	1178,8	1180,3	1173,9	1176	722,4	1176,8	1179,8	722,4
		1165	1177,1	1178,7	1179,4	1173,5	723,2	1178,5	1177,9	722,2

22

# Parameter Tuning

- HPC approach: Performance Analysis → Tuning → Analysis → Tuning → Analysis → Tuning → ...



Performance got better over time:  
System and Parameter Tuning

# Parameter Tuning – Going the Windows way

- o Excel application on a laptop controlled the whole cluster!

hpl parameters					Estimation data			Actual Results			Job		
N	NB	P	Q	Thread	Node	Expected Runtime	Total Work	Gflops	Efficiency	Job#	Status	Result File	
40704	192	1	8	1	1	623	41871.66	75.99	79.2%	45	Finished	45.log	
40960	128	1	8	1	1	635	42666.67	73.55	76.6%	47	Finished	47.log	
40896	144	1	8	1	1	632	42466.98	75.38	78.5%	48	Finished	48.log	
40960	160	1	8	1	1	635	42666.67	74.19	77.3%	49	Finished	49.log	
40832	176	1	8	1	1	629	42267.92	76.21	79.4%	50	Finished	50.log	
40768	208	1	8	1	1	626	42069.47	76.69	79.9%	52	Finished	52.log	
41280	240	1	8	1	1	650	43674.5	74.73	77.8%	54	Finished	54.log	
40960	256	1	8	1	1	635	42666.67	74.20	77.3%	55	Finished	55.log	
41344	272	1	8	1	1	653	43877.95	73.87	76.9%	56	Finished	56.log	
40320	288	1	8	1	1	606	40697.75	73.65	76.7%	57	Finished	57.log	
41344	304	1	8	1	1	653	43877.95	74.00	77.1%	58	Finished	58.log	
40960	320	1	8	1	1	635	42666.67	72.26	75.3%	59	Finished	59.log	
40320	336	1	8	1	1	606	40697.75	73.72	76.8%	60	Finished	60.log	
40832	352	1	8	1	1	629	42267.92	73.84	76.9%	61	Finished	61.log	
41216	368	1	8	1	1	647	43471.68	73.23	76.3%	62	Finished	62.log	
39936	384	1	8	1	1	588	39546	71.57	74.6%	63	Finished	63.log	
40000	400	1	8	1	1	591	39736.43	73.33	76.4%	64	Finished	64.log	
39936	416	1	8	1	1	599	39546	71.15	74.1%	65	Finished	65.log	

24

We easily reached 76,69 GFlop/s on one node.  
 Peak performance is: 8 cores \* 3 GHz \* 4 results per cycle = 96 Gflop/s  
 → That is 80% efficiency!

# Windows HPC Server 2008 in Action

- Job startup comparison – 2048 MPI processes:
  - Our Linux configuration: Order of Minutes
  - Our Windows configuration: Order of Seconds

Video: Job startup  
on Windows 2008

25

Center for

Computing and

Communication

HPC  
Overview

Windows  
Cluster

Top500

Case Studies

Summary

# Agenda

- High Performance Computing (HPC)
  - OpenMP & MPI & Hybrid Programming
- Windows-Cluster @ Aachen
  - Hardware & Software
  - Deployment & Configuration
- Top500 submission
- Case Studies
  - Dynamic Optimization: AVT
  - Bevel Gears: WZL
  - Application & Benchmark Codes
- Summary

26

Center for

Computing and

Communication

HPC  
OverviewWindows  
Cluster

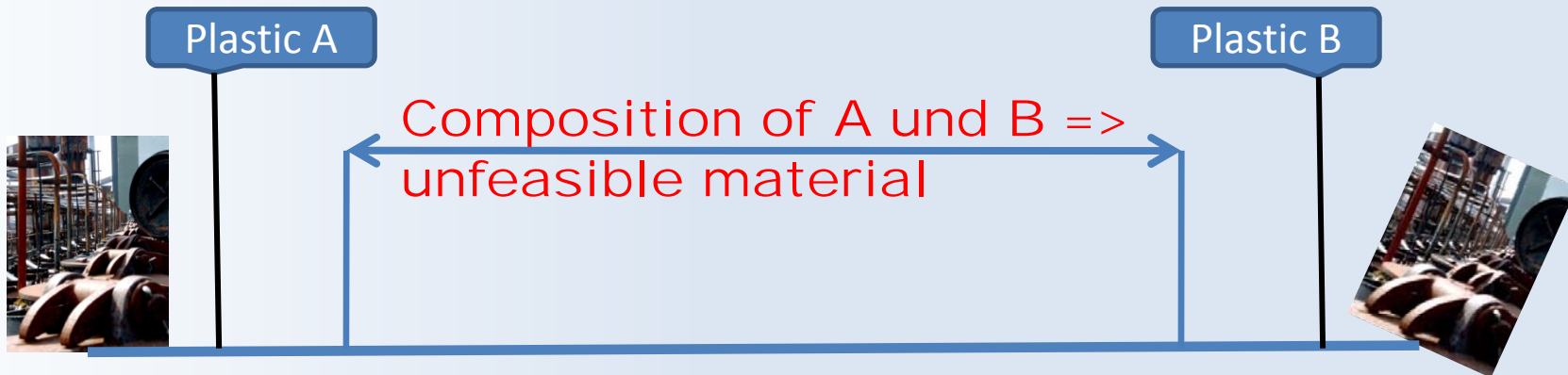
Top500

Case Studies

Summary

# Case Study: Dynamic Optimization (AVT)

## o Dynamic optimization in chemical industry



Task: Changing the product specification (from A to B) of a plastic manufactory in ongoing business

- Minimize the junk (composition of A and B)
- Search for economic and ecologic optimal operational mode!

This task is solved by the DyOS (Dynamic Optimization Software) tool, developed at the chair for process systems engineering (AVT: Aachener Verfahrenstechnik) at RWTH Aachen University.

27

Center for

Computing and

Communication

HPC  
Overview

Windows  
Cluster

Top500

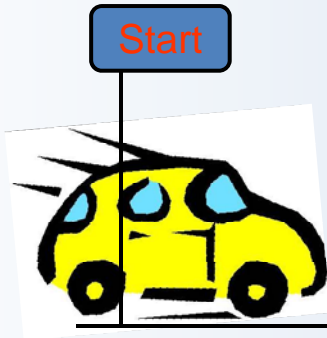
Case Studies

Summary

# Case Study: Dynamic Optimization (AVT)

## o What is dynamic optimization?

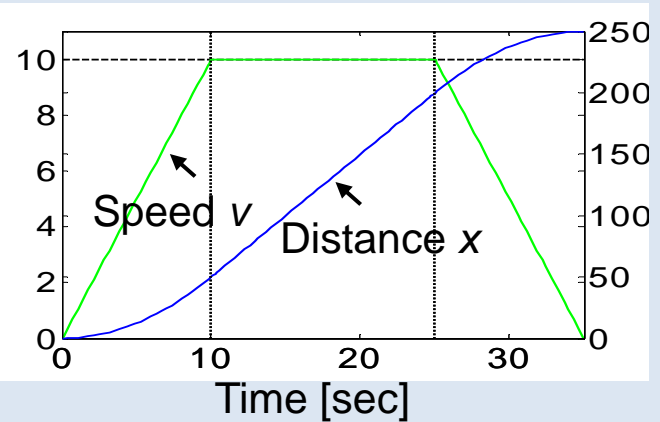
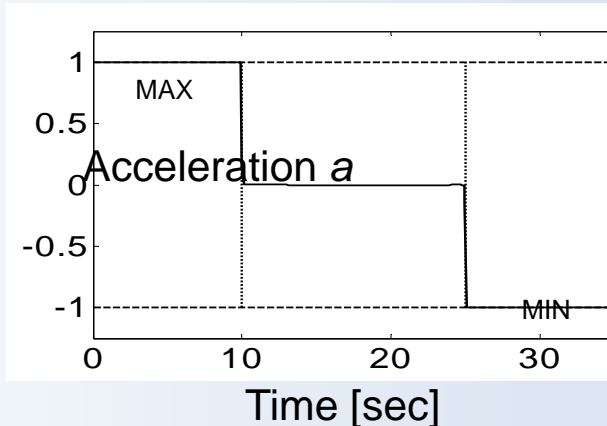
- Drive 250 m in minimal time!
- Start with  $v = 0$ , stop exactly at  $x = 250\text{m}$ !
- Maximal speed is  $v = 10\text{ m/s}$ !
- Maximal acceleration is  $a = 1\text{ m/s}^2$ !



$$v(t) \leq v_{\max}$$

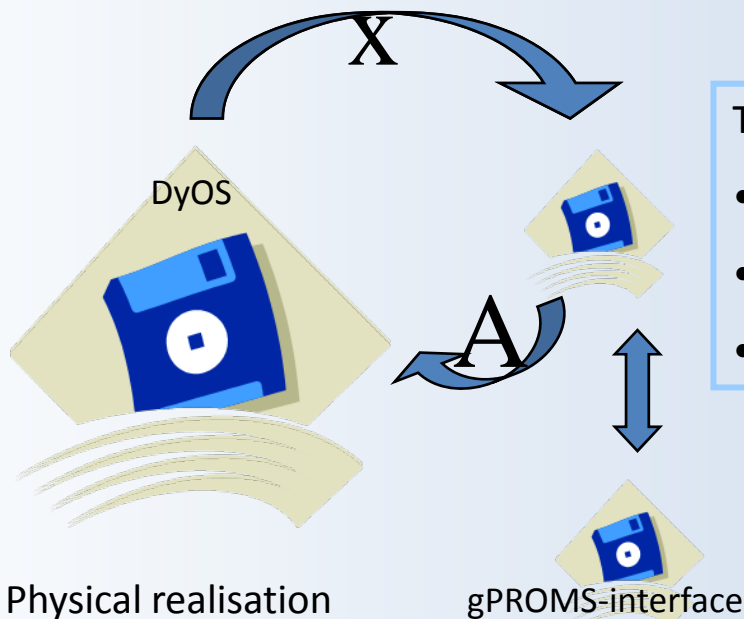
$$x(0), v(0) = 0$$

$$v(t_f) = 0, x(t_f) \geq x_{t_f}$$



## Case Study: Dynamic Optimization (AVT)

- One simulation typically takes up to two weeks and requires a significant amount of memory (>4GB) → MPI parallelization
- Challenge: Commercial software component depending on VS6, only one instance per machine allowed → outsourced into DLL



Tasks of MPI communicator (ROOT):

- storing data
- sending and receiving data
- control over entire software

29

## Case Study: Dynamic Optimization (AVT)

- A scenario consists of several simulations
- Projected compute time on desktop:
  - 5 scenarios à 2 months = 10 months
- Compute time on our cluster (faster machines, multiple jobs):
  - 5 scenarios à 1.5 months = 3 months
- MPI parallelization lead to factor 4.5 on 8 cores:
  - 5 scenarios à 0.3 months = 0.7 months
- Arndt Hartwich: Stability of RZ's compute nodes is significantly higher than stability of our desktops.

30

Center for

Computing and

Communication

HPC  
OverviewWindows  
Cluster

Top500

Case Studies

Summary

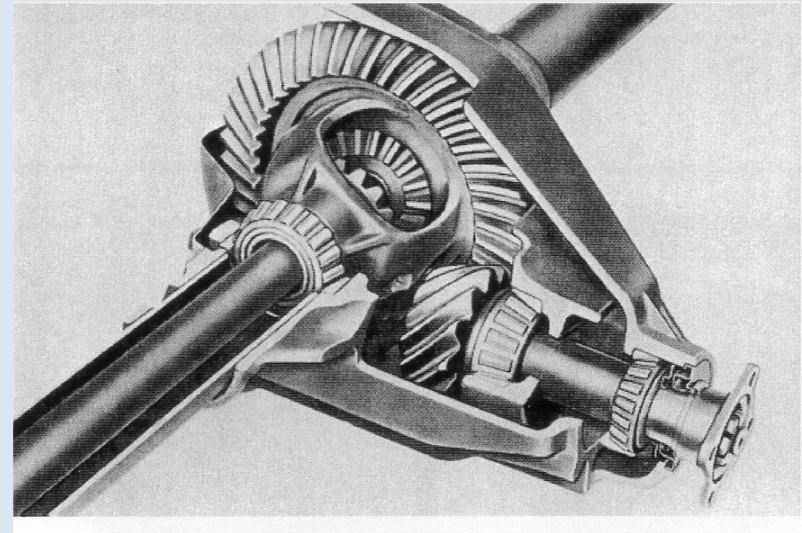
## Case Study: KegelToleranzen (WZL)

- Simulation of Bevel Gears
  - Written in Fortran, using Intel Fortran 10.1 compiler
  - Very cache-friendly → runs at high Mflop/s rates

Bevel Gear Pair



Differential Gear



31

Laboratory for Machine Tools and Production Engineering, RWTH Aachen

Center for

Computing and

Communication

HPC  
Overview

Windows  
Cluster

Top500

Case Studies

Summary

## Case Study: KegelToleranzen (WZL)

- Target
  - Pentium/Windows/Intel → Xeon/Windows/Intel
  - Serial Tuning + Parallelization with OpenMP
- Procedure
  - Get the tools: Porting to UltraSparc IV/Solaris/Sun Studio
  - Simulog Foresys: Convert to Fortran 90
    - 77000 Fortran77 lines → 91000 Fortran 90 lines
  - Sun Analyzer: Runtime Analysis with different datasets
    - Deduce targets for Serial Tuning and OpenMP Parallelization
  - OpenMP Parallelization: 5 Parallel Regions, 70 Directives
  - Get the tools: Porting new code to Xeon/Windows/Intel
  - Intel Thread Checker: Verification of OpenMP Parallelization
- Put new code in production on Xeon/Windows/Intel

32

Center for

Computing and

Communication

HPC  
OverviewWindows  
Cluster

Top500

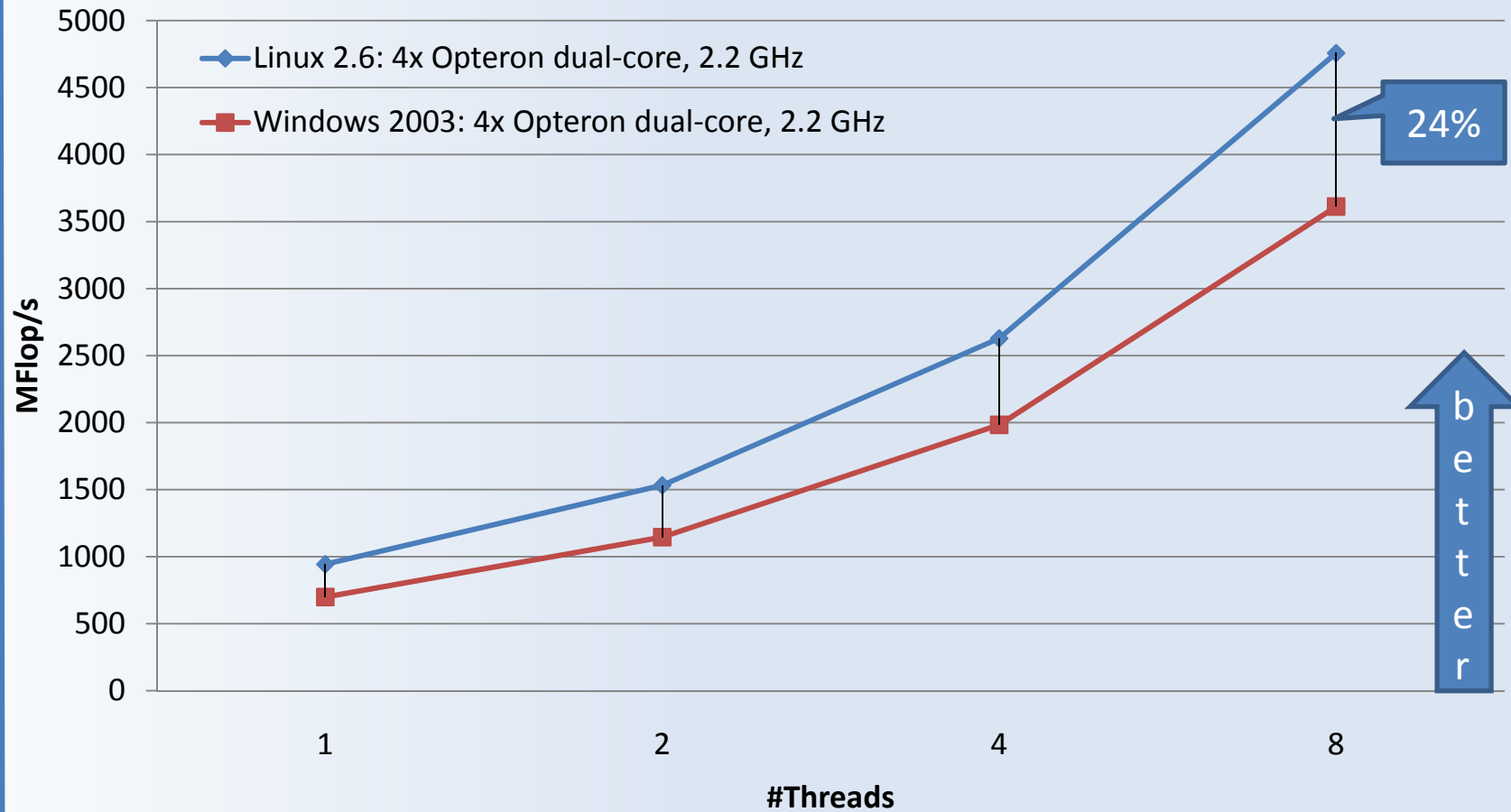
Case Studies

Summary

# Case Study: KegelToleranzen (WZL)

## Comparing Linux and Windows Server 2003:

Performance of KegelToleranzen

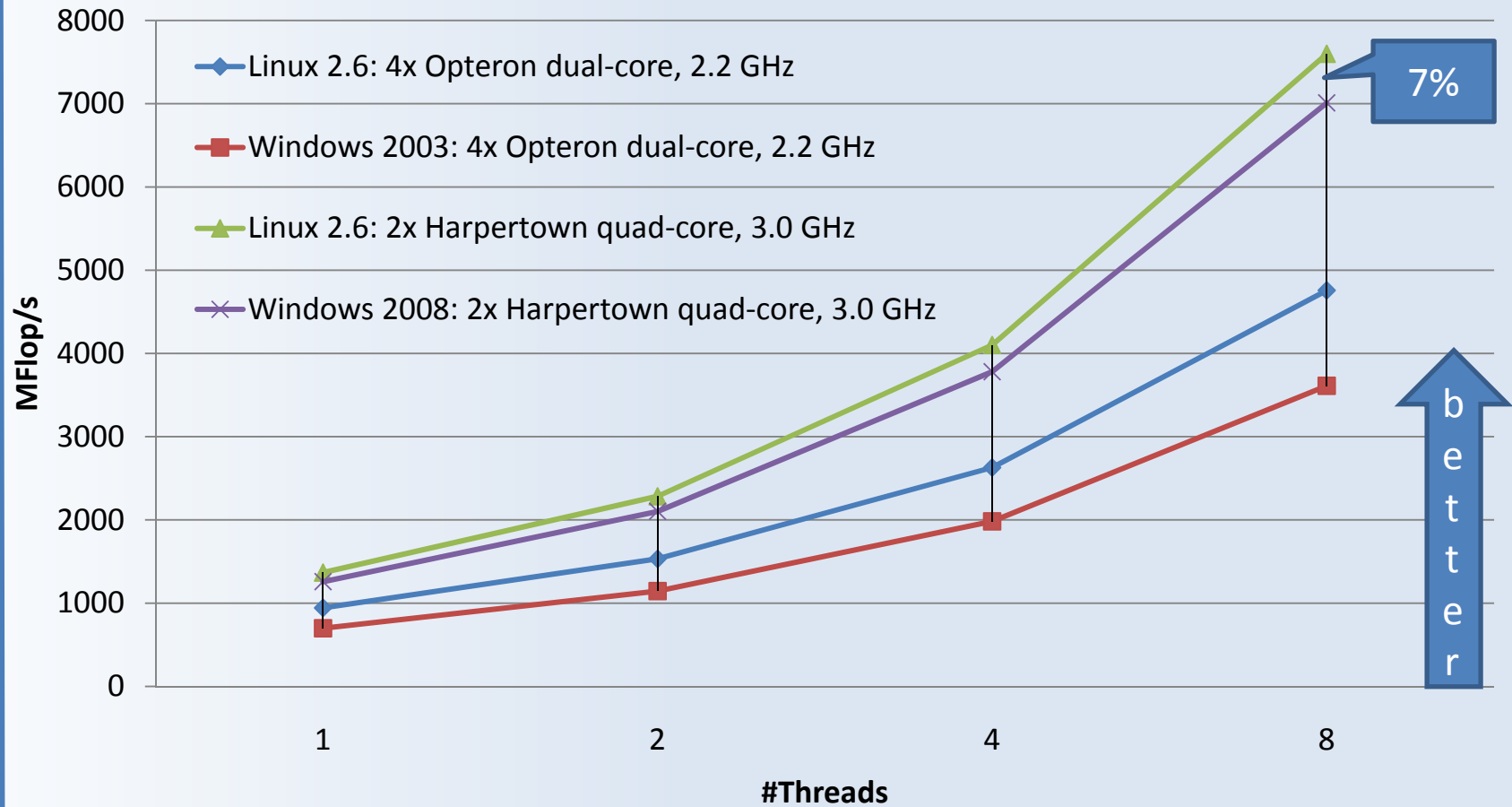


33

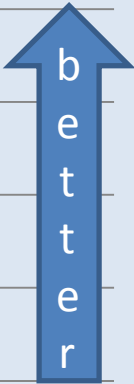
# Case Study: KegelToleranzen (WZL)

## Comparing Linux and Windows Server 2008:

Performance of KegelToleranzen



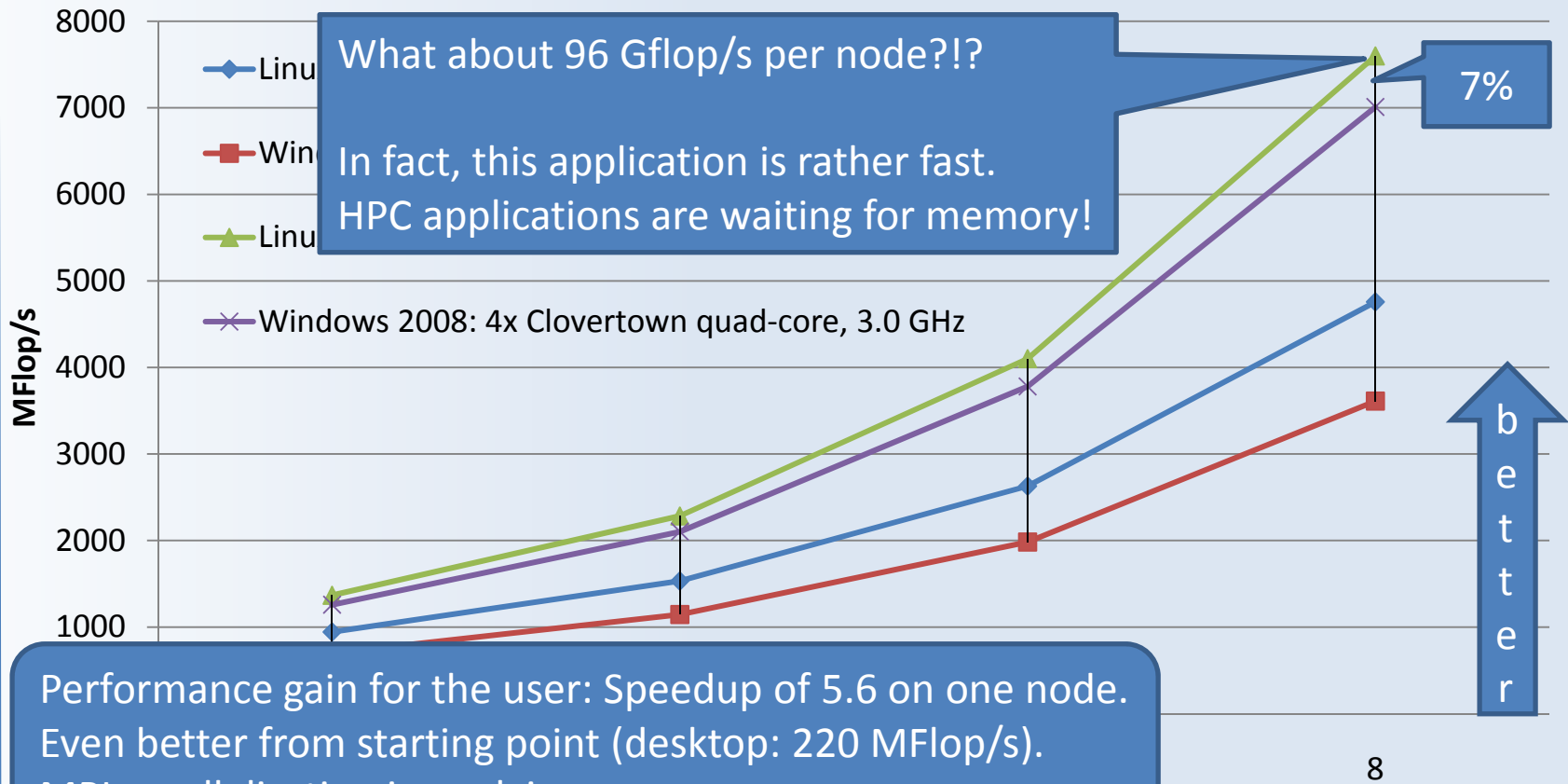
34



# Case Study: KegelToleranzen (WZL)

## Comparing Linux and Windows Server 2008:

Performance of KegelToleranzen



35

Performance gain for the user: Speedup of 5.6 on one node. Even better from starting point (desktop: 220 MFlop/s). MPI parallelization is work in progress.

# Agenda

- High Performance Computing (HPC)
  - OpenMP & MPI & Hybrid Programming
- Windows-Cluster @ Aachen
  - Hardware & Software
  - Deployment & Configuration
- Top500 submission
- Case Studies
  - Dynamic Optimization: AVT
  - Bevel Gears: WZL
  - Application & Benchmark Codes
- Summary

36

Center for

Computing and

Communication

HPC  
OverviewWindows  
Cluster

Top500

Case Studies

Summary

# Invitation: Windows-HPC User Group Meeting

Einladung zum 1. Treffen der  
Windows High Performance Computing  
Nutzergruppe im deutschsprachigen Raum

21./22. April 2008

Rechen- und Kommunikationszentrum, RWTH Aachen

<http://www.rz.rwth-aachen.de/winhhpcug>

LY PRODUCTIVE  
HIGH PERFORMANCE COMPUTING

37

Center for

Computing and

Communication

HPC  
Overview

Windows  
Cluster

Top500

Case Studies

Summary

# Invitation: Windows-HPC User Group Meeting

- Ziele
  - Informationsaustausch zwischen Nutzern untereinander
  - Informationsaustausch zwischen Nutzern und Microsoft
- Montag, 21. April
  - 17:00h: Domführung
  - 18:30h: Gemeinsames Abendessen
- Dienstag, 22. April
  - Präsentationen der Firmen Alinea, Cisco, Intel, The Mathworks und Microsoft
  - Präsentationen von Einrichtungen aus Forschung und Lehre
  - Frage-und-Antwort-Runde mit HPC-Experten von Microsoft

38

## Summary & Outlook

- The Windows-HPC environment has been well accepted
  - Growing interest and need of compute power on Windows.
- Windows HPC Server 2008 (beta) is pretty impressive!!!
  - Deploying and Configuring 256 nodes.
  - Job Startup and Job Management.
  - Performance improvements & Linpack efficiency.
- Interoperability in heterogeneous environments got easier.
  - E.g. WDS can interoperate with Linux DHCP and PXE-Boot.
- Performance: Need for Windows-specific tuning.

39

Center for

Computing and

Communication

HPC  
OverviewWindows  
Cluster

Top500

Case Studies

Summary

# The End

Thank you for  
your attention!

40

Center for

Computing and  
Communication

HPC  
Overview

Windows  
Cluster

Top500

Case Studies

Summary