

Windows-HPC Environment at RWTH Aachen University

Christian Terboven, Samuel Sarholz
{terboven, sarholz}@rz.rwth-aachen.de

Center for Computing and Communication
RWTH Aachen University



Agenda

- HPC @ RZ
- Cluster Overview
- Filesystems
- Software
- Batch System
- IDEs



2

Center for

Computing and

Communication

HPC @ RZ

Cluster

Filesystems

Software

Batch

IDEs

The RZ Compute Cluster History

- since 1958: Vector and other Super Computers
- 1994: The Unix Cluster started with IBM machines
- 2001-2007: SMP-Cluster with Sun UltraSparc-III/-IV systems
- 2004: First x86-based systems with 64 Opteron cluster nodes, mainly with Linux, some with Solaris for x86
- 2006: First Windows compute nodes on Opteron cluster
- 2008: Procurement of “intermediate” Intel Xeon Cluster with InfiniBand interconnection network
- 2009-2010: New procurement, new fileserver infrastructure



3

HPC @ RWTH Aachen: Objectives

- HPC on Unix and Windows is a service offered by the Center for Computing and Communication:
 - Account provisioning via TIM → one account (login+pw) for Solaris, Linux and Windows
 - Files on Unix are accessible from Windows because of same file infrastructure (\$HOME = H:, \$WORK = W:)
 - Operating Model: Interactive Machines + Batch System
 - Programming and Software Support:
 - Languages: C, C++, Fortran (, Java, C#)
 - Parallelization: MPI, OpenMP, Intel TBB, Native Threading
 - ISV-Codes: Matlab, Ansys, numerical libraries, ...
 - User training on all platforms!
- HPC service is open for employees and students as well!



4

Agenda

- HPC @ RZ
- Cluster Overview
- Filesystems
- Software
- Batch System
- IDEs



5

Center for

Computing and

Communication

HPC @ RZ

Cluster

Filesystems

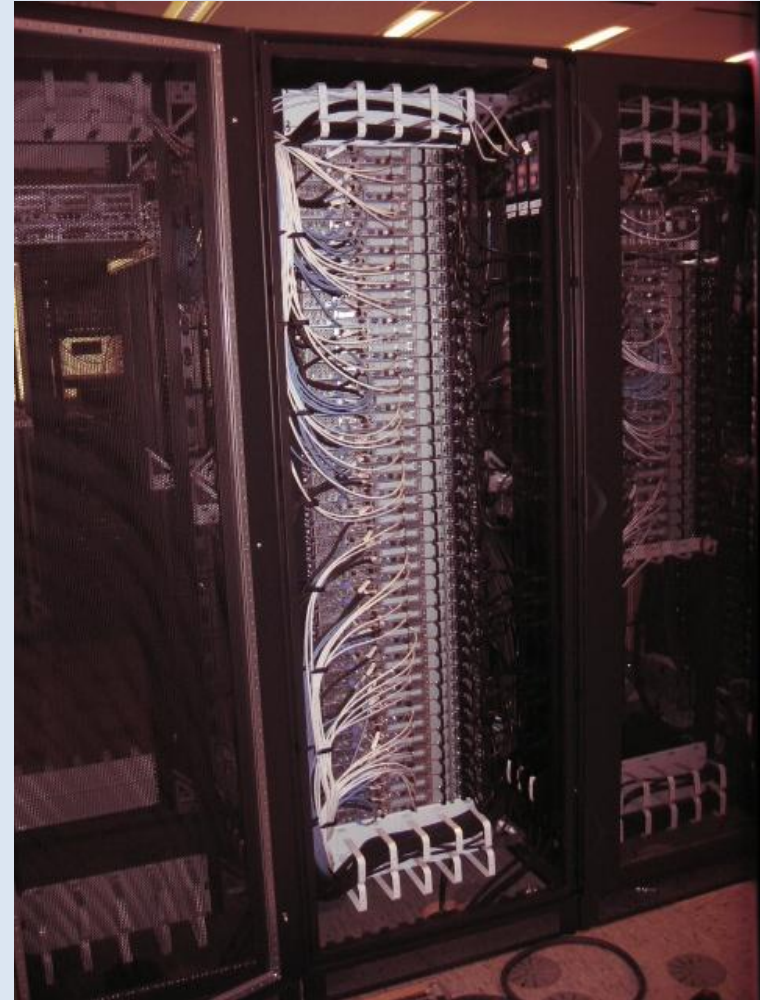
Software

Batch

IDEs

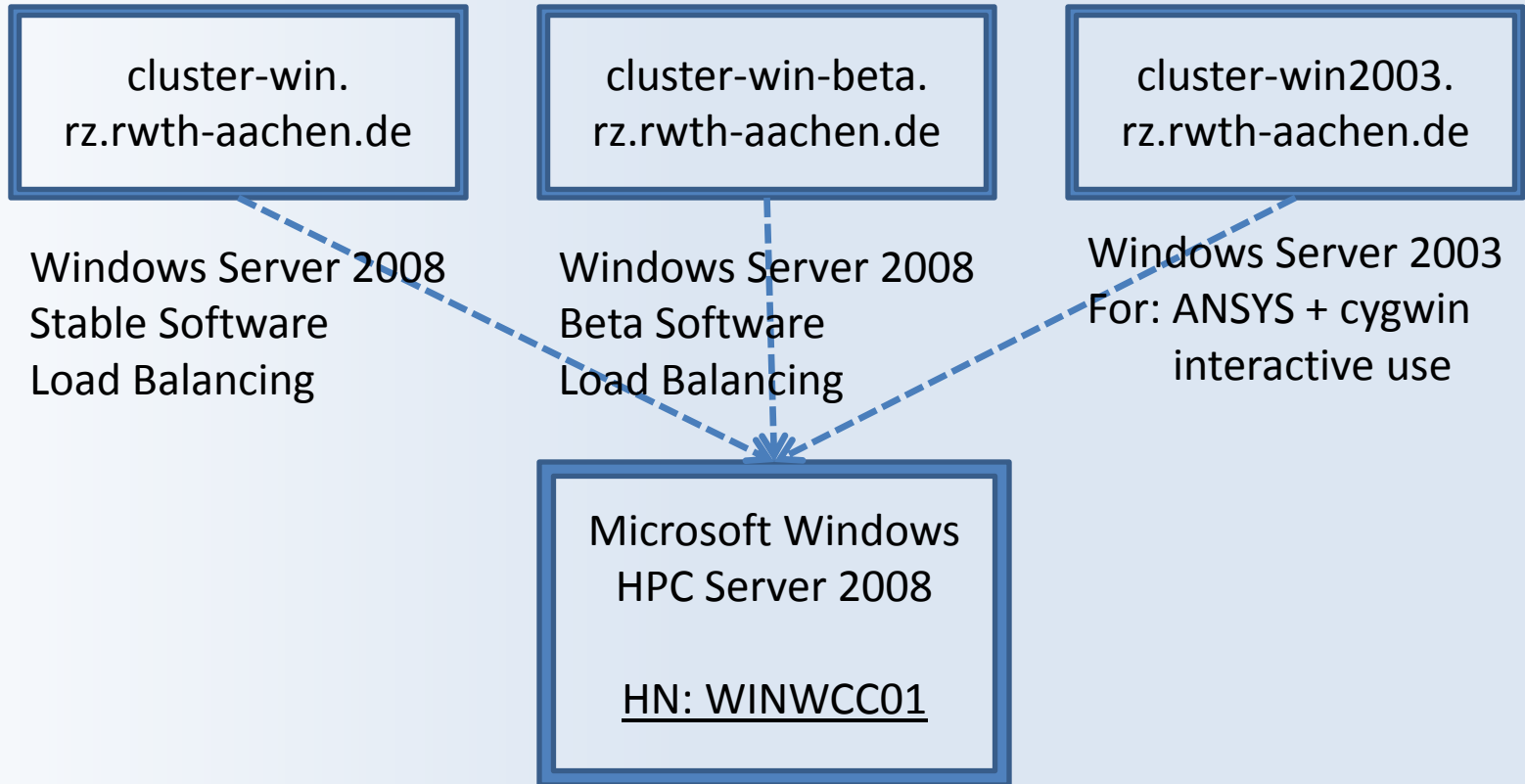
Intel Harpertown-based InfiniBand Cluster

- Cluster installed in Q1/2008:
 - 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:
 - MPI latency: 4.5 us
 - MPI bandwidth: 1250 MB/s
- Installation-on-demand:
Linux + Windows
- Rank 100 in Top500 in 06/2008!
 - 18.81 TFlop/s with 256 nodes
 - 195 Mflop/s per Watt



Windows-Cluster: Frontends

- Currently we are running three Frontends for the Cluster
 - Interactive Use for Software Development and the like

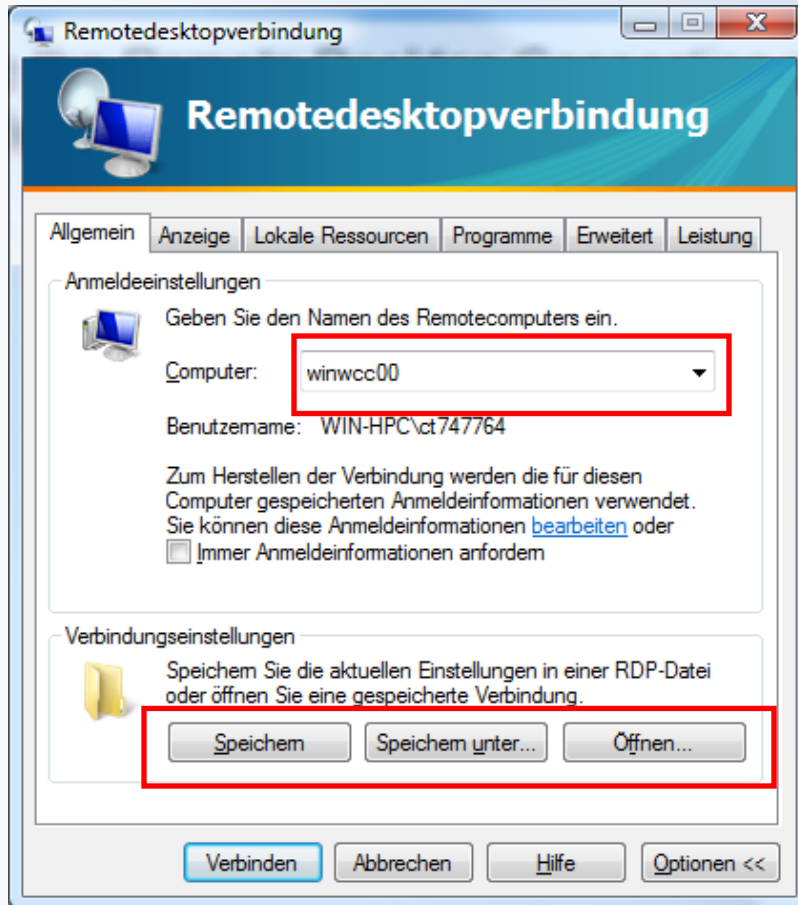


-----> Batch Job (long running Compute Job) Submission



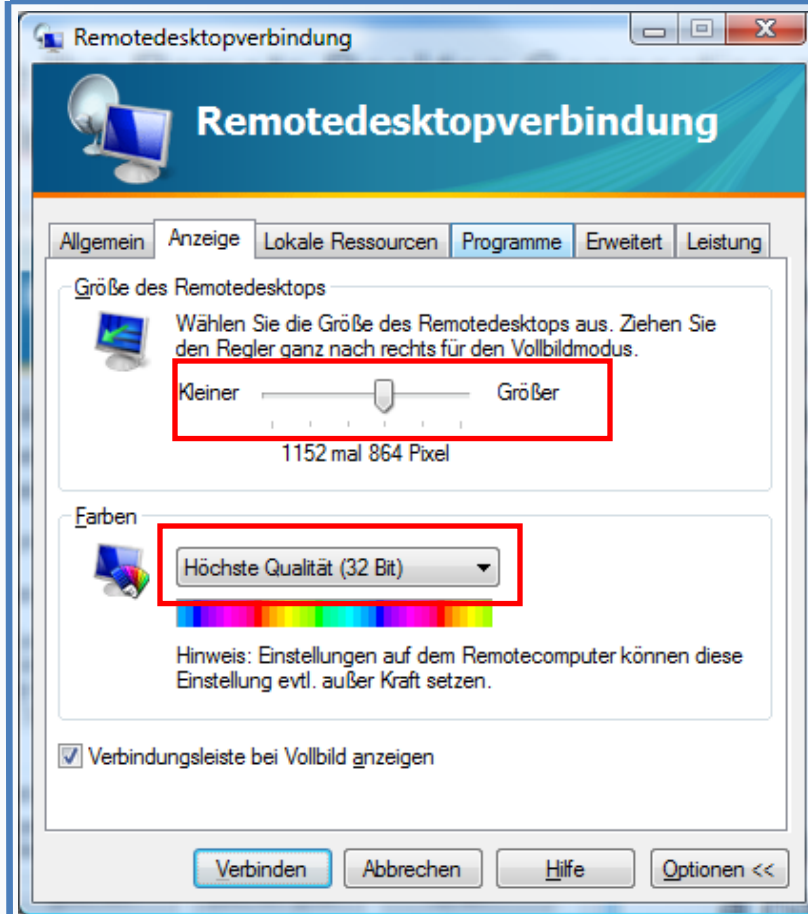
Login from Windows (1/3)

- Use the *Remote Desktop Connection* program, usually available under *All Programs* → *Accessoires* → *Communication*.



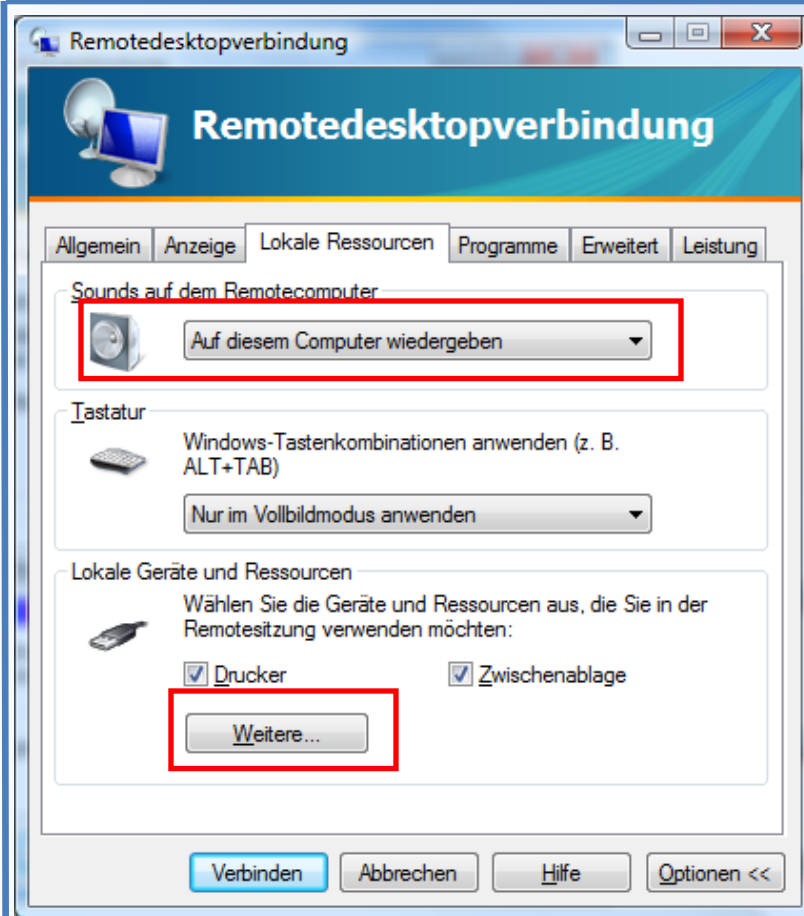
- Specify the computer name, in general `cluster-win.rz.rwth-aachen.de`.
- You can save a set of settings under a named profile / link.

Login from Windows (2/3)



- You can choose a resolution or fullscreen mode.
- You can choose the color depth.
- In fullscreen mode you should set this flag to ease the handling of the remote desktop window.

Login from Windows (3/3)



- Take resources of your local computer with you:
 - Audio device
 - Printer
 - Clipboard
 - Local hard disc drives
 - Locally mounted network drives

Login from Linux

- Use the `rdesktop` program available from www.rdesktop.org, probably already included in your distribution.

- Basic usage: `rdesktop [options] host` with
 - `-u <user>` Login as user `<user>`
 - `-d WIN-HPC` Login to domain WIN-HPC
 - `-4` Use protocol version 4 (often needed)
 - `-g WxH` Use resolution Width x Height
 - `-f` Use fullscreen resolution
 - `-a 24` Use 24bit color depth
 - `-k de` Use german keyboard layout
 - `-r sound:local` Play sound on local system



11

Enter for

Computing and
Communication

HPC @ RZ

Cluster

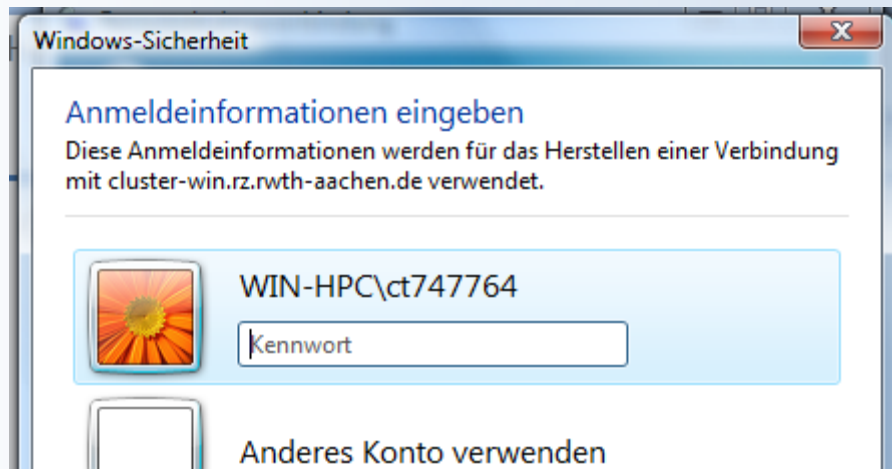
Filesystems

Software

Batch

IDEs

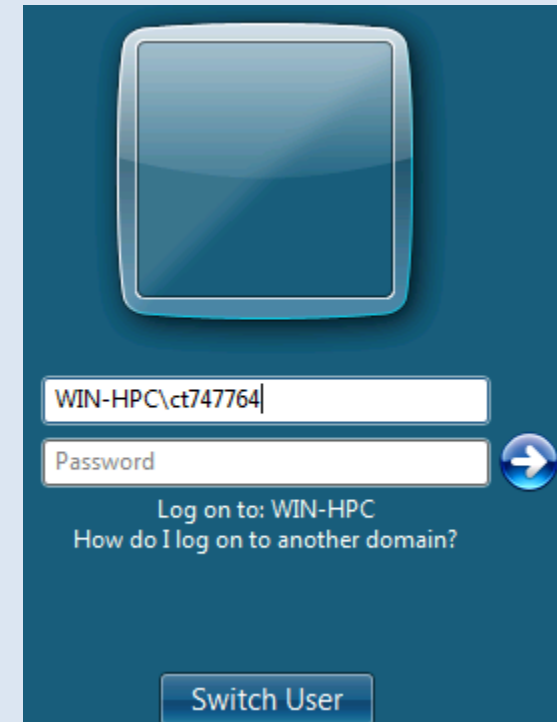
Login to Windows Server 2008



New Remote Desktop Connection program (e.g. with Vista) allows the specification of username before login.

If domain selection is not possible

- Username: WIN-HPC\.....
- Or: Specify in rdesktop program



Agenda

- HPC @ RZ
- Cluster Overview
- Filesystems
- Software
- Batch System
- IDEs



13

Center for

Computing and

Communication

HPC @ RZ

Cluster

Filesystems

Software

Batch

IDEs

File Storage Strategies (1/2)

- Home - H: on Windows and \$HOME on Unix
 - Permanent and long-term data (full backup)
- Work – W: on Windows and \$WORK on Unix
 - Large datasets or near-term data (no backup)
- Documents – X: on Windows
 - Windows „My Documents“ directory (full backup)
 - Also accessible via H:\WinDocuments
- Temp – D:\Temp\<>userid> on Windows and \$TMP on Unix
 - Temporary data (no backup,)
- All directories/shares have a quota, that is a size limitation.
Need more space? → email hpc@rz.rwth-aachen.de



14

Center for

Computing and

Communication

HPC @ RZ

Cluster

Filesystems

Software

Batch

IDEs

File Storage Strategies (2/2)

- Windows batch jobs cannot access H: or W: via the drive letter directly
 - Network paths have to be used:
 - H: is `\\cifs\cluster\home\userid`
 - W: is `\\cifs\cluster\work\userid`
 - X: is `\\cifs\cluster\documents\userid`
 - C:\Shared_Software is `\\cifs\cluster\software`
 - Either use those, or connect network drive in batch script:
 - `net use H: \\cifs\cluster\home\userid`

- Snapshots on H: and W are accessible via Windows Explorer:
 - Access to older - already overwritten - versions of a file
 - Right click on file → Properties → Previous Versions



15

Center for

Computing and

Communication

HPC @ RZ

Cluster

Filesystems

Software

Batch

IDEs

Agenda

- HPC @ RZ
- Cluster Overview
- Filesystems
- **Software**
- Batch System
- IDEs



16

Center for

Computing and

Communication

HPC @ RZ

Cluster

Filesystems

Software

Batch

IDEs

Software list (1/3)

- Complete set of Development Software:
 - cluster-win:
 - Visual Studio 2005 and Visual Studio 2008
 - Intel Compiler Suite 11 (C/C++ and Fortran)
 - Microsoft HPC Pack 2008 (= MS-MPI)
 - Intel Cluster Toolkit 3.1
 - Intel MPI 3.1 (= I-MPI)
 - Intel Trace Analyzer & Collector 7.1
 - Intel Threading Building Blocks 2.0
 - Intel VTune 9.0 + Intel Threading Tools 3.1
 - cluster-win-beta: same as above, but / plus (+)
 - + Visual Studio 2008 with Intel Parallel Studio
 - Visual Studio 2010 as soon as beta will be available
 - Intel Compiler Suite 11.1 beta (C/C++ and Fortran)



17

Center for

Computing and

Communication

HPC @ RZ

Cluster

Filesystems

Software

Batch

IDEs

Software list (2/3)

- Selected (and growing) list of tools and utilities:
 - All interactive machines:
 - Notepad++ editor
 - Subversion Client
 - Tortoise Subversion GUI / Explorer integration
 - X-Win32
 - cluster-win-beta: same as above, plus (+)
 - + Several SDKs and Windows Debugging / Analysis tools
- Selected (and growing) list of ISV-Software:
 - ANSYS (for interactive use go to `cluster-win2003`)
 - HyperWorks
 - Fluent
 - Maple



18

Center for

Computing and

Communication

HPC @ RZ

Cluster

Filesystems

Software

Batch

IDEs

Software list (3/3)

- Selected (and growing) list of ISV-Software:
 - Mathematica
 - Matlab
 - Microsoft Office 2003
 - Microsoft Excel Compute Services
 - MSC.Marc
 - MSC.Adams
 - Linear Algebra Libraries (e.g. Intel MKL 10.0)
 - ...
- If we have (floating) licenses and if the software is available on Windows, we will provide it.
- We make user-provided software available as well (if possible without giving administrator privileges away).

If there is something missing, please let us know ...



19

Center for

Computing and

Communication

HPC @ RZ

Cluster

Filesystems

Software

Batch

IDEs

ISV codes in the batch system (1/3)

- 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 sequential Matlab job:

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

WinHP3C

20

enter for

Computing and

Communication

HPC @ RZ

Cluster

Filesystems

Software

Batch

IDEs

ISV codes in the batch system (2/3)

- Exemplary usage instructions for sequential ANSYS job:

1. Create an Inputfile
2. Create a .CMD file containing the following lines

```
setlocal
set INPFILE=test.txt
set OUTFILE=%INPFILE%.out
set ANSCMD_NODIAG=TRUE
net use x: %CCP_WORKDIR%
x:
call "\\cifs\cluster\software\ansys inc
\v110\ansys\bin\winx64\ansys110.exe" -b nolist -j
jobname -p aa_r -i %INPFILE% -o %OUTFILE%
endlocal
```

- We have only two ANSYS parallel licenses. See examples:
\\cifs\cluster\software\ansys inc\v110\ANSYS\MSCCS



21

Center for

Computing and

Communication

HPC @ RZ

Cluster

Filesystems

Software

Batch

IDEs

ISV codes in the batch system (3/3)

- Exemplary usage instructions for sequential ANSYS CFX job:
 - Use the following command line with suited input file:
`\\cifs\cluster\software\"Ansys Inc"\v110\CFX\bin\cfx5solve.exe -def input.def`
- Parallel Jobs: Use either Solvermanager (GUI):
 1. Specify Definition File
 2. Run mode: „Submit to CCS Queue“
 3. Use „+“ to specify the number of cores (80 licenses)
 4. Ignore Hostname
 5. Take care: Result path `\\cifs\cluster\documents\%username%` is hard-coded
- or have full control (and knowledge) of what you are doing and adapt the job file on our homepage to your needs.

Agenda

- HPC @ RZ
- Cluster Overview
- Filesystems
- Software
- Batch System
- IDEs



23

Center for

Computing and

Communication

HPC @ RZ

Cluster

Filesystems

Software

Batch

IDEs

Running Parallel Programs

- Multi-Threading with OpenMP
 - Control the number of threads to be used with env. Variable
 - `set OMP_NUM_THREADS=8`
 - Batch job: Reserve full node or appropriate number of cores per process and set environment variable
- Message-Passing with MPI
 - Number of processes is determined by startup command
 - `mpiexec -n 8 ...`
 - Batch job: Reserve appropriate number of nodes or cores, number of processes is then specified implicitly
- Example Collection: You can find plenty examples of parallel programs in network drive P:



24

Center for

Computing and

Communication

HPC @ RZ

Cluster

Filesystems

Software

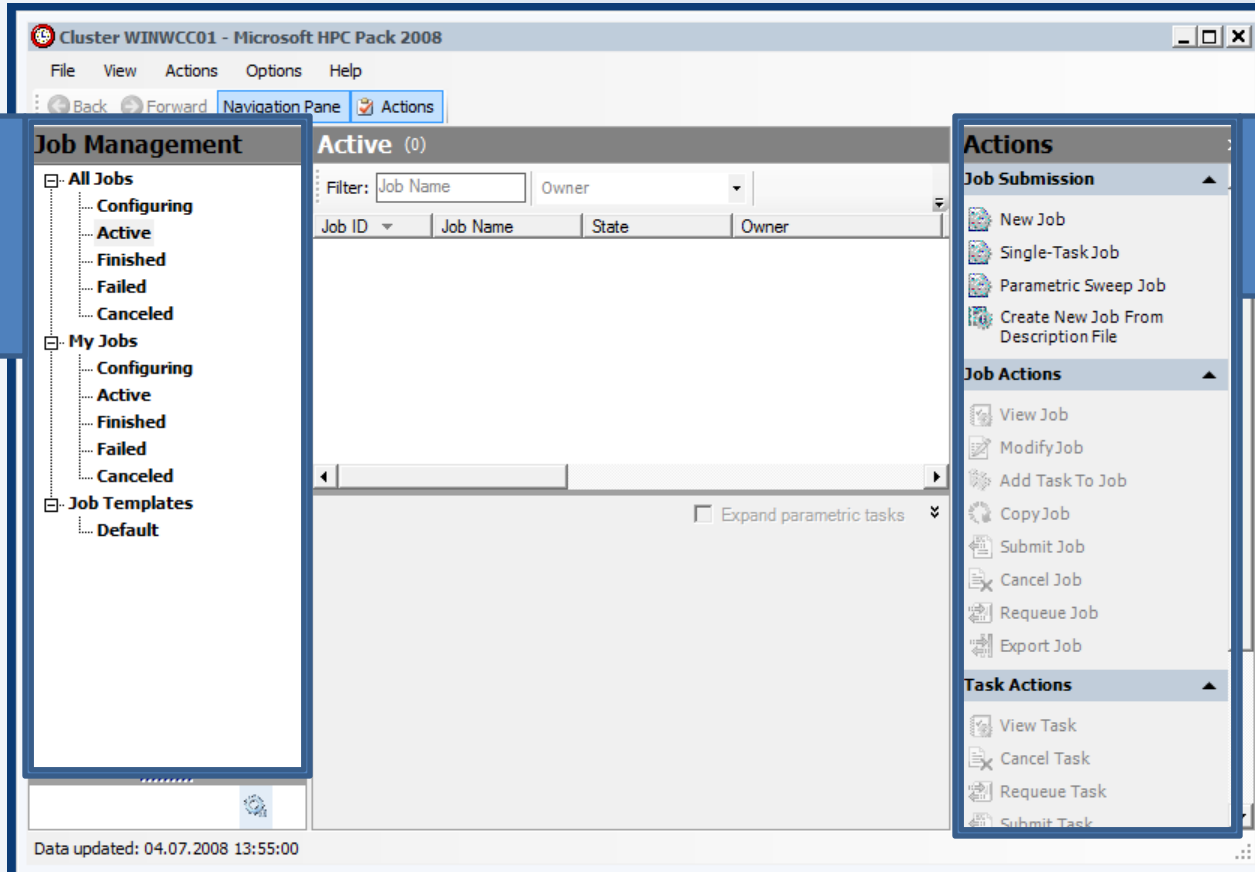
Batch

IDEs

Using the Batch System (1/5)

- Find the HPC Job Manager in the menu: *All Programs* → *Microsoft HPC Pack* → *HPC Job Manager*.

Several pre-conf. views available.



„Modern“ Action Pane.

Using the Batch System (2/5)

- To submit a new Job choose *Actions* → *Job Submission* →

Create New Job

Job Details

Job name:

Job template:

Project:

Priority:

Job run options

Do not run this job for more than:

Days: Hours: Minutes:

Run job until cancelled or run time expires

Fail the job if any task in the job fails

Job resources

Select the type of resource to request for this job:

Enter the minimum and/or maximum of the selected resource type that this job is allowed to use:

Minimum: Auto calculate

Maximum: Auto calculate

Use assigned resources exclusively for this job

No other jobs will be allowed to run on the selected nodes while the job is running.

Submit Save Job as ... Cancel

- You are free to choose a *Job Name* and a *Project Name* as you like.
- You might specify runtime and failure options for the job.
- Resource allocation changed significantly:
 - Per *Core*, or
 - Per *Socket*, or
 - Per *Node*.
- Saving of Job Templates possible.

Using the Batch System (3/5)

- Resource Allocation Granularity:
 - Per *Core*: Get n processor cores. No further restrictions, for example it cannot be assumed that a (sub)set of cores shares the same main memory (→ not suited for Shared-Memory).
 - Per *Socket*: Get n sockets. On our cluster, currently each socket has four cores (quad-core Xeon), thus it can be used for Hybrid or Shared-Memory (up to four threads per process).
 - Per *Node*: Get n nodes. On our cluster, currently each node has two sockets à four cores (dual-socket quad-core Xeon), thus it can be used for Hybrid or Shared-Memory (up to eight threads per node).
- If you use OpenMP: Set OMP_NUM_THREADS env. variable, otherwise you would get as many threads as there are cores
 - `mpiexec -genv OMP_NUM_THREADS 2`



27

Center for

Computing and
Communication

HPC @ RZ

Cluster

Filesystems

Software

Batch

IDEs

Using the Batch System (4/5)

- A Job consists of one or more Tasks.

The screenshot shows the 'Create New Job' dialog box. The 'Task Details and I/O Redirection' sub-dialog is open, showing the following fields:

- Task name: MyTask
- Command line: Task.exe
- Work directory: [Browse...]
- Standard input: [Browse...]
- Standard output: [Browse...]
- Standard error: [Browse...]

Below these fields, there are 'Minimum' and 'Maximum' resource settings, both set to 1. The background shows a table for 'Enter the tasks for this job' with columns for 'Task Name', 'Command Line', and 'Requested Resources'.

Command Line: You can specify the full path to a program including program options or to a .bat or .cmd file.

You have to use network paths instead of drive letters

(\\cifs\cluster\home
\... instead of H:) in any path.

For MPI Tasks just include mpiexec in the *Command Line*, do not specify any other MPI options.

Using the Batch System (5/5)

- Some restrictions for the node selection can be specified.

Select the resources to use for this job. Selecting a node group will filter the nodes available in the node selection list. Entering hardware preferences will limit the node groups and nodes you have selected to those that meet the specified hardware preferences.

Node preferences

Run this job only on nodes in the following node groups:

Available node groups: HeadNodes, ComputeNodes, WCFBrokerNodes

Selected node groups:

Add >> << Remove

Run this job only on nodes in the following list:

Node Name	Cores	Memory	State
<input type="checkbox"/> WINSCC002	8	16383	Ready
<input type="checkbox"/> WINSCC004	8	16383	Ready
<input type="checkbox"/> WINSCC005	8	16383	Ready
<input type="checkbox"/> WINSCC006	8	16383	Ready
<input type="checkbox"/> WINSCC007	8	16383	Ready
<input type="checkbox"/> WINSCC008	8	16383	Ready
<input type="checkbox"/> WINSCC009	8	16383	Ready
<input type="checkbox"/> WINSCC010	8	16383	Ready
<input type="checkbox"/> WINSCC011	8	16383	Ready

Hardware preferences

Minimum memory (MB): 0

Minimum cores: 0

Prefer nodes with: More Memory

Submit Save Job as ... Cancel

- Allow selected classes of nodes only.
- Allow a selected set of nodes only.
- Allow nodes with enough memory only.

Agenda

- HPC @ RZ
- Cluster Overview
- Filesystems
- Software
- Batch System
- IDEs



30

Center for

Computing and

Communication

HPC @ RZ

Cluster

Filesystems

Software

Batch

IDEs

Visual Studio: Motivation + Overview

- C / C++ / Fortran Software Development on Windows?
- My answer: Visual Studio 2008 w/ Intel Compiler Integration
- Visual Studio 2005 / 2008 for HPC Development
 - Usually command-line programs as HPC applications typically do not use GUIs. VS offers great support for GUI development on Windows, though.
 - Support for OpenMP for Shared-Memory parallel computing
 - Debugging of parallel programs: OpenMP and MPI and Hybrid
 - We provide DDTlite for improved MPI debugging experience
 - Intel Compiler Integration
 - Intel CPU-specific optimization
 - Intel Parallel Studio: Analyze + Tune + Parallelize + Check you code



31

Center for

Computing and

Communication

HPC @ RZ

Cluster

Filesystems

Software

Batch

IDEs

Visual Studio Teaser (1/3)

The screenshot displays the Microsoft Visual Studio IDE with three main components highlighted by blue boxes and labels:

- Code Editor:** Shows the source code for `gmres.cpp`. The code includes comments and function calls like `ctmrRoutine.Start()` and `ctmrRoutine.Stop()`. A callout box labeled "Code Editor" points to the main code area.
- Associated Code Definition:** Shows the definition for the `matrix_crs_builder` class in a template namespace. It includes a constructor that initializes `numrows`, `numcols`, and `numnonzeros`. A callout box labeled "Associated Code Definition" points to this window.
- Class Browser:** Shows a tree view of the `laperf` namespace, listing various classes and methods such as `getCols`, `getNumCols`, and `matrix_crs_builder`. A callout box labeled "Class Browser" points to this window.

Visual Studio Teaser (2/3)

The screenshot shows the Visual Studio IDE in debug mode. The main window displays the source code for `main.cpp`. The code implements a Jacobi stencil computation using OpenMP. A `#pragma omp for` directive is used to parallelize the inner loop over `j`. The code calculates the residual `fLRes` and updates the solution `U(j,i)`.

```

58     for (int i = 1; i < data.iCols - 1; i++)
59     {
60         UOLD(j,i) = U(j,i);
61     }
62 }
63
64 double fLRes;
65
66 /* compute stencil, residual and update */
67 #pragma omp for reduction(+:residual)
68 for (int j = data.iRowFirst + 1; j <= data.iRowLast - 1; j++)
69 {
70     for (int i = 1; i < data.iCols - 1; i++)
71     {
72         fLRes = ( ax * (UOLD(j, i-1) + UOLD(j, i+1))
73                 + ay * (UOLD(j-1, i) + UOLD(j+1, i))
74                 + b * UOLD(j, i) - F(j, i)) / b;
75
76         /* update solution */
77         U(j,i) = UOLD(j,i) - data.fRelax * fLRes;
78
79         /* accumulate residual error */
80         residual += fLRes * fLRes;
81     }
82 }
83 } /* end omp parallel */
84

```

The **Locals** window shows the following variables:

Name	Value	Type
afU	0x0013fd08	double *
residual	0.000000000000000000	double &
ax	999000.25000000023	double &
ay	999000.25000000023	double &
b	-3996001.80000000007	double &
aff	0x0013fdd0	double *

The **Threads** window shows three threads:

ID	Category	Name	Location	Priority	Suspend
7656	Main Thread	Main Thread	L_?Jacobi@@YAXAAUJacobiData@@@Z_53__par_region0_2_228	Normal	0
5044	Worker Thread	Win32 Thread	L_?Jacobi@@YAXAAUJacobiData@@@Z_53__par_region0_2_228	Normal	0
1688	Worker Thread	Win32 Thread	77225704	Highest	0

Locals

Name	Value	Type
afU	0x0013fd08	double *
residual	0.000000000000000000	double &
ax	999000.25000000023	double &
ay	999000.25000000023	double &
b	-3996001.80000000007	double &
aff	0x0013fdd0	double *

The screenshot shows the **Threads** and **Call Stack** windows. The **Threads** window shows three threads: a Main Thread (ID 7656) and two Worker Threads (IDs 5044 and 1688). The **Call Stack** window shows the current frame in `jacobi_omp.exe!L_?Jacobi@@YAXAAUJacobiData@@@Z_53__par_region0_2_228()` at line 77, with frames for `libiomp5md.dll!100011b5()` and `libiomp5md.dll!10007102()`.

Debugger w/ multi-threaded app



Visual Studio Teaser (3/3)

Microsoft's + Intel's Tools: Understand where the time is spent in your program, tune and parallelize it, check parallelization for correctness.

Function	Module	CPU Time
Jacobi	jacobi_omp.exe	25.489s
└─ Jacobi ← main ← _tmainCRTStartup ← mainCRTStartup	jacobi_omp.exe	25.489s
memset	jacobi_omp.exe	0.166s
InitializeMatrix	jacobi_omp.exe	0.150s
└─ InitializeMatrix ← main ← _tmainCRTStartup ← mainCRT	jacobi_omp.exe	0.150s
CheckError	jacobi_omp.exe	0.132s
_free_base	jacobi_omp.exe	0.010s
_write_nolock	jacobi_omp.exe	0.010s

Call Stack:

- jacobi_omp.exe!InitializeMatrix(struct JacobiData ...
- jacobi_omp.exe!InitializeMatrix(struct JacobiData ...
- jacobi_omp.exe!main - main.cpp:170
- jacobi_omp.exe!_tmainCRTStartup - crt0.c:266
- jacobi_omp.exe!mainCRTStartup - crt0.c:181
- kernel32.dll!BaseThreadInitThunk+0x11
- ntdll.dll!LdrInitializeThunk+0xec
- ntdll.dll!LdrInitializeThunk+0xbf

Summary:

- Elapsed Time: 26.298s
- CPU Time: 25.957s
- Logical CPU Count: 2



The End

Thank you for your attention!

Questions?



35