# MPI on Grid-Ireland

Brian Coghlan, John Walsh, Stephen Childs and Kathryn Cassidy
Trinity College Dublin

# Acknowledgements

- **Initial slides derived from slides by:**
  - Vered Kunik, Israeli Grid NA3 Team,

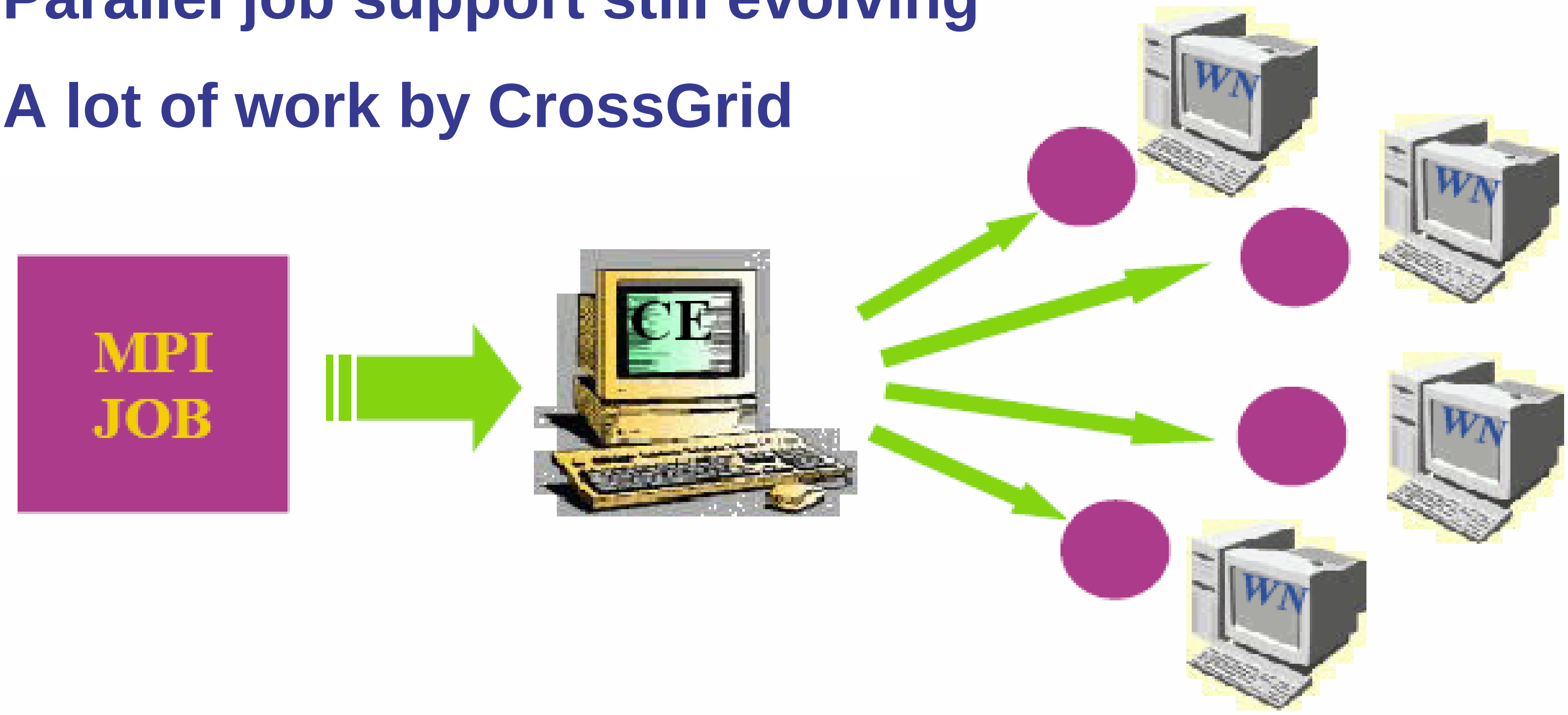
    for the Israeli Grid Workshop, Ra’anana, Israel, Sept. 2005
  - Miroslav Ruda, Masaryk University and CESNET,

    Grid for Complex Problems, Slovakia, 29 Nov. 2005
- **Extended by:**
  - Brian Coghlan, John Walsh, Stephen Childs and Kathryn Cassidy, TCD,

    for the Grid User’s Course, Trinity College Dublin, 14-15 March 2006

# Using MPI on the Grid

- The MPI job is run in parallel on several CPUs
- Libraries supported for parallel jobs: only MPICH so far
- Parallel job support still evolving
- A lot of work by CrossGrid



# Using MPI on the Grid

- **You can run your existing MPI applications with minimal modifications**
  - No need to change your MPI source code
  - Use wrapper script to compile and run your code
- **The Grid takes care of**
  - Finding suitable site to run your application
  - Running the application on multiple nodes

# Using MPI on Grid-Ireland

- **TCD, UCD (Rowan), DIAS (Leda) support MPICH:**

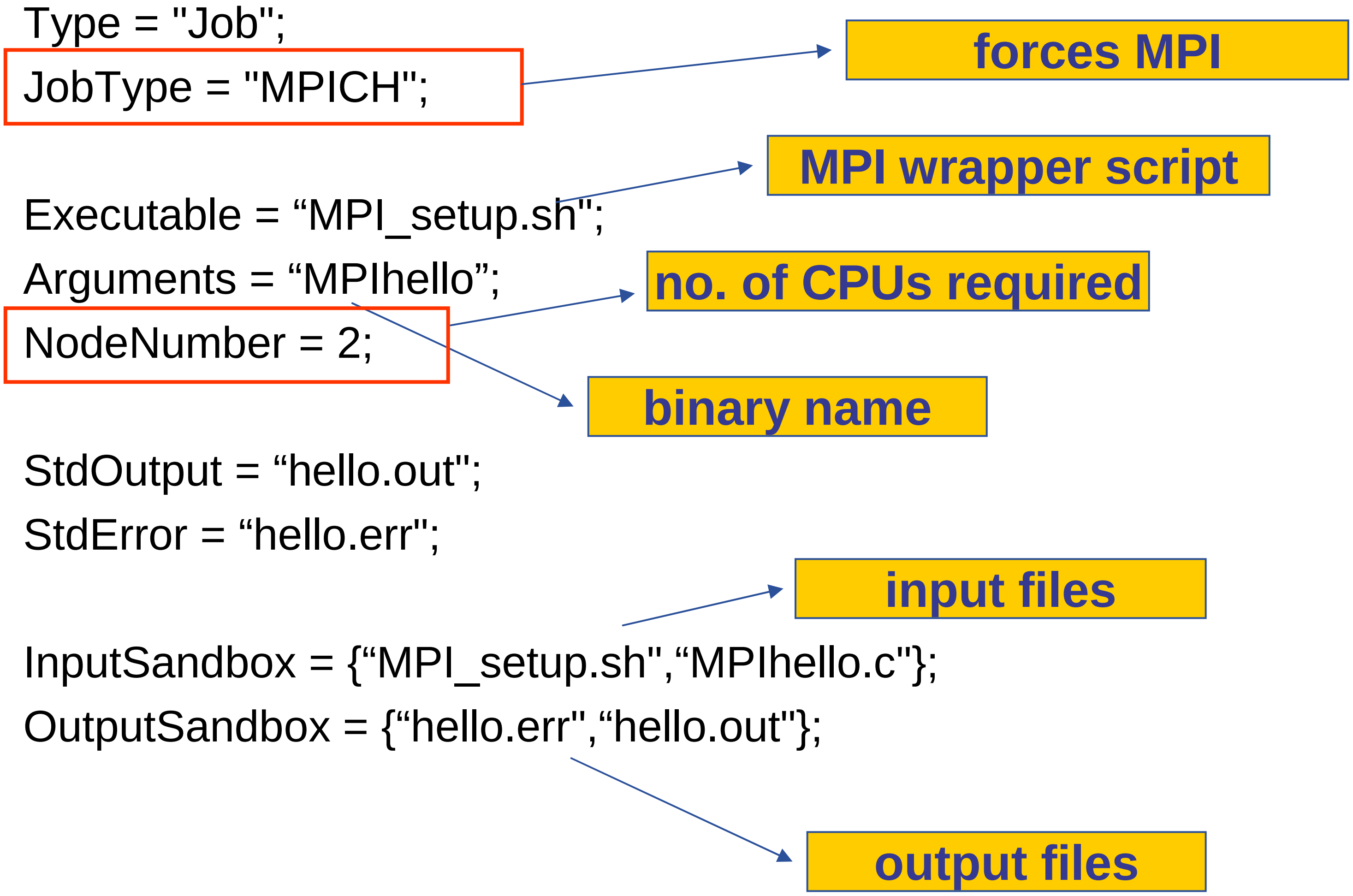
```
[childss@gridui example1]$ edg-job-list-match MPIhello.jdl

Selected Virtual Organisation name (from proxy certificate extension): cosmo
Connecting to host cagraidsvr18.cs.tcd.ie, port 7772

***************************************************************************
                         COMPUTING ELEMENT IDs LIST
 The following CE(s) matching your job requirements have been found:

                   *CEId*
 gridgate.cp.dias.ie:2119/jobmanager-lcgpbs-cosmo
 gridgate.cp.dias.ie:2119/jobmanager-lcgpbs-leda
 gridgate.cs.tcd.ie:2119/jobmanager-lcgpbs-cosmo
 gridgate.ucd.ie:2119/jobmanager-lcgpbs-rowan
***************************************************************************
```

# Using MPI on Grid-Ireland

- **Basic procedure:**
  - Write your code to use MPI
  - Set up appropriate JDL:
    - Specify JobType="MPICH"
    - Specify NodeNumber=<number_of_MPI_processes>
  - Write a wrapper script that:
    - (Optionally) compiles your code
    - Takes the filename of your executable as argument
    - Runs your executable using mpiexec

## Example: MPI "Hello world!"

# MPI Example 1: JDL



```
Type = "Job";
JobType = "MPICH";

Executable = “MPI_setup.sh";
Arguments = “MPIhello”;
NodeNumber = 2;

StdOutput = “hello.out";
StdError = “hello.err";

InputSandbox = {“MPI_setup.sh",“MPIhello.c"};
OutputSandbox = {“hello.err",“hello.out"};
```

forces MPI

MPI wrapper script

no. of CPUs required

binary name

input files

output files

# MPI Example 1: wrapper

- **Sample wrapper script**
  - compiles the application that was passed in as argument
  - then runs application using mpiexec

```
#!/bin/sh -x

# the binary to execute
EXE=$1

# compile the binary
mpicc -o ${EXE} ${EXE}.c

# run it using mpiexec
mpiexec `pwd`/$EXE
```

# MPI Example 1

- **Submit the MPI job to the Grid:**
  - edg-job-submit MPIhello.jdl
- **The Broker will automatically match the queue to the JDL**
  - JobType="MPICH"
    - Means that a MPI-capable queue will be chosen
- **The UI will automatically add the following to your JDL**
  - Member(other.RunTimeEnvironment, "MPICH");
    - Specifies that the queue's WNs have MPICH software installed
  - other.TotalCPUs >= NodeNumber;
    - Specifies the minimum number of CPUs on the queue
  - Rank = other.FreeCPUs;
    - Ranks the queues by number of free CPUs
    - Chooses queue with largest no. free CPUs matching all other requirements

# Limitations

- **Automatic site setup doesn’t yet work**
  - Site-specific MPI setup scripts aren’t yet automatically run
  - Special libraries might have to be set up in wrapper script
  - Working on a better solution to this problem

# MPI Example 2

- Write a wrapper script and JDL to submit the MPI cpi test program to calculate the value of pi

- Try this in the lab

# A real MPI example

- **Gareth Murphy of DIAS has a CFD application to model astrophysical jets flowing into molecular clouds**
  - Processes input files
  - Outputs a number of data files in HDF5 format
- **Consists of:**
  - a JDL file
  - a MPI wrapper script
  - a tgz file containing required libraries
  - a tgz file containing the executable source and data files

# A real MPI example

- **JDL file**
  - Specifies the MPI wrapper script as the executable
  - Specifies the library and code tarballs in the input sandbox
  - Specifies the tarred output files in the output sandbox

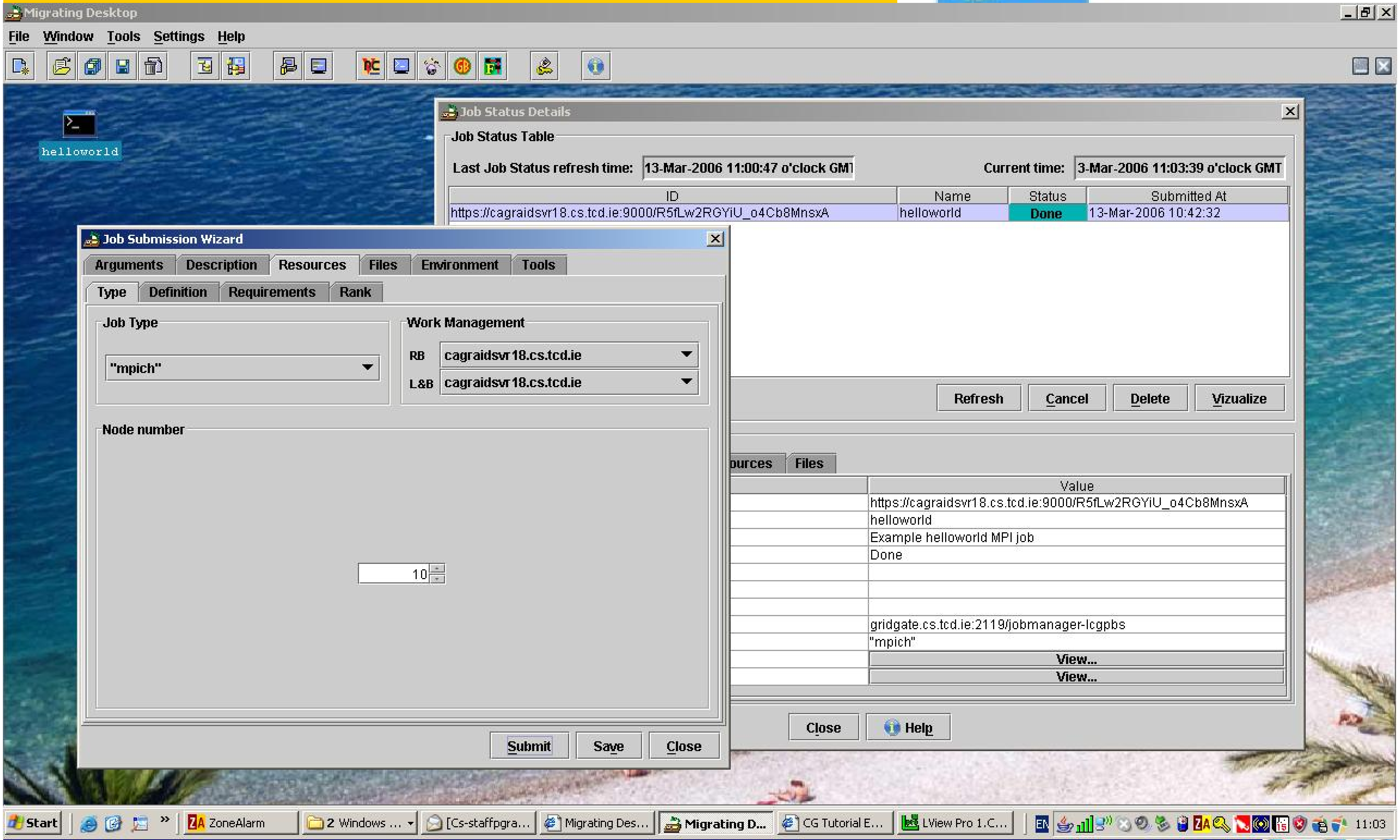
```
Type = "Job";
JobType = "MPICH";
NodeNumber = 10;
Executable = "mpi-application.sh";
StdOutput = "std.out";
StdError = "std.err";
InputSandbox = {"mpi-application.sh", "code.tgz",
"libraries.tgz"};
OutputSandbox = {"std.out","std.err", "mpi-output.tgz"};
Arguments = "";
RetryCount = 1;
```

# A real MPI example

- **MPI wrapper script**
  - Untars the libraries and code
  - Compiles the code
  - Runs the MPI executable
  - Tars the output files

```
#!/bin/bash
tar xzvf libraries.tgz
tar xzvf code.tgz
cp lib/* code/lib/
cd code/src/
make
cd ../bin/
export LD_LIBRARY_PATH="$LD_LIBRARY_PATH:$HOME/code/lib"
mpiexec ./mpi-executable
tar czvf ../../mpi-output.tgz outputfiles*
```

# Using MPI on the Grid



Job Status Table

Last Job Status refresh time: 13-Mar-2006 11:00:47 o'clock GMT

Current time: 3-Mar-2006 11:03:39 o'clock GMT

| ID | Name | Status | Submitted At |
|---|---|---|---|
| https://cagraidsvr18.cs.tcd.ie:9000/R5fLw2RGYiU_o4Cb8MnsxA | helloworld | Done | 13-Mar-2006 10:42:32 |

| | Value |
|---|---|
| | https://cagraidsvr18.cs.tcd.ie:9000/R5fLw2RGYiU_o4Cb8MnsxA |
| | helloworld |
| | Example helloworld MPI job |
| | Done |
| | |
| | |
| | gridgate.cs.tcd.ie:2119/jobmanager-lcgpbs |
| | "mpich" |
| | View... |
| | View... |

Job Type: "mpich"

Work Management

RB: cagraidsvr18.cs.tcd.ie

L&B: cagraidsvr18.cs.tcd.ie

Node number: 10

# Reality check

- **Middleware originally developed by and for high-energy physicists**
  - They don’t use MPI
  - So MPI support in Grid middleware has been neglected

- **Application areas now rapidly expanding**
  - Astrophysics, bio-medical, earth science users all want MPI
  - EGEE working group has been set up to improve support

- **We need your feedback!**
  - Try running your MPI jobs and let us know what is missing
  - We will feed this back into EGEE

# Using MPI on the Grid