Outline

- Production environment
  - Module environment
  - Job execution
- General porting of applications
  - Compilers and wrappers
  - Compiler/linker flags
- Porting and tuning – selected topics
  - Static and shared libraries
  - Mapping
  - MPMD
  - MPI extensions
  - QPX
## Module environment

<table>
<thead>
<tr>
<th>Modules – Features</th>
<th>Overview of available products</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Provides fast overview over all available products</td>
</tr>
<tr>
<td></td>
<td>• Provides fast overview over all available versions of a selected product</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Use of software and libraries</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Enables access to software packages and libraries</td>
</tr>
<tr>
<td>• Supports different versions of one product side by side</td>
</tr>
<tr>
<td>• Provides application-specific information</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dynamic modification of user environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Necessary environment variables are set appropriately, no need to set variables manually</td>
</tr>
<tr>
<td>• Detection of conflicts between applications</td>
</tr>
</tbody>
</table>
## Module environment

**Command**

```
module [options] [module]
```

**Selected options**

- `<no option>`
  - Lists all available options of the module command
- `avail`
  - Lists all available modules and versions
- `list`
  - Lists modules currently loaded
- `load/unload`
  - Loads/unloads a module
- `help`
  - Lists information about a module
- `show`
  - Displays information about environment settings applied by the module
- `purge`
  - Unloads all currently loaded modules
Module environment

Module categories on JUQUEEN

<table>
<thead>
<tr>
<th>Category</th>
<th>Software for compute nodes</th>
<th>Software for front-end nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMPILER</td>
<td>/bgsys/local</td>
<td>/usr/local</td>
</tr>
<tr>
<td>IO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MATH</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MISC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCIENTIFIC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOOLS</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Example

---

```
--------- /bgsys/local/modulefiles/TOOLS ---------
--------- /bgsys/local/modulefiles/SCIENTIFIC ---------
OpenFOAM/2.1.1-gcc(default)  cpmd/3.15.3
QuantumEspresso/5.0.1(default)  lammps/30Aug12
[...]
--------- /usr/local/modulefiles/COMPILER ---------
cmake/2.8.11(default)  python3/3.4.1(default)
--------- /usr/local/modulefiles/MATH ---------
[...]
```
# Module environment – available software

<table>
<thead>
<tr>
<th>Selected mathematical software</th>
<th>I/O libraries and tools</th>
<th>Science</th>
<th>Tools</th>
<th>Compiler (tools)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elemental (0.81)</td>
<td>cdo (1.6.1)</td>
<td>cp2k (2.4.0)</td>
<td>extrae (2.4.1)</td>
<td></td>
</tr>
<tr>
<td>FFTW (2.1.5/3.3.3)</td>
<td>darshan (2.2.8)</td>
<td>cpmd (3.15.3)</td>
<td>MUST (1.2.1)</td>
<td></td>
</tr>
<tr>
<td>GSL (1.15)</td>
<td>HDF5 (1.8.9)</td>
<td>lammmps (30AUG12)</td>
<td>Scalasca (2.1)</td>
<td></td>
</tr>
<tr>
<td>hypre (2.9.0)</td>
<td>Nco (4.3.4)</td>
<td>NAMD (2.9)</td>
<td>Scorep (1.3)</td>
<td></td>
</tr>
<tr>
<td>LAPACK (3.4.2)</td>
<td>netCDF (4.3)</td>
<td>OpenFOAM (2.1.1)</td>
<td>Tau (2.23.1)</td>
<td></td>
</tr>
<tr>
<td>MUMPS (4.10.0)</td>
<td>p-netCDF (1.3.1)</td>
<td>QE (5.0.1)</td>
<td>Totalview (8.14.0)</td>
<td></td>
</tr>
<tr>
<td>P3DFFT (2.5.2)</td>
<td></td>
<td></td>
<td>Vampir (8.4)</td>
<td></td>
</tr>
<tr>
<td>ParMETIS (4.0.2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PETSc (3.5.1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ScaLAPACK (2.0.2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPRNG (2.0)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sundials (2.5.0)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Not via modules
  - bgxlc (12.01)
  - bgxlf (14.01)
Job execution – Batch system

Execution of applications managed by LoadLeveler

- Users submit jobs using a job command file
- LoadLeveler allocates computing resources to run jobs
- The scheduling of jobs depends on
  - Availability of resources
  - Job priority (jobs with larger core counts are privileged)
- Jobs run in queues (job classes)
  - Chosen by LoadLeveler according to core count

More information about LoadLeveler on JUQUEEN

http://www.fz-juelich.de/ias/jsc/EN/Expertise/Supercomputers/JUQUEEN/UserInfo/LoadLeveler.html
**Job execution – Job command file**

ASCII file containing two major parts

<table>
<thead>
<tr>
<th>Job command file</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Keyword block</strong></td>
<td></td>
</tr>
<tr>
<td>• LoadLeveler keywords have the form</td>
<td></td>
</tr>
<tr>
<td>#@&lt;keyword&gt;</td>
<td></td>
</tr>
<tr>
<td>• # and @ can be separated by any number of blanks</td>
<td></td>
</tr>
<tr>
<td><strong>Shell script block</strong></td>
<td></td>
</tr>
<tr>
<td>• Regular shell script</td>
<td></td>
</tr>
<tr>
<td>• Can contain any shell command</td>
<td></td>
</tr>
</tbody>
</table>
### Job execution – LoadLeveler keywords

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>#@job_name=&lt;name&gt;</code></td>
<td>Name of the job</td>
</tr>
<tr>
<td><code>#@notification=</code></td>
<td>Send notification if the job is finished if the job returned an error code ≠ 0 never upon the start of the job combination of <code>start,end,error</code></td>
</tr>
<tr>
<td><code>end</code></td>
<td></td>
</tr>
<tr>
<td><code>error</code></td>
<td></td>
</tr>
<tr>
<td><code>never</code></td>
<td></td>
</tr>
<tr>
<td><code>start</code></td>
<td></td>
</tr>
<tr>
<td><code>always</code></td>
<td></td>
</tr>
<tr>
<td><code>#@notify_user=&lt;mailaddr&gt;</code></td>
<td>Mail address to send messages to</td>
</tr>
<tr>
<td><code>#@wall_clock_limit=hh:mm:ss</code></td>
<td>Requested wall time for the job</td>
</tr>
<tr>
<td><code>#@input=&lt;input file name&gt;</code></td>
<td>Specifies corresponding file names</td>
</tr>
<tr>
<td><code>#@output=&lt;file name for stdout&gt;</code></td>
<td></td>
</tr>
<tr>
<td><code>#@error=&lt;file name for stderr&gt;</code></td>
<td></td>
</tr>
<tr>
<td><code>#@environment=[&lt;variable&gt;,COPY_ALL]</code></td>
<td>Environment variable to be exported to job</td>
</tr>
<tr>
<td><code>#@queue</code></td>
<td>End of keyword section</td>
</tr>
</tbody>
</table>
# Job execution – LoadLeveler keywords

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>#@job_type=[serial, bluegene]</td>
<td>Specifies the type of job step to process. Must be set to bluegene for parallel applications.</td>
</tr>
<tr>
<td>#@bg_size=&lt;number of nodes&gt;</td>
<td>Size of the Blue Gene job, keywords bg_size and bg_shape are mutually exclusive.</td>
</tr>
<tr>
<td>#@bg_shape=&lt;A&gt;x&lt;B&gt;x&lt;C&gt;x&lt;D&gt;</td>
<td>Specifies the requested shape of a job (in midplanes).</td>
</tr>
<tr>
<td>#@bg_rotate=[True,False]</td>
<td>whether the scheduler should consider all possible rotations of the given shape</td>
</tr>
<tr>
<td>#@bg_connectivity=[TORUS,MESH,EITHER] Xa Xb Xc Xd</td>
<td>Type of wiring requested for the block (can be specified for each dimension separately)</td>
</tr>
</tbody>
</table>
Job execution – Shell script section

- Can contain any shell scripts commands
- Launching parallel applications:

**Command**

```
runjob [options]: <executable> [arguments]
```

**Selected options**

- `--envs <ENV_Var=Value>`
  
  Sets the environment variable `ENV_Var=Value`

- `--exp-env <ENV_Var>`
  
  Sets the environment variable `ENV_Var`

- `--np <number>`
  
  Total number of (MPI) tasks

- `--ranks-per-node <number>`
  
  Number of (MPI) tasks per compute node
Job execution – Example Job command file

```
#@job_name = MPI_code
#@comment = “32 ranks per node”
#@output = test_$(jobid).out
#@error = test_$(jobid).err
#@environment = COPY_ALL
#@job_type = bluegene
#@notification = never
#@bg_size = 512
#@bg_connectivity = torus
#@wall_clock_limit = 14:00:00
#@queue

runjob --np 16384 --ranks-per-node 32 : app.x
```

Pure MPI applications need to use 32 tasks per node in order to use the architecture efficiently, i.e., to fill all pipelines and registers and to avoid idle times!
Job execution – MPI/OpenMP hybrid jobs

```bash
# @job_name = hybrid_code
# @comment = "16x4 configuration"
# @output = test_$($jobid).out
# @error = test_$($jobid).err
# @environment = COPY_ALL
# @job_type = bluegene
# @notification = never
# @bg_size = 512
# @bg_connectivity = torus
# @wall_clock_limit = 14:00:00
# @queue

runjob --np 8192 --ranks-per-node 16\ 
     --env OMP_NUM_THREADS=4 : app.x -i input
```

Several process/thread configurations are possible:
- ntasks × nthreads = 64
- ntasks = 2^n, 0 ≤ n ≤ 6

Test which configuration gives the best performance for your application and setup!
# Job execution – Job classes

<table>
<thead>
<tr>
<th>Class name</th>
<th>#Nodes</th>
<th>Max. run time</th>
<th>Default run time</th>
</tr>
</thead>
<tbody>
<tr>
<td>n001</td>
<td>1 – 32</td>
<td>00:30:00</td>
<td>00:30:00</td>
</tr>
<tr>
<td>n002</td>
<td>33 – 64</td>
<td>00:30:00</td>
<td>00:30:00</td>
</tr>
<tr>
<td>n004</td>
<td>65 – 128</td>
<td>12:00:00</td>
<td>06:00:00</td>
</tr>
<tr>
<td>n008</td>
<td>129 – 256</td>
<td>12:00:00</td>
<td>06:00:00</td>
</tr>
<tr>
<td>m001</td>
<td>257 – 512</td>
<td>24:00:00</td>
<td>06:00:00</td>
</tr>
<tr>
<td>m002</td>
<td>513 – 1024</td>
<td>24:00:00</td>
<td>06:00:00</td>
</tr>
<tr>
<td>m004</td>
<td>1025 – 2048</td>
<td>24:00:00</td>
<td>06:00:00</td>
</tr>
<tr>
<td>m008</td>
<td>2049 – 4096</td>
<td>24:00:00</td>
<td>06:00:00</td>
</tr>
<tr>
<td>m016</td>
<td>4097 – 8192</td>
<td>24:00:00</td>
<td>06:00:00</td>
</tr>
<tr>
<td>m032*</td>
<td>8193 – 16384</td>
<td>24:00:00</td>
<td>06:00:00</td>
</tr>
<tr>
<td>m056*</td>
<td>24577 – 57344</td>
<td>24:00:00</td>
<td>06:00:00</td>
</tr>
</tbody>
</table>

* On demand only

You will be charged for full partitions (e.g. when requesting 513 nodes you will be charged for 1024 nodes!) → Always use full partitions!
# Job execution – Loadleveler commands

<table>
<thead>
<tr>
<th>Command</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>llssubmit &lt;jobfile&gt;</td>
<td>Sends job to the queuing system</td>
</tr>
<tr>
<td>llq</td>
<td>Lists all queued and running jobs</td>
</tr>
<tr>
<td>llsq -l &lt;job ID&gt;</td>
<td>Detailed information about the specified job</td>
</tr>
<tr>
<td>llsq -s &lt;job ID&gt;</td>
<td>Detailed information about a specific queued job, e.g. expected start time</td>
</tr>
<tr>
<td>llsq -u &lt;user&gt;</td>
<td>Lists all jobs of the specified user</td>
</tr>
<tr>
<td>llcancel &lt;job ID&gt;</td>
<td>Kills the specified job</td>
</tr>
<tr>
<td>llstatus</td>
<td>Displays the status of LoadLeveler</td>
</tr>
<tr>
<td>llclass</td>
<td>Lists existing classes and their properties</td>
</tr>
<tr>
<td>llqx</td>
<td>Shows detailed information about all jobs</td>
</tr>
</tbody>
</table>
Job execution – Monitoring of jobs

LoadLeveler

Command

```
llq [options]
```

Ilview

- Client-server based application
- compact summary of different information (e.g. current usage of system, job prediction, expected and average waiting times, ...)
- Customizable
- Developed by W. Frings (JSC)

Command

```
llview
```
Job execution – Ilview
Outline

- Production environment
  - Module environment
  - Job execution
- General porting of applications
  - Compilers and wrappers
  - Compiler/linker flags
- Porting and tuning – selected topics
  - Static and shared libraries
  - Mapping
  - MPMD
  - MPI extensions
  - QPX
Compilers and wrappers

- Different compilers for front-end and compute nodes
- GNU and IBM XL family of compilers available
- For C/C++ also Clang compiler for compute nodes available
  - Full support of C++ 11 standard and special BGQ hardware features like QPX and vector intrinsics

<table>
<thead>
<tr>
<th>Front- end nodes</th>
<th>Language</th>
<th>XL (thread-safe: add _r)</th>
<th>GNU</th>
<th>Clang</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C</td>
<td>xlc</td>
<td>gcc</td>
<td>---</td>
</tr>
<tr>
<td></td>
<td>C++</td>
<td>xlc++, xlc</td>
<td>g++</td>
<td>---</td>
</tr>
<tr>
<td></td>
<td>Fortran</td>
<td>xlf, xlf90, xlf95, xlf2003</td>
<td>gfortran</td>
<td>---</td>
</tr>
</tbody>
</table>
## Compilers and wrappers

<table>
<thead>
<tr>
<th>Compute nodes</th>
<th>Language</th>
<th>Compiler</th>
<th>MPI wrapper</th>
</tr>
</thead>
<tbody>
<tr>
<td>XL</td>
<td>C</td>
<td>bgxlc, bgc89, bgc99</td>
<td>mpixlc</td>
</tr>
<tr>
<td></td>
<td>C++</td>
<td>bgx1c++, bgx1c</td>
<td>mpixlcxx</td>
</tr>
<tr>
<td></td>
<td>Fortran</td>
<td>bgxlf, bgxlf90, bgxlf95,</td>
<td>mpixlf77, mpixlf90, mpixlf95, mpixlf2003</td>
</tr>
<tr>
<td>GNU</td>
<td>C</td>
<td>powerpc64-bq-linux-gcc</td>
<td>mpigcc</td>
</tr>
<tr>
<td></td>
<td>C++</td>
<td>powerpc64-bq-linux-g++</td>
<td>mpi++</td>
</tr>
<tr>
<td></td>
<td>Fortran</td>
<td>powerpc64-bq-linux-gfortran</td>
<td>mpifortran</td>
</tr>
<tr>
<td>Clang</td>
<td>C</td>
<td>bgclang</td>
<td>mpiclang</td>
</tr>
<tr>
<td></td>
<td>C++</td>
<td>bgclang++</td>
<td>mpiclang++</td>
</tr>
<tr>
<td></td>
<td>Fortran</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>

- If working with the XL suite of compilers, always use the thread-safe mpi wrappers when compiling for the compute nodes!

- If you use OpenMP tasks, the gcc suite might lead to executables with a noticeable better performance than the XL suite of compilers!
MPI setups

Six different MPI setups available on JUQUEEN

<table>
<thead>
<tr>
<th>&lt;name&gt;</th>
<th>Description</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>gcc</td>
<td>MPI library compiled with gcc, fine-grained locking, error checking and assertions enabled</td>
<td></td>
</tr>
<tr>
<td>gcc.legacy</td>
<td>MPI library compiled with gcc, coarse-grained locking, error checking and assertions enabled → slightly better latency for single-threaded codes</td>
<td>Initial porting</td>
</tr>
<tr>
<td>xl</td>
<td>MPI library compiled with XL compilers, PAMI compiled with gcc, error checking and assertions enabled (default wrappers)</td>
<td></td>
</tr>
<tr>
<td>xl.legacy</td>
<td>MPI library compiled with XL compilers, PAMI compiled with gcc, coarse-grained locking, error checking and assertions enabled</td>
<td></td>
</tr>
<tr>
<td>xl.ndebug</td>
<td>MPI library compiled with XL compilers, PAMI compiled with gcc, fine-grained locking, error checking and assertions disabled</td>
<td>Correctly running production code</td>
</tr>
<tr>
<td>xl.legacy.ndebug</td>
<td>MPI library compiled with XL compilers, PAMI compiled with gcc, coarse-grained locking, error checking and assertions disabled</td>
<td></td>
</tr>
</tbody>
</table>
## XL: basic compiler and linker options

<table>
<thead>
<tr>
<th>Optimization Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>-O2 -qarch=qp -qtune=qp</td>
<td>Basic optimization</td>
</tr>
<tr>
<td>-O3 -qstrict -qarch=qp -qtune=qp</td>
<td>More aggressive, not impact on acc.</td>
</tr>
<tr>
<td>-O3 -qhot -qarch=qp -qtune=qp</td>
<td>More aggressive, may influence acc. (high-order transformations of loops)</td>
</tr>
<tr>
<td>-O4 -qarch=qp -qtune=qp</td>
<td>Interprocedural optimization at compile time</td>
</tr>
<tr>
<td>-O5 -qarch=qp -qtune=qp</td>
<td>Interprocedural optimization at link time, whole program analysis</td>
</tr>
</tbody>
</table>

Some flags need to be used for compilation AND linking – Check the compiler manual. If you are not sure, include all flags used in the compile step in the linking step as well!
**XL: basic compiler and linker options**

<table>
<thead>
<tr>
<th>Compiler/linker flag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>-qsmp=omp -qthreaded</td>
<td>Switch on OpenMP support</td>
</tr>
<tr>
<td>-qreport -qlist</td>
<td>Generates for each source file <code>&lt;name&gt;</code> a file <code>&lt;name&gt;</code>.lst with pseudo code and a description of the kind of code optimizations which were performed</td>
</tr>
<tr>
<td>-qessl -lessl[smp]bg</td>
<td>Compiler attempts to replace some intrinsic FORTRAN 90 procedures by essl routines where it is safe to do so</td>
</tr>
</tbody>
</table>

**More information about using ESSL**

Example: compiler diagnostics

```
subroutine mult(c,a,ndim)

implicit none
integer :: ndim,i,j
double precision :: a(ndim),c(ndim,ndim)

! Loop
do i=1,1000
   do j=1,1000
      c(i,j) = a(i)
   enddo
enddo

end subroutine mult

>>>>> LOOP TRANSFORMATION SECTION <<<<<<

1| SUBROUTINE mult (c, a, ndim)

[...]

Id=1   DO $$CIV2 = $$CIV2,124
   10| IF (.FALSE.) GOTO lab_11
      $$LoopIV1 = 0
Id=2   DO $$LoopIV1 = $$LoopIV1,999

[...]

0 9 1   Loop interchanging applied to loop nest.
0 9 1   Outer loop has been unrolled 8 time(s).
```
## XL: further compiler and linker options

<table>
<thead>
<tr>
<th>Compiler/linker flag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>-qinline=auto:level=&lt;number&gt;</code></td>
<td>Function inlining</td>
</tr>
<tr>
<td><code>-qinline+procedure1[:procedure2[...]]</code></td>
<td>0 &lt;= &lt;number&gt; &lt;= 10</td>
</tr>
<tr>
<td>`-qunroll[={auto</td>
<td>yes}]`</td>
</tr>
<tr>
<td><code>-qipa[=&lt;suboptions_list&gt;]</code></td>
<td>Intra/inter-procedural optimization, see manual</td>
</tr>
<tr>
<td>`-qsimd[=auto</td>
<td>noauto]`</td>
</tr>
<tr>
<td></td>
<td><code>-qreport</code> <code>-qlist</code> !!)</td>
</tr>
</tbody>
</table>

**Warning:** Intra-procedural optimization (-O5 or `-qipa=2`) can cause very long compilation times, especially for C++ codes!
IBM XL compiler support for QPX

Usage of QPX

- Compiler flag `-qsimd=auto`
- Check that SIMD vectorization is actually done!
  - `-qreport`
  - `-qlist`

```plaintext
>>> LOOP TRANSFORMATION SECTION <<< [...]
-------------------------------------
0 9 1 Loop with nest-level 1 and iteration count 1000 was SIMD vectorized
[...]
```

```plaintext
>>> LOOP TRANSFORMATION SECTION <<< [...]
-------------------------------------
0 9 1 Loop was not SIMD vectorized because the loop is not the innermost loop.
0 10 1 Loop was not SIMD vectorized because it contains memory references with non-vectorizable alignment.
```
Outline

- Production environment
  - Module environment
  - Job execution
- General porting of applications
  - Compilers and wrappers
  - Compiler/linker flags
- Porting and tuning – selected topics
  - Static and shared libraries
  - Mapping
  - MPMD
  - MPI extensions
  - QPX
User Information and Support

Information about JUQUEEN

- JSC websites at http://www.fz-juelich.de/ias/jsc/EN/Expertise/Supercomputers/JUQUEEN/JUQUEEN_node.html

Dispatch and User Support

- Applications for accounts (for approved projects)
  Forschungszentrum Jülich GmbH, JSC, Dispatch, 52425 Jülich
  Tel: +49 2461 61 5642, Fax: +49 2461 61 2810
  email: dispatch.jsc@fz-juelich.de

- User Support
  Tel: +49 2461 61 2828
  email: sc@fz-juelich.de
Creating static libraries

Preferred kind of library on BG/Q

```
bgxlc -c pi.c
bgxlc -c main.c
#
### Create the library
ar rcs libpi.a pi.o
#
### Create the executable program
bgxlc -o pi main.o -L. -lpi
```
Creating shared libraries

```plaintext
bgxlc -c pi.c
bgxlc -c main.c
### Create the dynamic library
   bgxlc -qmkshrobj -Wl,-soname, libpi.so.0 \
   -o libpi.so.0.0 libpi.o
### Set up the soname
   ln -s libpi.so.0.0 libpi.so.0
### Create a linker name
   ln -s libpi.so.0 libpi.so
### Create the executable program
   bgxlc -o pi main.o -L -lpi -qnostaticlink \ 
   -qnostaticlink=libgcc
```

Shared libraries might become a bottleneck when using large core counts on Blue Gene systems! Try to avoid them!
Mapping

The placement of MPI ranks on the cores of a partition, i.e., the physical layout of the MPI ranks, is called mapping.

Network topology of JUQUEEN

- 5 dimensional hypercube or torus
- Each node is connected to 10 nearest neighbors in directions \( \pm A, \pm B, \pm C, \pm D, \pm E \)
- A 6\(^{th}\) dimension \( T \) is added to specify the hardware thread ID within one node (\( 0 < T <= N-1 \), where \( N \) is the number of MPI ranks per node)

<table>
<thead>
<tr>
<th>Node location in the torus</th>
<th>( &lt;A, B, C, D, E&gt; )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Process or thread location in the torus</td>
<td>( &lt;A, B, C, D, E, T&gt; )</td>
</tr>
</tbody>
</table>
Mapping – Receiving partition information

Command: runjob : /bgsys/local/samples/personality/personality.elf

Partition Information:
block shape : <2,2,2,2,2>
torus links enabled : <0,0,0,0,1>

IO Bridge Information:
rank 0000017 of 0000032 location <01,00,00,00,01> core 00 hthread 0 procid 00 ioBridge <01,00,00,00,01> cn ‘R63-M0-N06-J11’ ion ‘R63-ID-J03’
rank 0000023 of 0000032 location <01,00,01,01,01> core 00 hthread 0 procid 00 ioBridge <01,00,01,01,01> cn ‘R63-M0-N06-J06’ ion ‘R63-ID-J01’

Compute Node Information:
rank 0000023 location <01,00,01,01,01> core 00 hthread 0 procid = 00 ( Task 23 of 32 (1,0,1,1,0) R63-M0-N06-J06 )
rank 0000017 location <01,00,00,00,01> core 00 hthread 0 procid = 00 ( Task 17 of 32 (1,0,0,0,1,0) R63-M0-N06-J11 )
rank 0000022 location <01,00,01,00,00> core 00 hthread 0 procid = 00 ( Task 22 of 32 (1,0,1,1,0,0) R63-M0-N06-J01 )
rank 0000000 location <00,00,00,00,00> core 00 hthread 0 procid = 00 ( Task 0 of 32 (0,0,0,0,0,0) R63-M0-N06-J12 )
rank 0000016 location <01,00,01,01,00> core 00 hthread 0 procid = 00 ( Task 16 of 32 (1,0,0,0,0,0) R63-M0-N06-J17 )
rank 0000007 location <00,00,00,00,00> core 00 hthread 0 procid = 00 ( Task 7 of 32 (0,0,1,1,1,0) R63-M0-N06-J27 )
rank 0000021 location <01,00,00,01,01> core 00 hthread 0 procid = 00 ( Task 21 of 32 (1,0,1,0,1,0) R63-M0-N06-J10 )
rank 0000013 location <00,01,01,00,01> core 00 hthread 0 procid = 00 ( Task 13 of 32 (0,1,1,0,1,0) R63-M0-N06-J20 )
rank 0000025 location <01,01,00,00,01> core 00 hthread 0 procid = 00 ( Task 25 of 32 (1,1,0,0,1,0) R63-M0-N06-J08 )
rank 0000001 location <00,00,00,00,00> core 00 hthread 0 procid = 00 ( Task 1 of 32 (0,0,0,0,0,1) R63-M0-N06-J22 )
rank 0000027 location <01,01,00,01,00> core 00 hthread 0 procid = 00 ( Task 27 of 32 (1,1,0,1,1,0) R63-M0-N06-J04 )
rank 0000012 location <00,01,00,00,00> core 00 hthread 0 procid = 00 ( Task 12 of 32 (0,1,1,0,0,0) R63-M0-N06-J19 )
rank 0000019 location <01,00,00,01,01> core 00 hthread 0 procid = 00 ( Task 19 of 32 (1,0,1,1,0,0) R63-M0-N06-J07 )
rank 0000005 location <00,00,00,00,00> core 00 hthread 0 procid = 00 ( Task 5 of 32 (0,0,1,0,1,0) R63-M0-N06-J23 )
rank 0000029 location <01,01,01,00,01> core 00 hthread 0 procid = 00 ( Task 29 of 32 (1,1,1,0,1,0) R63-M0-N06-J09 )
rank 0000010 location <00,01,00,01,00> core 00 hthread 0 procid = 00 ( Task 10 of 32 (0,1,0,1,0,0) R63-M0-N06-J30 )
rank 0000031 location <01,01,01,01,01> core 00 hthread 0 procid = 00 ( Task 31 of 32 (1,1,1,1,1,0) R63-M0-N06-J05 )
Mapping

Changing the Mapping

Pre-defined mappings

Command

```
runjob --mapping <permutation of ABCDET>
```

The rightmost coordinate is varied fastest. If no mapping is specified, the default is ABCDET, i.e., the processes are mapped first within one node before occupying the next node, etc.

User-defined mappings

Command

```
runjob --mapping <mapfile>
```

<mapfile> is an ASCII file, where the \( n \)th line specifies the 6 coordinates of the MPI rank \( n \):

```
4 0 0 0 0 0  # task 0
0 0 0 2 0 0  # task 1
0 0 1 0 0 0  # task 2
[...]  
```
Mapping

Hints for choosing an optimal mapping

- Can increase performance considerably
- Depends on the communication pattern of applications
- Ranks communicating with each other should be close
- Folding of dimensions according to decomposition

<table>
<thead>
<tr>
<th>Job size: 1 midplane with 16 MPI ranks/node</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mapping ABCDET</td>
</tr>
<tr>
<td>Dimensions: 4x4x4x4x2x16</td>
</tr>
<tr>
<td>- Good for simulations with, e.g.,</td>
</tr>
<tr>
<td>following 2D decompositions:</td>
</tr>
<tr>
<td>256x32 = (4^4) x (2x16)</td>
</tr>
<tr>
<td>64x128 = (4^3) x (4x2x16)</td>
</tr>
</tbody>
</table>
MPMD – Multiple Program Multiple Data

Instead of using only one executables several different executables can be used at the same time sharing one JUQUEEN partition
# MPMD job execution – Job command file

<table>
<thead>
<tr>
<th>Example</th>
<th>MPMD setup</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>app1.x: 4 MPI processes</td>
</tr>
<tr>
<td></td>
<td>app2.x: 4 MPI processes</td>
</tr>
<tr>
<td></td>
<td>app3.x: 8 MPI processes</td>
</tr>
</tbody>
</table>

**Job command file**

```plaintext
# @job_name = MPMD
# @comment = "MPMD Example"
# @environment = COPY_ALL
# @job_type = bluegene
# @notification = never
# @bg_size = 32
# @bg_connectivity = torus
# @wall_clock_limit = 00:30:00
# @queue

runjob -n 16 -p 4 -mapping mpmd_mapfile : dummy.x
```
# Mapfile for MPMD – Section I

<table>
<thead>
<tr>
<th>Example</th>
<th>MPMD setup</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>app1.x: 4 MPI processes</td>
</tr>
<tr>
<td></td>
<td>app2.x: 4 MPI processes</td>
</tr>
<tr>
<td></td>
<td>app3.x: 8 MPI processes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>mpmd_mapfile (ASCII file) – Section I</th>
</tr>
</thead>
<tbody>
<tr>
<td>#mpmdbegin 0-3</td>
</tr>
<tr>
<td>#mpmdcmd app1.x</td>
</tr>
<tr>
<td>#mpmdend</td>
</tr>
<tr>
<td>#mpmdbegin 4-7</td>
</tr>
<tr>
<td>#mpmdcmd app2.x</td>
</tr>
<tr>
<td>#mpmdend</td>
</tr>
<tr>
<td>#mpmdbegin 8-15</td>
</tr>
<tr>
<td>#mpmdcmd app3.x</td>
</tr>
<tr>
<td>#mpmdend</td>
</tr>
</tbody>
</table>

- app1.x with rank 0 to 3
- app2.x with rank 4 to 7
- app3.x with rank 8 to 15
## Mapfile for MPMD – Section II

<table>
<thead>
<tr>
<th>MPMD setup</th>
<th>Example</th>
<th>mpmd_mapfile (ASCII file) – Section II positions in the 6D torus (ABCDE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>app1.x: 4 MPI processes</td>
<td></td>
<td>app1.x with rank 0 to 3, first node, core 0 to 3</td>
</tr>
<tr>
<td>app2.x: 4 MPI processes</td>
<td></td>
<td>app2.x with rank 4 to 7, second node in E, core 0 to 3</td>
</tr>
<tr>
<td>app3.x: 8 MPI processes</td>
<td></td>
<td>app3.x with rank 8 to 15, first node in D and C, cores 0 to 3 on each node</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Example</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0 0 0 0 0</td>
<td></td>
</tr>
<tr>
<td>0 0 0 0 0 1</td>
<td></td>
</tr>
<tr>
<td>0 0 0 0 0 2</td>
<td></td>
</tr>
<tr>
<td>0 0 0 0 0 3</td>
<td></td>
</tr>
<tr>
<td>0 0 0 0 1 0</td>
<td></td>
</tr>
<tr>
<td>0 0 0 0 1 1</td>
<td></td>
</tr>
<tr>
<td>0 0 0 0 1 2</td>
<td></td>
</tr>
<tr>
<td>0 0 0 0 1 3</td>
<td></td>
</tr>
<tr>
<td>0 0 0 1 0 0</td>
<td></td>
</tr>
<tr>
<td>0 0 0 1 0 1</td>
<td></td>
</tr>
<tr>
<td>0 0 0 1 0 2</td>
<td></td>
</tr>
<tr>
<td>0 0 0 1 0 3</td>
<td></td>
</tr>
<tr>
<td>0 0 1 0 0 0</td>
<td></td>
</tr>
<tr>
<td>0 0 1 0 0 1</td>
<td></td>
</tr>
<tr>
<td>0 0 1 0 0 2</td>
<td></td>
</tr>
<tr>
<td>0 0 1 0 0 3</td>
<td></td>
</tr>
</tbody>
</table>
MPMD – important hints

- On one node only one executable can run.
- The number of ranks specified in the `runjob` command (`-p` or `--ranks-per-node`) must be the largest number of ranks per node needed.
- The memory available to an MPI task depends on the number of ranks per node specified in the `runjob` command (`-p 8` → 2 GB per rank, regardless how many ranks per node are actually started).
- The executable specified in the `runjob` command is a dummy argument.

Obtaining information about the allocated partition and the distribution of processes

http://www.fz-juelich.de/ias/jsc/EN/Expertise/Supercomputers/JUQUEEN/UserInfo/ShapeMapping.html#doc1413730bodyText4
MPI extensions for Blue Gene/Q

IBM offers extensions to the MPI standard for Blue Gene/Q

- *Not* part of the official MPI standard!
- C (and Fortran77 interface for most functions)
- Functions start with `MPIX_` instead of `MPI_`

Overview over all available extensions

http://www.fz-juelich.de/ias/jsc/EN/Expertise/Supercomputers/JUQUEEN/UserInfo/MPIextensions.html

<table>
<thead>
<tr>
<th>C</th>
<th>#include &lt;mpix.h&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fortran</td>
<td>include <code>mpif.h</code></td>
</tr>
</tbody>
</table>
MPI extensions for Blue Gene/Q

```c
typedef struct {
    unsigned prank;                        // Physical rank of node
    unsigned psize;                        // Size of partition
    unsigned ppn;                          // Processes per node
    unsigned coreID;                       // Process ID
    unsigned clockMHz;                     // Frequency in MHz
    unsigned memSize;                      // Memory in MB
    unsigned torus_dimension;              // Actual torus dimension
    unsigned Size[MPIX_TORUS_MAX_DIMS];    // Max. torus dimensions
    unsigned Coords[MPIX_TORUS_MAX_DIMS];  // Node’s coordinated
    unsigned isTorus[MPIX_TORUS_MAX_DIMS]; // Wrap-around dims?
    unsigned rankInPset;                   // Zero on Blue Gene/Q
    unsigned sizeOfPset;                   // Zero on Blue Gene/Q
    unsigned idOfPset;                     // Zero on Blue Gene/Q
} MPIX_Hardware_t;
```
Blue Gene/Q: I/O Node Cabling (8 ION/Rack)
MPI BGQ Extensions – I/O related functions

- Collective call on `MPI_COMM_WORLD`
- Returns a communicator which contains only MPI ranks which run on nodes belonging to different/same I/O Bridge Nodes
- The name of this function is chosen for backwards compatibility, since there are no psets on the Blue Gene/Q anymore

### C

```c
int MPIX_Pset_diff_comm_create(MPI_Comm *pset_comm_diff)
int MPIX_Pset_same_comm_create(MPI_Comm *pset_comm_same)
```

### Fortran

```fortran
MPIX_PSET_DIFF_COMM_CREATE(PSET_COMM_DIFF, IERROR)
MPIX_PSET_SAME_COMM_CREATE(PSET_COMM_SAME, IERROR)
INTEGER PSET_COMM_DIFF, PSET_COMM_SAME, IERROR
```
MPI BGQ Extensions – I/O related functions

- Returns the ID of the associated I/O node

C

```c
int MPIX_IO_node_id()
```

Fortran

```fortran
MPIX_IO_NODE_ID(IO_NODE_ID)
INTEGER :: IO_NODE_ID
```

- NO IERROR parameter for Fortran
MPI BGQ Extensions – I/O related functions

- Returns the distance to the associated I/O node in number of hops

```
C
int MPIX_IO_distance()

Fortran
MPIX_IO_DISTANCE(IO_DISTANCE)
INTEGER :: IO_DISTANCE
```

**NO IERROR parameter for Fortran**
MPI BGQ Extensions – I/O related functions

- Returns the ID of the link to the associated I/O node

```c
int MPIX_IO_link_id();
```

```fortran
MPIX_IO_LINK_ID(IO_LINK_ID)
INTEGER :: IO_LINK_ID
```

NO IERROR parameter for Fortran
Quad-processing extension (QPX)

- Allows to perform 8 concurrent floating point operations (fused multiply-add, FMA) + load + store

- Four double precision pipelines, usable as
  - Scalar FPU
  - 4-wide double SIMD
  - 2-wide complex SIMD
IBM XL compiler support for QPX

Usage of QPX

- Compiler flag `-qsimd=auto`
- Check that SIMD vectorization is actually done!
  - `-qreport`
  - `-qlist`

```plaintext
>>> LOOP TRANSFORMATION SECTION <<<
[...]
-------------------------------
0 9 1  Loop with nest-level 1 and iteration count 1000 was SIMD vectorized
[...]
```

```plaintext
>>> LOOP TRANSFORMATION SECTION <<<
[...]
-------------------------------
0 9 1  Loop was not SIMD vectorized because the loop is not the innermost loop.
0 10 1  Loop was not SIMD vectorized because it contains memory references with non-vectorizable alignment.
```
IBM XL compiler support for QPX

Provide hints for the compiler

- Point to compiler likely iteration counts
- Instruct compiler to align fields
- Tell that FORTRAN assumed-shape arrays are contiguous

```
FORTRAN
real*8 :: x(:), y(:), a
! ibm* align(32, x, y)
! ibm* assert(itercnt(100))

do i=m, n
   z(i) = x(i) + a*y(i)
enddo

C/C++
double __align(32) *x, *y;
double a;
#pragma disjoint(*x, *y)
#pragma disjoint(*x, a)
#pragma ibm iterations(100)
for (int i=m; i<n; i++)
   z[i] = x[i] + a*y[i]
void foo(double* restrict a1,
         double* restrict a2){
   for (int i=0; i<n; i++)
      a1[i]=a2[i];}
```
IBM XL compiler QPX intrinsics

<table>
<thead>
<tr>
<th>XL intrinsics</th>
<th>New intrinsic variable type</th>
</tr>
</thead>
<tbody>
<tr>
<td>C/C++</td>
<td>vector4double</td>
</tr>
<tr>
<td>Fortran</td>
<td>vector(real(8))</td>
</tr>
</tbody>
</table>

Wide set of elemental functions available

- LOAD, STORE, MULT, MULT-ADD, ROUND, CEILING, SQRT, …

**Strengths:**

- User may layout calculation by hand, if compiler not smart enough

**Easy to use:**

- Leave stack, register layout, load/store scheduling to compiler
QPX example using compiler intrinsics

```c
typedef vector4double qv;
qv dx, dy, dz, dx2, dy2, dz2
for (i=0; i<4; i++)
{
    xd[i] = xdipl[j];
    yd[i] = ydipl[j];
    zd[i] = zdipl[j];
}
dx2 = vec_mul(dx, dx);
dy2 = vec_mul(dy, dy);
dz2 = vec_mul(dz, dz);
d = vec_swsqrt(dx2 + dy2 + dz2);
...```

Source: IBM Corporation