NPB3.3-MZ-MPI/BT tutorial example OpenMP+MPI application (generic cluster version)

Brian Wylie
Jülich Supercomputing Centre
b.wylie@fz-juelich.de

September 2011

Tutorial exercise objectives

- Familiarise with usage of VI-HPS tools
  - complementary tools' capabilities & interoperatibilty
- Prepare to apply tools productively to *your* application(s)
- Exercise is based on a small portable benchmark code
  - unlikely to have significant optimization opportunities

- Optional (recommended) exercise extensions
  - analyze performance of alternative configurations
  - investigate effectiveness of system-specific compiler/MPI optimizations and/or placement/binding/affinity capabilities
  - investigate scalability and analyze scalability limiters
  - compare performance on different HPC platforms
  - ...

Local installation (UNITE)

- Load the UNITE module and then the tool of interest

```
% module load UNITE
% module avail
kcachegrind/1.4         marmot/2.4.0              paraver/4.0
periscope/1.4b          scalasca/1.3.3            tau/2.20.3
vampir/7.3.0            vampirserver/2.3.0        vampirtrace/5.11
% module load scalasca/1.3.3
```

- Often distinct modules for each MPI/compiler combination
  - MPI libraries typically not binary compatible
  - OpenMP run-time libraries not binary compatible
  - Compiler-based instrumentation is compiler-specific
    - measurement collection, analysis & examination shouldn't be
- Tutorial sources should be copied to your own directory where you can then work with them

```
% cp -r ~hpclab01/tutorial/NPB3.3-MZ-MPI $WORK
```

NPB-MZ-MPI suite

- NAS Parallel Benchmark suite (sample MZ-MPI version)
  - Available from `http://www.nas.nasa.gov/Software/NPB`
  - 3 benchmarks (all in Fortran77, using OpenMP+MPI)
  - Configurable for various sizes & classes
- Move into the NPB3.3-MZ-MPI root directory

```
% cd NPB3.3-MZ-MPI; ls
BT-MZ/  LU-MZ/  SP-MZ/
bin/    common/ config/ jobscript/      Makefile        README  sys/
```

- Subdirectories contain source code for each benchmark
  - plus additional configuration and common code
- The provided distribution has already been configured for the tutorial, such that it's ready to “make” benchmarks and install them into a (tool-specific) “bin” subdirectory

Building an NPB-MZ-MPI benchmark

- Type “make” for instructions

```
% make
  ========================================
  =      NAS Parallel Benchmarks 3.3     =
  =      MPI+OpenMP Multi-Zone versions  =
  ========================================

  To make a NAS multi-zone benchmark type

        make <benchmark-name> CLASS=<class> NPROCS=<number>

  To make a set of benchmarks, create the file config/suite.def
  according to the instructions in config/suite.def.template and type

        make suite

 ************************************************************
 * Custom build configuration is specified in config/make.def  *
 * Suggested tutorial benchmark specification:                 *
 *       make bt-mz CLASS=B NPROCS=4                           *
 ************************************************************
```

Building the NPB-MZ-MPI/BT benchmark

- Specify the benchmark configuration
  - benchmark name: **bt-mz**, lu-mz, sp-mz
  - the number of MPI processes: NPROCS=**4**
  - the benchmark class (S, W, A, B, C, D, E, F): CLASS=**B**

```
% make bt-mz CLASS=B NPROCS=4
cd BT-MZ; make CLASS=B NPROCS=4 VERSION=
gmake: Entering directory 'BT-MZ'
cd ../sys; cc  -o setparams setparams.c -lm
../sys/setparams bt-mz 4 B
mpif77 -c  -O3 -openmp  bt.f
...
mpif77 -c  -O3 -openmp  setup_mpi.f
cd ../common;  mpif77 -c  -O3 -openmp  print_results.f
cd ../common;  mpif77 -c  -O3 -openmp  timers.f
mpif77 -O3 -openmp  -o ../bin/bt-mz_B.4 \
    bt.o make_set.o initialize.o exact_solution.o exact_rhs.o \
    set_constants.o adi.o define.o copy_faces.o rhs.o solve_subs.o \
    x_solve.o y_solve.o z_solve.o add.o error.o verify.o setup_mpi.o \
    ../common/print_results.o ../common/timers.o
Built executable ../bin/bt-mz_B.4
gmake: Leaving directory 'BT-MZ'
```

NPB-MZ-MPI/BT (Block Tridiagonal solver)

- What does it do?
  - Solves a discretized version of unsteady, compressible Navier-Stokes equations in three spatial dimensions
  - Performs 200 time-steps on a regular 3-dimensional grid using ADI and verifies solution error within acceptable limit
  - Intra-zone computation with OpenMP, inter-zone with MPI
- Implemented in 20 or so Fortran77 source modules

- Runs with any number of MPI processes & OpenMP threads
  - bt-mz_B.4 x4 is reasonable (increase to 4x6 as appropriate)
    - excess processes idle when run with more than compiled
  - bt-mz_B.4 x4 should run in around 30 seconds
    - typically runs more efficiently with more processes than threads
  - CLASS=C does much more work and takes much longer!

NPB-MZ-MPI/BT reference execution

- Set OMP_NUM_THREADS and launch as an MPI application

```
% cd bin;  OMP_NUM_THREADS=4  mpiexec -np 4 ./bt-mz_B.4
 NAS Parallel Benchmarks (NPB3.3-MZ-MPI) - BT-MZ MPI+OpenMP Benchmark
 Number of zones:    8 x   8
 Iterations:  200    dt:   0.000300
 Number of active processes:      4

 Time step    1
 Time step   20
 Time step   40
 Time step   60
 Time step   80
 Time step  100
 Time step  120
 Time step  140
 Time step  160
 Time step  180
 Time step  200
 Verification Successful

 BT-MZ Benchmark Completed.
 Time in seconds = 28.86
```

Hint: copy/edit example batch scripts from jobscript directory:
% qsub ../jobscript/run.pbs

Hint: save the benchmark output (or note the run time) to be able to refer to it later

Tutorial exercise steps

- The tutorial steps are similar and repeated for each tool
- Use the provided NPB3.3-MZ-MPI tutorial directory

```
% cd NPB3.3-MZ-MPI; ls
BT-MZ/  LU-MZ/  SP-MZ/
bin/    common/ config/ jobscript/      Makefile        README  sys/
```

- Edit config/make.def to adjust build configuration
  - Modify specification of compiler/linker: MPIF77
- Make clean and build new tool-specific executable

```
% make clean
% make bt-mz CLASS=B NPROCS=4
Built executable ../bin.$(TOOL)/bt-mz_B.4
```

- Change to the directory containing the new executable before running it with the desired tool configuration

```
% cd bin.$(TOOL)
% export OMP_NUM_THREADS=4 ...
% mpiexec -np 4 ./bt-mz_B.4
```

NPB-MZ-MPI/BT build configuration definition

- config/make.def

```
#            SITE- AND/OR PLATFORM-SPECIFIC DEFINITIONS
#-----------------------------------------------------------------------
# Items in this file may need to be changed for each platform.
…
#OPENMP = -fopenmp  # GCC
OPENMP  = -openmp   # Intel
#-----------------------------------------------------------------------
# The Fortran compiler used for hybrid MPI programs
#-----------------------------------------------------------------------
MPIF77 = mpif77

# Alternative variants to perform instrumentation
#MPIF77 = marmotf77
#MPIF77 = psc_instrument mpif77
#MPIF77 = scalasca -instrument mpif77
#MPIF77 = tau_f90.sh
#MPIF77 = vtf77 -vt:hyb -vt:f77 mpif77

# PREP is a generic preposition macro for instrumentation preparation
#MPIF77 = $(PREP) mpif77

...
```

Set flag according to compiler

Default (no instrumentation)

Hint: uncomment one of these alternative compiler wrappers to perform instrumentation ...

… or this for generic variant

Workshop systems (hardware)

System	***juropa***	***cluster-beta***	***jugene***
Domain	fz-juelich.de	rz.rwth-aachen.de	fz-juelich.de
Vendor	Sun/Bull	Bull	IBM
Type	Constellation	Bullx	BlueGene/P
Network	Infiniband	Infiniband	BlueGene/P
Processors	Xeon X5570	Xeon X5675 / X7550	PowerPC 450
Frequency	2930 MHz	3060 / 2000 MHz	850 MHz
Compute nodes	3288	1358 / 346	2304
Chips per node	2	2 / 4	32
Cores per chip	4	6 / 8	4
Threads per core	2	1 / 1	1
Memory per node	24 GB	24-96 / 64-256 GB	2 GB

Workshop systems (software environment)

System	***juropa***	***cluster-beta***	***jugene***
domain	fz-juelich.de	rz.rwth-aachen.de	fz-juelich.de
Filesystem	*Lustre*	*Lustre*	*GPFS*
Parallel filesys	$WORK	/hpcwork/$USER	$WORK
Compiler	*Intel*	*Intel*	*IBM XL*
OpenMP	-openmp	-openmp	-qsmp=omp
MPI	*ParaStation*	*IntelMPI / OpenMPI*	*BG MPICH*
C compiler	mpicc	mpiicc / mpicc	mpixlc_r
C++ compiler	mpicxx	mpiicpc / mpicxx	mpixlcxx_r
F9x compiler	mpif9x	mpiifort / mpif90	mpixlf9x_r
Queue	*Moab/PBS*	*LSF*	*LoadLeveler*
job submit	msub job	bsub < job	llsubmit job
list jobs	qstat -u	bjobs	llq -u $USER