Science & Technology Facilities Council

Code Saturne on POWER8 clusters: First Investigations

C. MOULINEC, V. SZEREMI, D.R. EMERSON (STFC Daresbury Lab., UK)
Y. FOURNIER (EDF R&D, FR)
P. VEZOLLE, L. ENAULT (IBM Montpellier, FR)
B. ANLAUF, M. BUEHLER (IBM Boeblingen, GE)

CODE SATURNE

Contents

Science & Technology Facilities Council

Code_Saturne

Technology

-Co-located finite volume, arbitrary unstructured meshes, predictor-corrector
-350 000 lines of code, 37% Fortran, 50% C, 13% Python
-MPI for distributed-memory and some OpenMP for shared-memory machines

Physical modeling

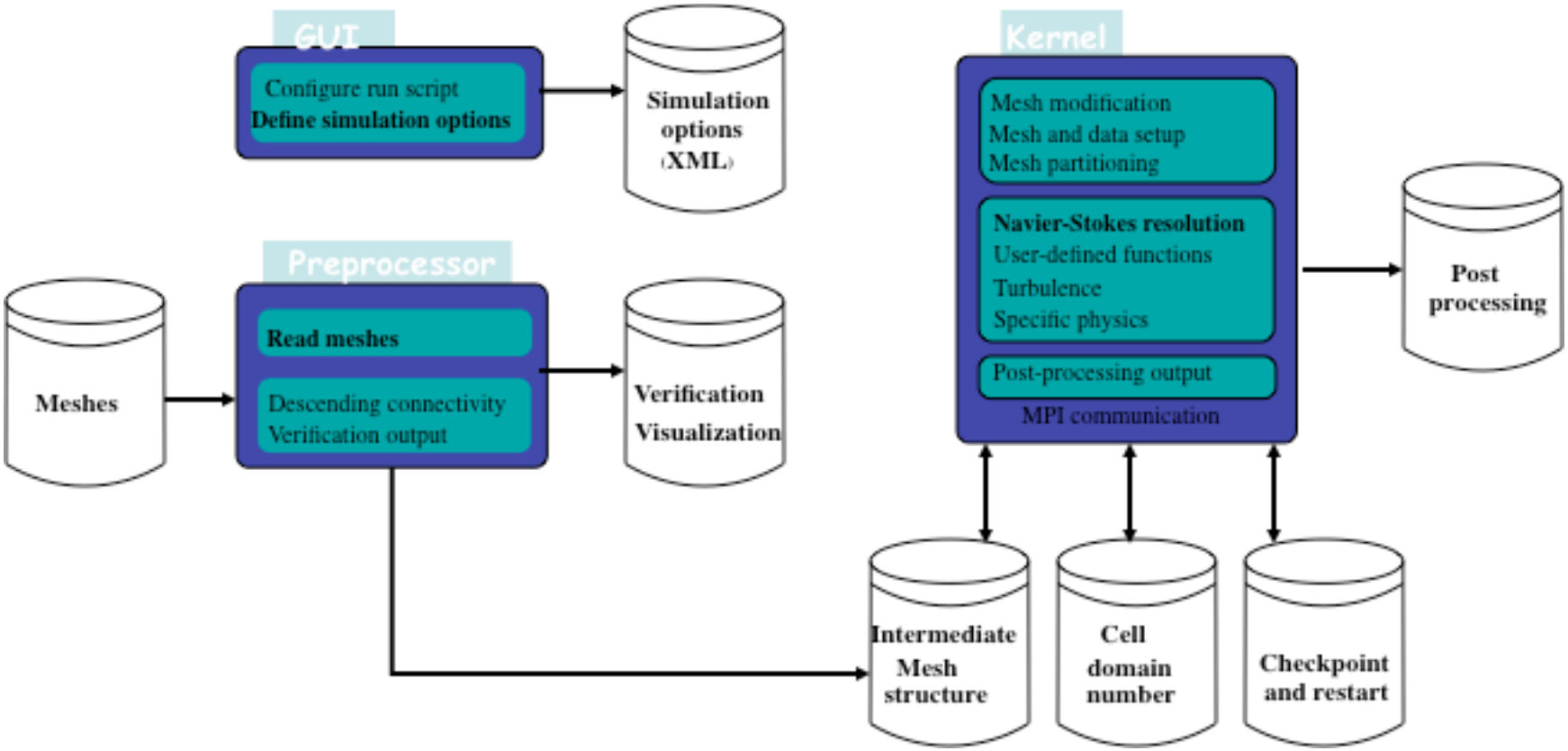
-Laminar and turbulent flows: k-epsilon, k-omega, SST, v2f, RSM, LES models
-Radiative transfer
-Coal, heavy-fuel and gas combustion
-Electric arcs and Joule effect
-Lagrangian module for particles tracking
-Atmospheric modeling
-ALE method for deformable meshes
-Rotor / stator interaction for pump modeling, for marine turbines

Flexibility

-Portability (Unix, Linux, MacOS X and now Windows)
-Graphical User Interface with possible integration within the SALOME platform

Reduced number of tools

- Each with rich functionality
- Natural separation between interactive and potentially long-running parts
- In-line (pdf) documentation

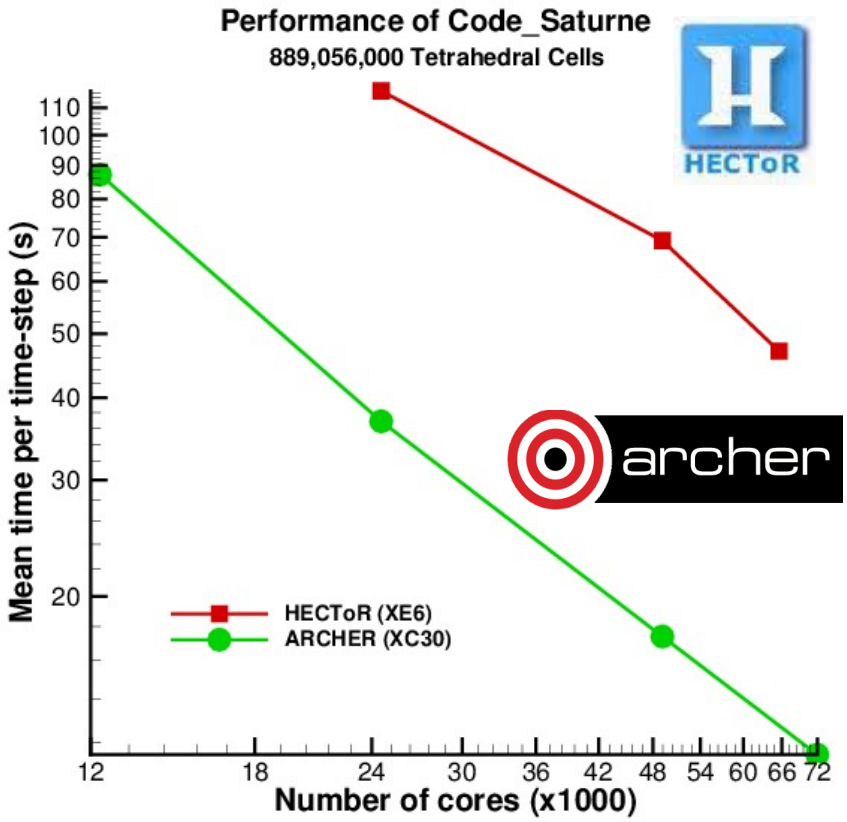


Science & Technology Facilities Council

Solver Scalability at Scale

Comparison Cray XE6 – XC30

Mesh generated by Mesh Multiplication
Cube meshed using tetrahedral cells only



105B Cell Mesh (MIRA, BGQ)

Cores	Time in Solver
262,144	789.79 s
524,288	403.18 s

13B Cell Mesh (MIRA, BGQ)

MPI Tasks	Time in Solver
524,288	70.114 s
1,048,576	52.574 s
1,572,864	45.731 s

- **105B**: 16 ranks/node
- **13B**: 32 ranks/node

Argonne NATIONAL LABORATORY

Science & Technology Facilities Council

Parallel IO Performance

mesh: dumped to disk after MM:

19.853 TiB

checkpoint: 2 files

2.338 TiB and **3.053TiB**

mesh: dumped to disk after MM:

2.483 TiB

checkpoint: 2 files

0.306 TiB and **0.391 TiB**

105B Cell Mesh (MIRA, BGQ)

Cores	mesh	checkpoint
262,144	2697.15 s	**659.72 s**
524,288	2536.23 s	**827.26 s**

13B Cell Mesh (MIRA, BGQ)

MPI Tasks	mesh	checkpoint
524,288	458.51 s	**409.60 s**
1,048,576	595.52 s	**541.26 s**
1,572,864	732.28 s	**591.66 s**

- **105B**: 16 ranks/node
- **13B**: 32 ranks/node

Argonne NATIONAL LABORATORY

Science & Technology Facilities Council

Description of the Machine

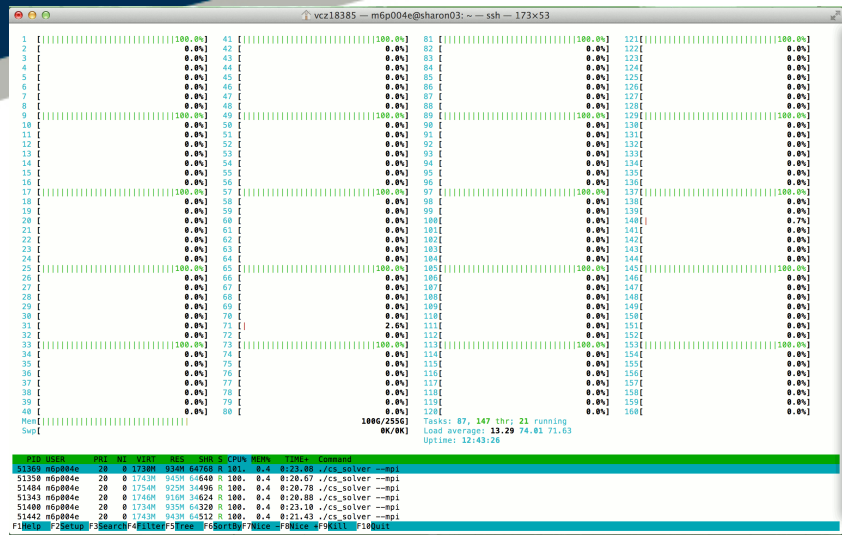
Two types of IBM POWER8 nodes are tested

IBM Power System S822LC	IBM Power System S824L
2 processors with 10 cores	2 processors with 12 cores
Up to 8 logical cores	Up to 8 logical cores
4 on-chip memory controllers (SCM)	8 on-chip memory controllers (DCM)
256 GiB RAM/node	256 GiB RAM/node
~2.92 GHz	~3.00GHz
2 NVIDIA K80	2 NVIDIA K40
2x GK210 per K80	1x GK180 per K40
2,496 stream processors	2,880 stream processors
12 GiB RAM	12 GiB RAM

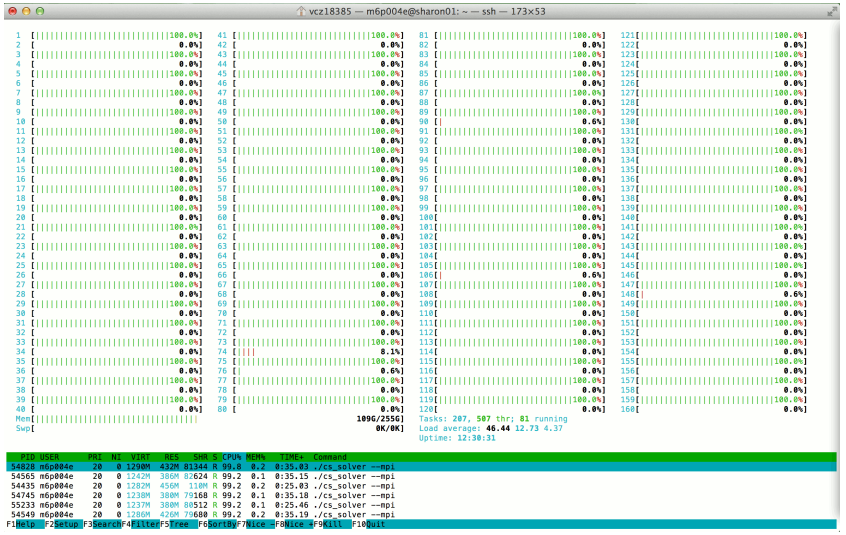
SMT expresses the number of virtual (hardware) cores or number of concurrent threads per physical core. This ratio can be set up without system reboot.

Science & Technology Facilities Council

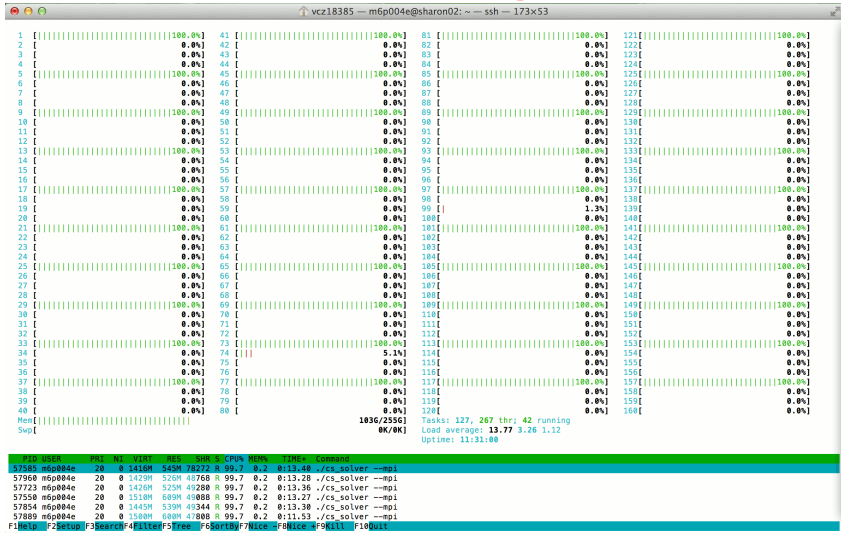
CPU Binding



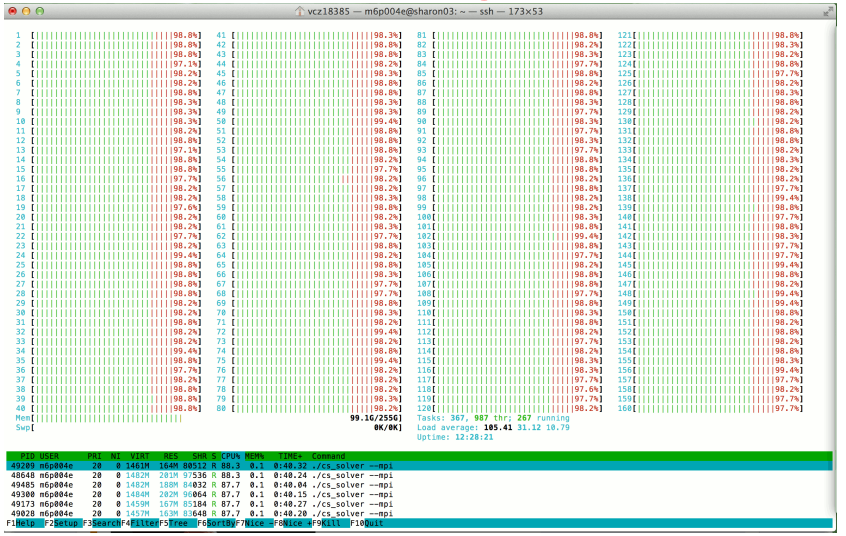
SMT8 – 1 MPI/physical core



SMT8 – 4 MPIs/physical core



SMT8 – 2 MPIs/physical core



SMT8 – 8 MPIs/physical core

Science & Technology Facilities Council

Test Case Configuration

3-D lid-driven cavity test case

Box 1x1x1

Mesh of 13 million tetrahedral cells

Boundary conditions:

Top lid: horizontal velocity

Spanwise direction: symmetry

Last 3 boundaries: wall - Dirichlet

Re=100

Settings:

Code_Saturne V 4.2.1

METIS as a partitioner

Pressure Poisson equation:

All the cases use the native AMG of the code (as a preconditioner)

Except the last one which uses the PETSc library for comparison with with GPUs (Conjugate Gradient)

POWER8 vs x86

Comparison between 2 different half/full POWER8 nodes and 1 half/full x86 node

The x86 node is an Ivy Bridge E5-2697v2 2.7GHz

S822LC			S824L			x86		
#C	T (s)	E (%)	#C	T (s)	E (%)	#C	T (s)	E (%)
10	26.50	-	12	21.51	-	12	31.61	-
20	16.43	**81**	24	11.80	**91**	24	25.33	**62**

All the runs are perfomed using SMT8 on the POWER8 nodes

For the “half node” simulations, the runs are using 2 sockets half loaded

N.B: The simulation using 1 S824L node is more than twice as fast as the one using the x86 node.

PURE MPI

The tests are carried out from 20 (24) MPI tasks on

IBM Power System S822LC		IBM Power System S824L	
	PE/XL		PE/XL
#C	T (s)	#C	T (s)
20	16.43	24	11.80
40	14.35	48	9.64
80	14.38	96	9.62
160	14.75	192	11.10

The runs are performed are all using SMT8, with respectively 20 (24) MPIs, 40 (48) MPIs, 80 (96) MPIs and 160 (192) MPIs

Performance is lost when not fully utilising the system. If logical cores are unused, it is better to reconfigure (dynamically) the system to dedicate all hardware resources to the running threads.

The tests are carried out using 20 MPI tasks

IBM Power System S822LC		
	OPENMPI/GNU	
#T	T (s)	SP
1	18.49	1.00
2	16.01	1.16
4	14.20	1.30

All the runs are using SMT8

Performance is lost when not fully utilising the system. If logical cores are unused, it is better to reconfigure (dynamically) the system to dedicate all hardware resources to the running threads.

CPU vs CPU + GPU

The PETSc Library is used to compute the pressure equation, using their Conjugate Gradient as a solver.
We use OPENMPI/GNU/CUDA.

IBM Power System S822LC						
2x P8 10-cores + 2x K80 (2 G210 per K80)						
	CPU	CPU/GPU		CPU	CPU/GPU	
#C	T_{pres} (s)	T_{pres} (s)	SP	T_{total} (s)	T_{total} (s)	SP
1	951.04	519.48	**1.83**	1022.18	630.54	**1.62**
2	595.06	280.64	**2.12**	621.20	337.12	**1.84**
4	245.62	145.73	**1.68**	263.61	173.95	**1.51**
20	72.26	*108.80*	*0.66*	76.38	*109.75*	*0.70*

If the GPUs are not overloaded, CPU+GPUs is cheaper.

Science & Technology Facilities Council

CPU vs CPU + GPU

The PETSc Library is used to compute the pressure equation, using their Conjugate Gradient as a solver.
We use OPENMPI/GNU/CUDA.

IBM Power System S824L						
2x P8 12-cores + 2x K40 (1 G180 per K40)						
	CPU	CPU/GPU		CPU	CPU/GPU	
#C	T_{pres} (s)	T_{pres} (s)	SP	T_{total} (s)	T_{total} (s)	SP
1	1012.97	512.54	**1.97**	1087.75	637.86	**1.71**
2	632.45	267.82	**2.36**	659.71	327.81	**2.01**
4	248.33	*163.51*	*1.52*	267.82	*189.04*	*1.42*
24	54.12	*106.19*	*0.65*	57.81	*112.82*	*0.51*

If the GPUs are not overloaded, CPU+GPUs is cheaper.

Conclusions – Future work

Conclusions

For 1 node only, using physical cores, much better performance on 2 different types of POWER8 nodes than on 1 x86 node

Some speedup observed using hardware threads (MPI only or MPI+OpenMP)

Comparison between PETSc (CPU) and PETSc (CPU+GPU) favourable to the latter

Future work

Test the code on several nodes & using meshes with other types of cells

Test the performance of the code when linked to AmgX (NVIDIA solver library)

Test newer hardware / software capability:
- GP100 "Pascal" GPU (more compute, better scheduling), NVLINK (GPU transfer performance)
- CUDA8 (better scheduling), Unified Memory (programmability, efficiency?)

Acknowledgements

- **UKTC (EPSRC – EP/L000261/1)**
- **Hartree Centre for using their NextScale machine**
- **PRACE 4iP**
- **INCITE PEAC, PEAC projects – ALCF –DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357.**

THANK YOU !

Any QUESTIONS ?