



MAX School on Advanced Materials and Molecular Modelling with QUANTUM ESPRESSO

### Quantum ESPRESSO on HPC and GPU systems: parallelization and hybrid architectures

Pietro Bonfà Università di Parma and CNR-Nano, Italy MaX School on Advanced Materials and Molecular Modellingwith Quantum ESPRESSO May 17-28 2021, ICTP Virtual Meeting









Jožef Stefan Institute
 Ljubljana, Slovenia



cecam

## Running on a laptop

• You go home and type ./configure && make pw



## On a real HPC node...

## 2 x AMD Epyc7352 + 4 x NVIDIA A100 ./configure && make pw



## On a real HPC node...

## $2 \times AMD Epyc7352 + 4 \times NVIDIA A100$

☑ ▲ https://www.phoronix.com/scan.php?page=article&item=intel-xeon-8380-linux&num=4

thanks in part to going from 28 to 40 cores.



Written by Michael Larabel in Processors on 12 May 2021. Page 4 of 7. 13 Comments



## QE running on HPC systems

- Parallel Message Passing Interface
- Parallel OpenMP
- Hierarchical levels of parallelization for fine grain performance tuning
- Accelerated CUDA Fortran for NVIDIA GPUs

Parallel computing (a concise introduction)

## Amdahl's law

- A task takes the time **T** to run.
- A portion p of T may benefit from parallel execution. <u>That portion</u> becomes s times faster.
- The original task now takes

$$T'(p,s) = (1-p)T + \frac{p}{s}T$$

## Amdahl's law

The speedup of the whole task is  $S = \frac{T}{T'}$ 20 and from 15  $T'(p,s) = \left[ (1-p) + \frac{p}{s} \right] T$ (s) S 10 it is easily obtained 5  $S(p,s) = \frac{1}{(1-p) + \frac{p}{s}}$ 0 1 2



## mpirun -np 3 pw.x



Time









## Message Passing & OpenMP



## Message Passing & OpenMP

# Export OMP\_NUM\_THREADS=3 mpirun -np 3 pw.x



Time

## Message Passing & OpenMP

## Export OMP\_NUM\_THREADS=3 mpirun -np 3 pw.x



Time

• From P. Giannozzi, day 2

$$\psi_i(\mathbf{r}) = \frac{1}{\sqrt{V}} \sum_{\mathbf{G}} c_{\mathbf{k}+\mathbf{G}} e^{i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}}, \quad \frac{\hbar^2}{2m} |\mathbf{k}+\mathbf{G}|^2 \le E_{cut}$$

The code computes an unsymmetrized charge density

$$\tilde{n}(\mathbf{r}) = \sum_{\mathbf{k}\in IBZ} \sum_{v} w_{\mathbf{k}} \left| \psi_{\mathbf{k},v}(\mathbf{r}) \right|^2$$





- Split KS states expansion in PW
- Default
- A lot of messages

 $N_b$ 



- Work on multiple k-points at the same time
- Less communication
- More memory
- Unbalance

 $N_b$ 



- Work on multiple k-points at the same time and split WFs.
- Less communication
- Optimal memory usage
- Reduced unbalance

 $N_h$ 

## Workload decomposition

Again from P. Giannozzi, day 2 $H\psi \equiv (T + \hat{V}_{NL} + V_{loc} + V_H + V_{xc})\psi$ 

 $(T\psi)$ : easy in **G**-space,  $T_{CPU} = \mathcal{O}(N)$  $(V_{loc} + V_H + V_{xc})\psi$ : easy in **r**-space,  $T_{CPU} = \mathcal{O}(N)$  $(\hat{V}_{NL}\psi)$ : easy in **G**-space (also in **r**-space) if  $\hat{V}$  is written in separable form  $T_{CPU} = \mathcal{O}(mN)$ , m =number of projectors

FFT is used to jump from real to reciprocal space. Operations are performed where it is easier.



Eric Pascolo, master thesis

## An example: BCC Fe

#### Parallelization info

sticks:	dense	smooth	PW	G-vecs:	dense	smooth	PW
Min	111	37	13		2974	573	129
Max	112	38	14		2976	575	132
Sum	1781	599	219		47597	9189	2093

#### Using Slab Decomposition

Dense	grid:	47597	G-vectors	FFT	dimensions:	(	50,	50,	<b>50</b> )
Smooth	grid:	9189	G-vectors	FFT	dimensions:	(	30,	30,	<b>30</b> )

## Parallel levels



## Parallel levels



## Parallel Diagonalization

- Diagonalization options
  - Davidson
  - CG
  - PPCG
  - ParO

. . .

Speed

Scaling

Memory

Davidson

 $H_{KS}\psi_j$ 

PPCG

CG

Picture from Anoop Chandran

## Parallel Diagonalization

- Diagonalization options
  - <u>Davidson</u>
  - CG
  - PPCG

- ParO

. . .

$$\tilde{H}_{ij} = \left\langle \psi_i^{(n)} \left| H_{KS} \right| \psi_j^{(n)} \right\rangle, \quad \tilde{S}_{ij} = \left\langle \psi_i^{(n)} \left| S \right| \psi_j^{(n)} \right\rangle$$
$$\left| \tilde{\psi}_i^{(n)} \right\rangle = \left( H_{diag} - \varepsilon_i S_{diag} \right)^{-1} \left( H_{KS} - \varepsilon_i S \right) \left| \psi_i^{(n)} \right\rangle$$

 $\mathbf{H}\mathbf{v} = \varepsilon \mathbf{S}\mathbf{v}$ 

 $\left\{ \left| \psi_{i}^{(n)} \right\rangle, \varepsilon_{i}^{(n)} \right\}$ 



## Trend of Parallel Diagonalization

- Results for all eigenstates.
- Additional communications in parallel Davidson.

## Fine Grained Parallelization

- OpenMP
  - In the code and in the libraries.
  - Only when MPI is saturated.
  - Generally no more than 8 threads.
  - Don't forget about it on HPC systems!
- Multithreading
  - Generally not useful

## Image parallelism

### Nudged Elastic Band

PHonon (linear response)





Images from DOI: 10.1016/j.cpc.2007.09.011 DOI: 10.1103/PhysRevB.93.174419

## Image parallelism

Nudged Elastic Band

PHonon (linear response)



Images from DOI: 10.1016/j.cpc.2007.09.011 DOI: 10.1103/PhysRevB.93.174419

## Image parallelism

## Distribution of images

mpirun neb.x -nimage I -inp neb.in > neb.out

Output:

path-images division: nimage = I

Max value: total number of images in the simulation.

Constraints:

• Depend on code using this "abstract" parallelism

Tentative optimal value: **nimage** = max possible value

## K-points aka Pools

#### mpirun pw.x -npool X -inp pw.in > pw.out

Output:
---------

K-points division: npool =

Х

Distribute k points among X pools of MPI procs. Max value: n(k)

Constraints:

- at least 1 k point per pool
- Must be a divisor of the total number of processes

Tentative optimal value: npool = max(n(k))

## Parallel diagonalization

mpirun pw.x -npool X -ndiag Y -inp pw.in > pw.out

Distribute and parallelize matrix diagonalization and matrix-matrix multiplications needed in iterative diagonalization (pw.x) or orthonormalization(cp.x).

Max value: n(MPI)/X Constraints:

• Must be square

• Must be smaller than band-group size

Tentative optimal value:

- Use it for inputs with more than 100 KS;
- depends on many architectural parameters

Output Subspace diagonalization (size of subgroup: **sqrt(Y) \* sqrt(Y)** procs)

## Finding the right balance

- Pools:
  - Very effective, low communication
  - Memory hungry!
- G vectors:
  - Lower memory footprint
  - More communication
- OpenMP:
  - Practically no memory duplication
  - When MPI is saturated
- Diagonalization method:
  - Davidson: faster, more memory
  - CG: slower, less memory



## Libraries

• A few libraries you may need



Intel<sup>®</sup> oneAPI Math Kernel Library

## AMD Optimizing CPU Libraries (AOCL)

## Input/Output

A parallel filesystem is essentially a parallel application.

- Different performance have different cost:
  - Home directory
  - Long term storage
  - Scratch space

Always set the **outdir** folder in input to a fast scratch space.

## **GPU** Acceleration

Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)						
1	Supercomputer Fugaku - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442,010.0	537,212.0	29,899	6	Tianhe-2A - TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000, NUDT National Super Computer Center in Guangzhou China	4,981,760	61,444.5	100,678.7	18,482
2	Summit - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148,600.0	200,794.9	10,096	7	JUWELS Booster Module - Bull Sequana XH2000, AMD EPYC 7402 24C 2.8GHz, NVIDIA A100, Mellanox HDR InfiniBand/ParTec ParaStation ClusterSuite, Atos Forschungszentrum Juelich (FZJ) Germany	449,280	44,120.0	70,980.0	1,764
3	Sierra - IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM / NVIDIA / Mellanox DOE/NNSA/LLNL United States	1,572,480	94,640.0	125,712.0	7,438	8	HPC5 - PowerEdge C4140, Xeon Gold 6252 24C 2.1GHz, NVIDIA Tesla V100, Mellanox HDR Infiniband, <b>Dell EMC</b> Eni S.p.A. Italy	669,760	35,450.0	51,720.8	2,252
4	Sunway TaihuLight - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway, NRCPC National Supercomputing Center in Wuxi China	10,649,600	93,014.6	125,435.9	15,371	9	Frontera - Dell C6420, Xeon Platinum 8280 28C 2.7GHz, Mellanox InfiniBand HDR, Dell EMC Texas Advanced Computing Center/Univ. of Texas United States	448,448	23,516.4	38,745.9	
5	Selene - NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100, Mellanox HDR Infiniband, Nvidia NVIDIA Corporation United States	555,520	63,460.0	79,215.0	2,646	10	Dammam-7 - Cray CS-Storm, Xeon Gold 6248 20C 2.5GHz, NVIDIA Tesla V100 SXM2, InfiniBand HDR 100, HPE Saudi Aramco Saudi Arabia	672,520	22,400.0	55,423.6	

#### https://top500.org/lists/top500/2020/11/

## Different architecture



#### More transistors devoted to data processing (but less optimized memory access and speculative execution)

## CUDA Programming

# In order to compile QE-GPU you'll need to know a bit about *CUDA* ...



## CUDA Programming

The *compute capabilities* codify the features and specifications of the target device.

- Tesla K40: 3.5
- Tesla K80: 3.7
- Tesla P100: 6.0
- Tesla V100: 7.0
- Tesla A100: 8.0

Feature Support	Compute Capability						
(Unlisted features are supported for all compute capabilities)	3.0	3.2	3.5, 3.7, 5.0, 5.2	5.3	6.x	7.x	
Atomic addition operating on 32-bit floating point values in global and shared memory (atomicAdd())			Y	es			
Atomic addition operating on 64-bit floating point values in global memory and shared memory (atomicAdd())		١	10		Y	es	
Warp vote and ballot functions (Warp Vote Functions)							
threadfence_system() (Memory Fence Functions)							
syncthreads_count(), syncthreads_and(), syncthreads_or() (Synchronization Functions)			Yı	es			
Surface functions (Surface Functions)							
3D grid of thread blocks							
Unified Memory Programming							
Funnel shift (see reference manual)	No			Yes			
Dynamic Parallelism	1	NO		Y	es		
Half-precision floating-point operations: addition, subtraction, multiplication, comparison, warp shuffle functions, conversion	No		No Yes		Yes		
Tensor Core		0	No			Yes	

From CUDA C Programming guide v10

## Compilers

# The accelerated version requires NVIDIA HPC SDK



Freely available at https://developer.nvidia.com/hpc-sdk



## **Execution flow**

Typical code progression

- 1) Memory allocated on host and device
- 2)Data is transferred **from** the *Host* **to** the *Device*
- 3)Kernel is lunched by the Host on the Device
- 4)Data is transferred **from** the
  - *Device* to the *Host*.
- 5)Memory is deallocated.



From https://commons.wikimedia.org/wiki/File:CUDA\_processing\_flow\_(En).PNG

### Quantum ESPRESSO

Wannier90 Optimize memory duplication, WanT PLUMED Directive based programming syncronization. SaX YAMBO TDDFPT Xspectra GIPAW GWL NEB allocation and PHonon PWCOND Atomic CP PWscf CORE modules Applications **Domain Specific Libraries** concurrenc) Optimize computational fficiency and concurrenc LAXLib **FFTXlib KS\_Solvers** Explicit accelerator programming  $|\delta\psi_i\rangle = \frac{1}{D-\epsilon_i}(H-\epsilon_i)|\psi_i\rangle$  $A\mathbf{v} = \lambda B\mathbf{v}$ 

TASK: parallel linear algebra LIBS: ELPA, MKL, cuBLAS, cuSOLVER, ESSL, ...

TASK: Parallel distributed FFT LIBS: FFTW, MKL, ESSL, cuFFT, ...

**TASK: Iterative solvers** LIBS: LAXLib, MKL, cuBLAS, ...

Programming model Objective

## Quantum ESPRESSO

Features	v 6.4	v 6.4.1	v 6.5a1	v 6.5a2	v6.7
Total Energy (K points)	1	1	1	1	1
Total Energy (Gamma point)	1	1	1	1	1
Spin polarized systems	1	1	1	1	\$
Non collinear simulations	1	1	1	1	\$
Forces	=	=	1	1	1
Stress	=	=	=	1	1
Exact exchange	=	=	1	1	1
LDA+U	=	=	1	1	\$
metaGGA	=	=	=	=	=
Parallel eigenproblem (ndiag)	×	×	×	×	×

## The GPUs of Marconi100

## NVIDIA Tesla V100



16GB memory per GPU

5120 CUDA Cores

Running at 1200-1300 MHz

## Performance Ratio

• Let's consider Marconi100@CINECA (ranked 11)





Icons By Misha Petrishchev, RU and iconsmind.com, GB







~31.2 TFlops

## QE with GPU acceleration



## QE with GPU acceleration



## How to run the GPU code

# 1 GPU <-> 1 MPI Fill the CPU with OpenMP threads No parallel eigensolver (-ndiag 1) yet

K-point pools are great, but device memory is limited.

## A few practical advices

#### Configure options:

enable-openmp enable-parallel	compile for openmp execution if possible (default: no) compile for parallel execution if possible (default: yes)
enable-cuda-env-check=	=yes The configure script will check CUDA installation
with-cuda=PATH with-cuda-cc=VAL with-cuda-runtime=VAL	and report problems [default=no] prefix where CUDA is installed [default=no] GPU architecture (Kepler: 35, Pascal: 60, Volta: 70) [default=35] CUDA runtime (Pascal: 8+, Volta: 9+) [default=10.1]
with-scalapack	(ves no  <b>intel</b> ) Use scalapack if available. Set to
-	"intel" to use Intel MPI and blacs (default: use openMPI)

## A few practical advices

- Checking compilation options
  - Example for an Intel based platform...

MANUAL_DFLAGS DFLAGS FDFLAGS	= = -D_DFTI -D_MPI -D_SCALAPACK -D_ELPA_2016 = \$(DFLAGS) \$(MANUAL_DFLAGS)
[]	
MPIF90 F90 CC F77 []	<pre>= mpiifort = ifort = icc = ifort</pre>
BLAS_LIBS	= -lmkl_intel_lp64 -lmkl_intel_thread -lmkl_core

## A few practical advices

- Checking compilation options
  - Example for an Intel based platform...
  - Example for a (NVIDIA) GPU platform

```
MANUAL_DFLAGS =
DFLAGS = -D_PGI -D_CUDA -D_USE_CUSOLVER -D_FFTW -D_MPI
FDFLAGS = $(DFLAGS) $(MANUAL_DFLAGS)
[...]
# GPU architecture (Kepler: 35, Pascal: 60, Volta: 70)
GPU_ARCH=70
# CUDA runtime (Pascal: 8.0, Volta: 9.0)
CUDA_RUNTIME=11.0
# CUDA F90 Flags
CUDA_F90FLAGS=-Mcuda=cc70,cuda11.0 [...]
```

#### Something very bad...

Program PWSCF v.6.2 starts on 29Nov2017 at 15:22:59

This program is part of the open-source Quantum ESPRESSO suite
for quantum simulation of materials; please cite
 "P. Giannozzi et al., J. Phys.:Condens. Matter 21 395502 (2009);
 "P. Giannozzi et al., J. Phys.:Condens. Matter 29 465901 (2017);
 URL http://www.quantum-espresso.org",
in publications or presentations arising from this work. More details at
http://www.quantum-espresso.org/quote

Parallel version (MPI & OpenMP), running on<br/>Number of MPI processes:15552 processor coresMumber of MPI processes:432Threads/MPI process:36

MPI processes distributed on 12 nodes
K-points division: npool = 2
R & G space division: proc/nbgrp/npool/nimage = 216

Subspace diagonalization in iterative solution of the eigenvalue problem: one sub-group per band group will be used custom distributed-memory algorithm (size of sub-group: 4* 4 procs)					
Message from routine setu DEPRECATED: symmetry with	up: n ibrav=0, use correct ibrav instead				
Parallelization info					
sticks: dense smooth Min 1199 640	PW G-vecs: dense smooth PW 159 383982 149548 18695				
Max 1202 642 Sum 28829 15389	16238400414956218714385592157993589319448895				
Title: DyOtBuClTHF_100K.cif					
bravais-lattice index	= 0				
lattice parameter (alat) unit-cell volume	= 25.6474 a.u. = 37134.3792 (a.u.)^3				
number of atoms/cell	= 608				
number of atomic types number of electrons	= 6 = 1512.00				
number of Kohn-Sham state	es= 756				
kinetic-energy cutoff	= 80.0000 Ry				
charge density cutoff convergence threshold	= 500.0000 Ry = 1.0F-09				
mixing beta	= 0.5000				
number of iterations used	d = 8 plain mixing				
Exchange-correlation	= SLA PW PBE PBE (1 4 3 4 0 0)				
	Subspace diagonalization one sub-group per band gr custom distributed-memory Message from routine setu DEPRECATED: symmetry with Parallelization info 				

init_run : electrons : update_pot : forces :	42.99s CPL 60819.95s CPL 1461.58s CPL 17437.52s CPL	46.16s 63107.94s 1522.64s 17714.01s	WALL ( WALL ( WALL ( WALL (	1 calls) 83 calls) 82 calls) 82 calls) 82 calls)
Called by init wfcinit :	_run: 28.44s CPL	29.07s	WALL (	1 calls)
potinit : hinit0 :	0.79s CPL 8.50s CPL	2.21s 8.60s	WALL ( WALL (	1 calls) 1 calls)
Called by elec	trons:			
c_bands : sum_band : v_of_rho : newd : mix_rho :	37126.13s CPL 9663.72s CPL 501.54s CPL 2620.20s CPL 116.23s CPL	37854.42s 10448.81s 536.25s 3367.58s 122.01s	WALL ( WALL ( WALL ( WALL ( WALL (	889 calls) 889 calls) 890 calls) 890 calls) 889 calls)
Called by c_ba init_us_2 : regterg :	nds: 296.76s CPL 36350.79s CPL	297.22s 36971.49s	WALL ( WALL (	1779 calls) 889 calls)
Called by sum_ sum_band:bec : addusdens :	band: 6.01s CPL 3042.08s CPL	6.08s 3745.04s	WALL ( WALL (	889 calls) 889 calls)
Called by *egt h_psi : s_psi : g_psi :	erg: 24521.86s CPL 3235.74s CPL 40.31s CPL	24722.48s 3235.97s 40.48s	WALL ( WALL ( WALL (	3704 calls) 3704 calls) 2814 calls)
rdiaghg :	2592.35s CPL	2678.62s	WALL (	3540 calls)
Called by h_ps	1:	24627 22-	WALL Z	2704 0011->
n_psi:pot : h_psi:calbec : vloc_psi : add_vuspsi :	24426.825 CPU 3349.63s CPU 17839.75s CPU 3237.38s CPU	24627.23s 3389.39s 17998.35s 3239.45s	WALL ( WALL ( WALL ( WALL (	3704 calls) 3704 calls) 3704 calls) 3704 calls) 3704 calls)



init_run electrons update_pot forces	: 42.99s CPU 46.16 : 60819.95s CPU 63107.94 : 1461.58s CPU 1522.64 : 17437.52s CPU 17714.01	s WALL ( 1 calls) s WALL ( 83 calls) s WALL ( 82 calls) s WALL ( 82 calls)		
Called by in wfcinit potinit	nit_run: : 28.44s CPU 29.07 :	s WALL ( 1 calls)		
hinit0 Called by el	: .ectr 80000 -	Calle	ed by PWSCF	
sum_band v_of_rho newd mix_rho	80000 -	<b>C</b> alled by ele	ectrons	
Called b init_us_ regterg	60000 - जु			
Called b sum_band b addusden	20000 -			
Called b h_psi s_psi g_psi rdiaghg	0 <b>-</b>	bands sum_ban	d newd	
Called by h_ h_psi:pot h_psi:calbec vloc_psi add_vuspsi	psi: : 24426.82s CPU 24627.23 : 3349.63s CPU 3389.39 : 17839.75s CPU 17998.35 : 3237.38s CPU 3239.45	s WALL ( 3704 calls) s WALL ( 3704 calls) s WALL ( 3704 calls) s WALL ( 3704 calls) s WALL ( 3704 calls)		]

