HECToR – CSE technical meeting , Oxford 2009

Parallel Algorithms for the Materials Modelling code CRYSTAL

Dr Stanko Tomić

Computational Science & Engineering Department,
STFC Daresbury Laboratory, UK

Acknowledgements

Dr Ian Bush (NAG)
Dr Barry Searle (DL-STFC)
Prof Nic Harrison (Imperial-DL-STFC)

dCSE-EPSRC Initiative

Outline

Motivation

CRYSTAL code: Theoretical background

CRYSTAL code: Numerical implementation

CRYSTAL code: Parallelisation

CRYSTAL code: Memory economisation

Concussions

Motivation for using highly parallel CRYSTAL code

a) Miniaturisation of devices and advance in current technology, requires material description at atomistic level and detailed quantum mechanical knowledge of its properties.

b) New materials can not be described from empirical method because lack of parameters. Experimentally to expensive or “parameter space” is too big to be completely examined.

c) Ab-initio model (parameter free) is the only option then

d) Community that I support is interested in very big (computationally) systems:

   1. bulks with dilute level of impurities (supercells) as a new materials, a building blocks, for the active region of novel PV or optoelectronics devices
   2. Band alignments (offsets) at the surface interfaces: new materials, transport properties
   3. Electronic structure of the QD (nano-clusters) form the first principles: new materials, PV, chemistry, biology

e) The (1-3) are very challenging task and NEW algorithms are needed now!

Outline

Motivation

CRYSTAL code: Theoretical background

CRYSTAL code: Numerical implementation

CRYSTAL code: Parallelisation

CRYSTAL code: Memory economisation

Concussions

Hohenberg-Kohn Theory

For any legal wave-function (anti-symmetric and normalised)

$$E[\Psi] = \int \Psi^* \hat{H} \Psi d\mathbf{r} = \langle \Psi^* | \hat{H} | \Psi \rangle$$

The total energy is functional of Ψ

$$E[\Psi] > E_0$$

Search all Ψ(3N-coordinates) to minimize *E* in order to find ground state E_0!!!

Theorem 1: External potential and the number of electrons in the system are uniquely determined by the density ρ(r), so that the total energy is a unique functional of density: $E=E[\rho(\mathrm{r})]$

Theorem 2: The density that minimises the total energy is the ground state density and the minimum energy is the ground state energy, min{$E[\rho(\mathrm{r})]$}=E_0

Density Functional Theory-LDA

$$\left\{-\frac{1}{2}\sum_i \nabla_i^2 + \sum_{i,j}\frac{1}{|r_i - r_j|} + \sum_{i,\alpha}\frac{Z_\alpha}{|r_i - R_\alpha|}\right\}\Psi(r_1,\ldots,r_N) = E\Psi(r_1,\ldots,r_N)$$

For the <u>ground state energy and density</u> there is an *exact* mapping between many body system and *fictitious* non-interacting system

$$\left\{-\frac{1}{2}\nabla^2 + \int\frac{\rho(r')}{|r - r'|}dr' + \sum_{\alpha}\frac{Z_\alpha}{|r - R_\alpha|} + V_{XC}\right\}\psi_i(r) = E_i\psi_i(r)$$

- **The *fictitious* system is subject to an unknown potential V_{XC} derived from exchange-correlation functional**
- **In LDA the energy functional can be approximated as a local function of density:**

$$V_{XC}[\rho(r)] = \frac{\delta E_{XC}[\rho]}{\delta\rho} \qquad \textbf{and} \qquad \rho(r) = \sum_i |\psi_i(r)|^2$$

The DFT Eq. describe non-interacting electrons under the influence of a mean field potential consisting of the classical Coulomb potential and local exchange-correlation potential

Hartree-Fock Theory

Specific structure of HF WF: $\Psi_{\text{HF}} \frac{1}{\sqrt{N!}} \det[\psi_1\psi_2 ... \psi_N]$

$$E_{HF} = \int \psi_i^*(r)\left(-\frac{1}{2}\nabla + \sum_\alpha \frac{Z_\alpha}{|r-R_\alpha|}\right)\psi_i(r)dr \underbrace{+\frac{1}{2}\sum_{i,j}\int\frac{\psi_i^*(r_1)\psi_i(r_1)\psi_j^*(r_2)\psi_j(r_2)}{|r_i-r_j|}dr_1dr_2}_{\textbf{Coulomb energy}} \underbrace{-\frac{1}{2}\sum_{i,j}\int\frac{\psi_i^*(r_1)\psi_j(r_1)\psi_j^*(r_2)\psi_i(r_2)}{|r_i-r_j|}dr_1dr_2}_{\textbf{Exchange energy}}$$

After application of the variation theorem and under: $\langle\psi_i|\psi_j\rangle = \delta_{ij}$

The ground state Hartree-Fock Equation:

$$\left\{-\frac{1}{2}\nabla^2 + \int\frac{\rho(r')}{|r-r'|}dr' + \sum_\alpha\frac{Z_\alpha}{|r-R_\alpha|}\right\}\psi_i(r) + \underbrace{\int v_X(r,r')\psi_i(r')dr'}_{} = E_i\psi_i(r)$$

Exact non-local exchange potential: $= -\sum_j \frac{\psi_j(r)\psi_j^*(r')}{|r-r'|}\psi_i(r')dr'$

The Hartree-Fock Eq. describe non-interacting electrons under the influence of a mean field potential consisting of the classical Coulomb potential and NON-LOCAL exchange potential

Self Interaction Energy

- In DFT-LDA each individual electron contributes to the total density
- An electron in an occupied orbital interacts with *N* electrons, when it should interact with *N*-1 — this is the *self interaction*
- DFT-LDA: all occupied bands are pushed up in energy by this interaction — consequence of an on-site/diagonal Coulomb/Exchange repulsion. DFT-LDA band gap are TOO SMALL
- Hartree-Fock theory: Coulomb and exchange interactions cancels exactly, i.e. $J_{ii} = K_{ii}$, no self interaction, + absent of correlations: HF band gaps TOO BIG
- Hybrid functional — 20% HF and 80% DFT ⇒
  Energy gaps appear to be good!!!

Functionals: PBE0 & B3LYP

Hybrid Functional: An exchange-correlation functional in DFT that incorporates portion of exact exchange from HF ($E_x^{HF} \equiv exact$) with exchange and correlation from other sources (LDA, GGA, etc.)

PBE0 = Perdew-Burke-Ernzerhof

$$E_{xc}^{\mathrm{PBE0}} = E_{xc}^{GGA} + 0.25(E_x^{HF} - E_x^{GGA})$$

B3LYP = Becke exchange + (Lee-Yang-Parr + Vosko-Wilk-Nusair) correlation

$$E_{xc}^{\mathrm{B3LYP}} = E_{xc}^{LDA} + a_1(E_x^{HF} - E_x^{GGA}) + a_2(E_x^{LDA} - E_x^{GGA}) + a_3(E_c^{LDA} - E_c^{GGA})$$

$$E_x^{\mathrm{B3LYP}} = 0.8E_x^{LDA} + 0.2E_x^{HF} + 0.72(E_x^{\mathrm{B88}(GGA)} - E_x^{LDA})$$

$$E_c^{\mathrm{B3LYP}} = 0.19E_c^{VWN3(LDA)} + 0.81E_c^{LYP(GGA)}$$

$$E_x^{HF} \equiv \text{exact} \qquad E_{xc}^{LDA} = \int d^3\mathbf{r}\, f[\rho(\mathbf{r})] \qquad E_{xc}^{GGA} = \int d^3\mathbf{r}\, f[\rho(\mathbf{r}), \nabla\rho(\mathbf{r})]$$

Outline

Motivation

CRYSTAL code: Theoretical background

CRYSTAL code: Numerical implementation

CRYSTAL code: Parallelisation

CRYSTAL code: Memory economisation

Concussions

Implementation: Gaussian Basis Set

Anti-symmetric many body wave function

$$\Phi_N = \hat{A}\prod_{i=1}^{N}\psi_i(\mathbf{r}_i,\mathbf{k})$$

One-electron wave-function=LC of Bloch functions

$$\psi_i(\mathbf{r}_i,\mathbf{k}) = \sum_{\mu}\boxed{a_{\mu,i}(\mathbf{k})}\chi_\mu(\mathbf{r}_i,\mathbf{k})$$

needs to be found

Bloch function=LC of Atomic Orbitals

$$\chi_\mu(\mathbf{r}_i,\mathbf{k}) = \sum_{\mathbf{R}}\varphi_\mu(\mathbf{r}_i-\mathbf{r}_\mu-\mathbf{R})e^{i\mathbf{kR}}$$

Atomic Orbitals = LC of n_G Gaussian type functions (orbitals) (GTO)

coeficients

$$\varphi_\mu(\mathbf{r}_i-\mathbf{r}_\mu-\mathbf{R}) = \sum_{j=1}^{n_G}\boxed{d_j}\mathcal{G}(\alpha_j,\mathbf{r}_i-\mathbf{r}_\mu-\mathbf{R})$$

exponents

$$\mathcal{G}(\alpha_j,\mathbf{r}_i-\mathbf{r}_\mu) = \exp[\boxed{-\alpha_j}(\mathbf{r}_i-\mathbf{r}_\mu)^2](x_i-x_\mu)^l(y_i-y_\mu)^m(z_i-z_\mu)^n$$

$$\mathcal{G}_{\mathbf{r}} = \mathcal{G}_x\cdot\mathcal{G}_y\cdot\mathcal{G}_z$$

Linear dependency problem in basis set

Typically Gaussian Atomic Orbitals, $\varphi(r_i-r_\mu-R)$, are localized and non-orthogonal !!!

Very "diffused" Gaussians (very small exponent α_j) can cause "linear dependency problem" and trouble SCF cycles !!!

Two ways to cure this problem:

1. To limit basis set to less diffused Gaussians: typically up to ~0.1 Bohr^{-1}
2. To project out several smallest eigenvalues of the overlap matrix S(k) that enters Fock equation: $F(k)A(k) = S(k)E(k)A(k)$. In this way possible singularities in Fock matrix F(k) are avoided during SCF cycles

Implementation: Gaussian Basis Set

Solution of secular equation in k- space gives: $a_{\mu,i}(\mathbf{k})$

$$\mathbf{F}(\mathbf{k})\mathbf{A}(\mathbf{k}) = \mathbf{S}(\mathbf{k})\mathbf{E}(\mathbf{k})\mathbf{A}(\mathbf{k})$$

While we have basis functions (Gaussians) in r- space

$$\mathbf{F}(\mathbf{k}) = \sum_{\mathbf{g}} \mathbf{F}(\mathbf{g}) e^{i\mathbf{k}\cdot\mathbf{g}}$$

Fock matrix sum of one- and two- electron contributions

$$F_{i,j}(\mathbf{g}) = H_{i,j}(\mathbf{g}) + B_{i,j}(\mathbf{g})$$

One electron contributions: kinetic & nuclear

$$H_{i,j}(\mathbf{g}) = T_{i,j}(\mathbf{g}) + Z_{i,j}(\mathbf{g}) = \langle i \mid \hat{T} \mid j \rangle + \langle i \mid \hat{Z} \mid j \rangle$$

Two electron contribution: Coulomb and exchange

$$B_{i,j}(\mathbf{g}) = C_{i,j}(\mathbf{g}) + X_{i,j}(\mathbf{g}) = \sum_{k,l} P_{k,l}[\langle i,j \mid k,l \rangle - \frac{1}{2}\langle i,k \mid j,l \rangle]$$

NON-LOCAL term

Outline

Motivation

CRYSTAL code: Theoretical background

CRYSTAL code: Numerical implementation

CRYSTAL code: Parallelisation

CRYSTAL code: Memory economisation

Concussions

CRYSTAL scalability on HECToR

Test 1: Bulk material (periodic solid) with dilute amount of impurities (1.62%): $Ga_{432}Sb_{426}N_6$

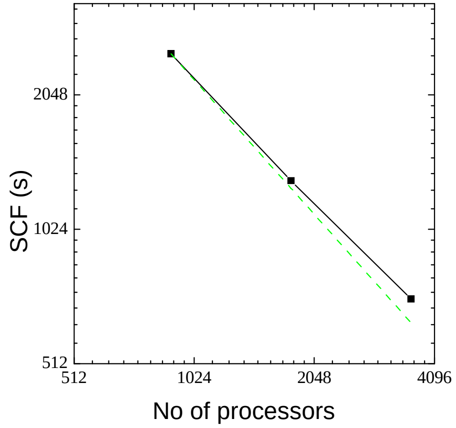
No of atoms: 864

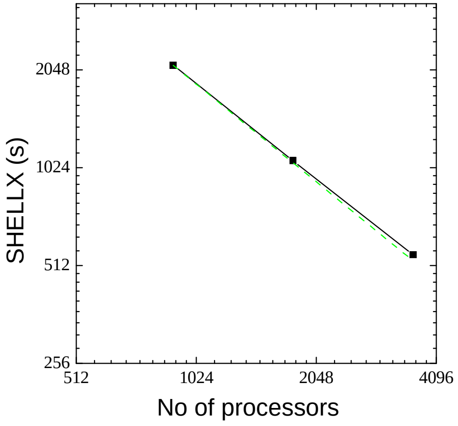
Basis sets: m-pVDZ-PP relativistic eff. core for Ga, Sb; and m-6-311G for N

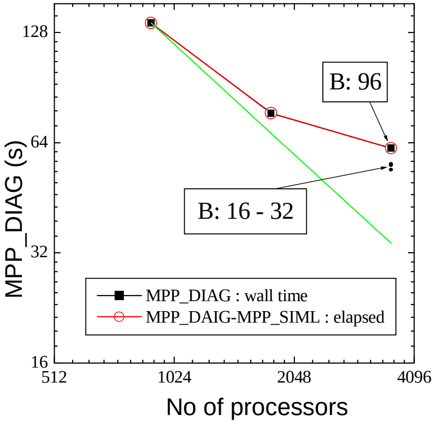
No of AO: 19837

Symmetry : 1

K-points: 1 X R + 1 x C







Problem identified: system run only when: #PBS -l mppnppn=1

Bad memory management => unnecessary expensive to run

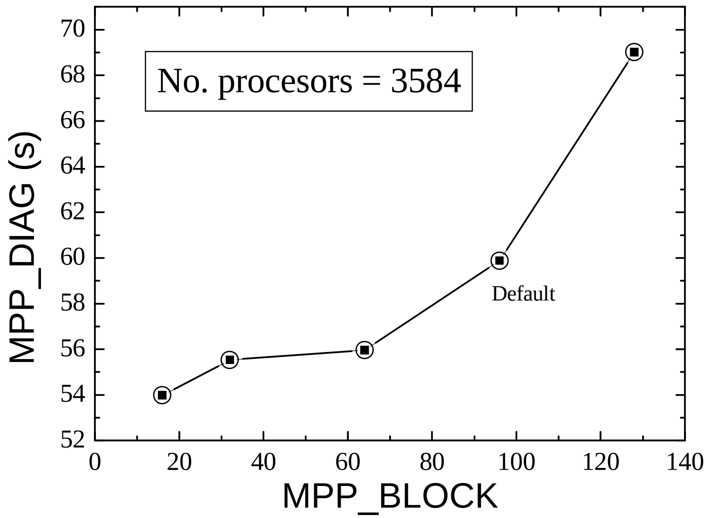
CRYSTAL scalability on HECToR

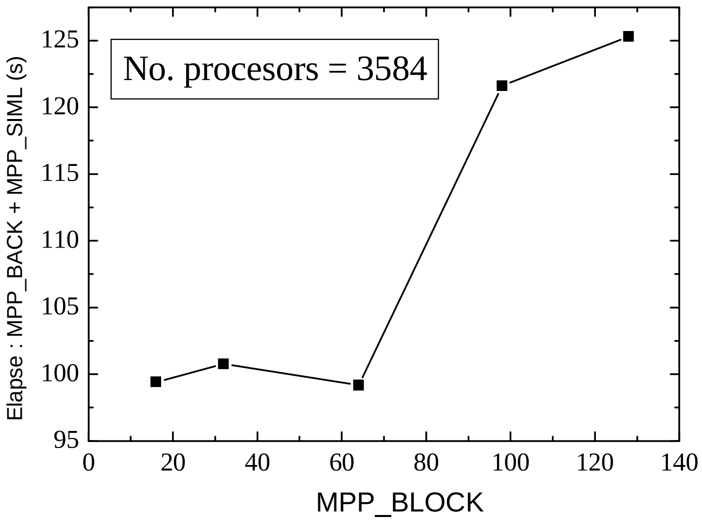
Test 1:

New command is implemented in CRYSTAL:

```
....
MPP_BLOCK
16
....
```

if omitted in INPUT file
“block” is calculated in
CRYSTAL code





CRYSTAL scalability on HECToR

Test 2: Slab CdSe/ZnS

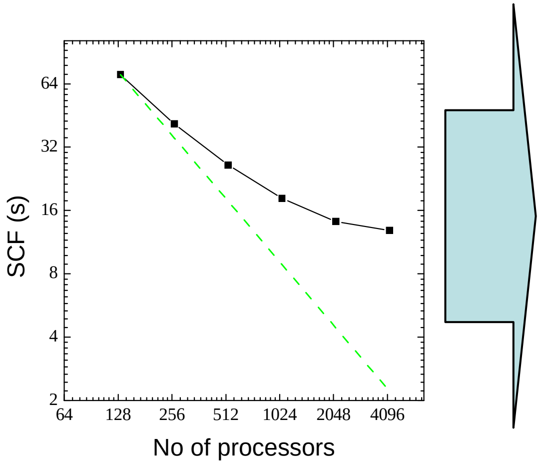
No of atoms: 128

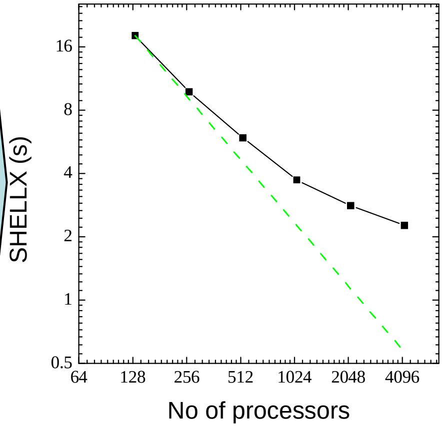
Basis sets: All electron basis sett for all: Cd, Se, Zn, and S

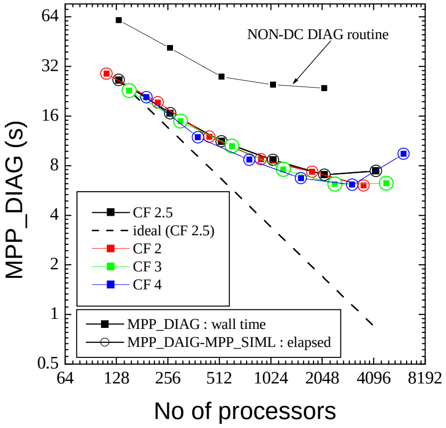
No of AO: 4024

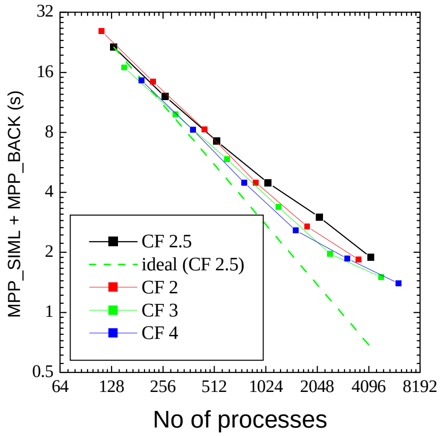
Symmetry : 2

K-points: 4 X R + 5 x C









CMPLXFAC was changed in order to find optimal load balance between complex and real k-points diagonalization

CRYSTAL scalability on HECToR

Test 3: Cluster (QD) $Cd_{159}Se_{166}$

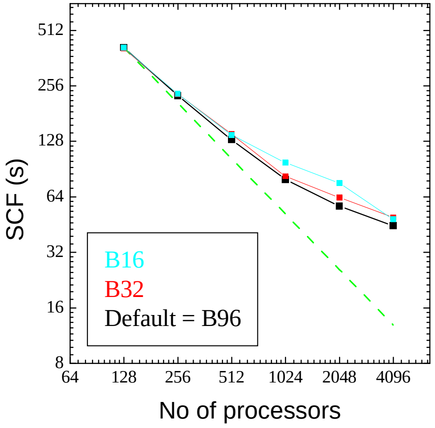
No of atoms: 325

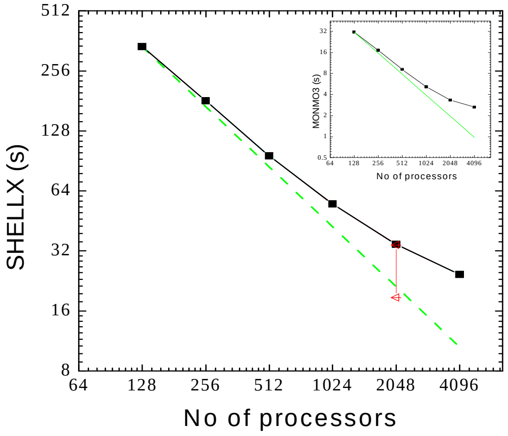
Basis sets: m-pVDZ-PP relativistic eff. core for Cd and Se

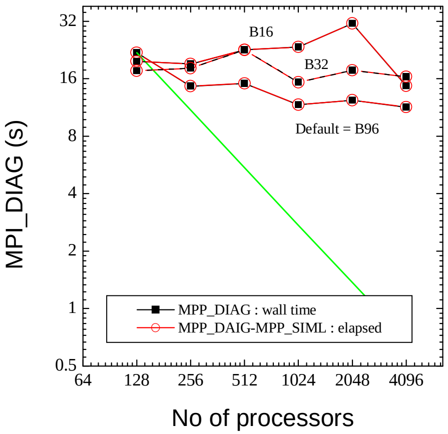
No of AO: 8747

Symmetry : 6

K-points: 1 X R







Outline

- Motivation
- CRYSTAL code: Theoretical background
- CRYSTAL code: Numerical implementation
- CRYSTAL code: Parallelisation
- **CRYSTAL code: Memory economisation**
- Concussions

CRYSTAL memory economization on HECToR

Test 1: Bulk material (periodic solid) with dilute amount of impurities (1.62%): $Ga_{432}Sb_{426}N_6$

No of atoms: 864

No of AO: 19837

Symmetry : 1

Problem identified: system run ONLY when: #PBS -l mppnppn=1

Bad memory management => unnecessary expensive to run !!!

Solution: Replicated data structure of the Fock and density matrices is replaced with direct space IRREDUCIBLE representation of the Fock and density matrices for all class of SCF calculations

CRYSTAL memory economization on HECToR

```
st53@nid00015:~/work/Bulk_GaSbN/896_sym> grep CYC mppnppn_1/out.dat
 INFORMATION **** MAXCYCLE **** MAX NUMBER OF SCF CYCLES SET TO        2
 MAX NUMBER OF SCF CYCLES       2  CONVERGENCE ON DELTAP       10**-16
 CYC   0 ETOT(AU) -2.147182650300E+05 DETOT -2.15E+05 tst  0.00E+00 PX  0.00E+00
 CYC   1 ETOT(AU) -2.145996462515E+05 DETOT  1.19E+02 tst  0.00E+00 PX  0.00E+00
 CYC   2 ETOT(AU) -2.146025490714E+05 DETOT -2.90E+00 tst  1.60E-04 PX  0.00E+00
 == SCF ENDED - TOO MANY CYCLES           E(AU) -2.1460254907141E+05 CYCLES   2
```

#PBS -l mppwidth=896
#PBS -l mppnppn=1

```
st53@nid00015:~/work/Bulk_GaSbN/896_sym> grep CYC mppnppn_2/out.dat
 INFORMATION **** MAXCYCLE **** MAX NUMBER OF SCF CYCLES SET TO        2
 MAX NUMBER OF SCF CYCLES       2  CONVERGENCE ON DELTAP       10**-16
 CYC   0 ETOT(AU) -2.147182650300E+05 DETOT -2.15E+05 tst  0.00E+00 PX  0.00E+00
 CYC   1 ETOT(AU) -2.145996462515E+05 DETOT  1.19E+02 tst  0.00E+00 PX  0.00E+00
 CYC   2 ETOT(AU) -2.146025490302E+05 DETOT -2.90E+00 tst  1.60E-04 PX  0.00E+00
 == SCF ENDED - TOO MANY CYCLES           E(AU) -2.1460254903022E+05 CYCLES   2
```

#PBS -l mppwidth=896
#PBS -l mppnppn=2

```
st53@nid00015:~/work/Bulk_GaSbN/896_sym> grep CYC mppnppn_4/out.dat
 INFORMATION **** MAXCYCLE **** MAX NUMBER OF SCF CYCLES SET TO        2
 MAX NUMBER OF SCF CYCLES       2  CONVERGENCE ON DELTAP       10**-16
 CYC   0 ETOT(AU) -2.147182650300E+05 DETOT -2.15E+05 tst  0.00E+00 PX  0.00E+00
 CYC   1 ETOT(AU) -2.145996462515E+05 DETOT  1.19E+02 tst  0.00E+00 PX  0.00E+00
 CYC   2 ETOT(AU) -2.146025573510E+05 DETOT -2.91E+00 tst  1.60E-04 PX  0.00E+00
 == SCF ENDED - TOO MANY CYCLES           E(AU) -2.1460255735103E+05 CYCLES   2
```

#PBS -l mppwidth=896
#PBS -l mppnppn=4

CRYSTAL memory economization on HAPU

```
[st53@hapu33 test_Si]$ grep CYC std/out.dat
 INFORMATION **** MAXCYCLE **** MAX NUMBER OF SCF CYCLES SET TO       200
 MAX NUMBER OF SCF CYCLES    200  CONVERGENCE ON DELTAP       10**-17
 CYC   0 ETOT(AU) -4.567509511167E+03 DETOT -4.57E+03 tst  0.00E+00 PX  1.00E+00
 CYC   1 ETOT(AU) -4.570469348171E+03 DETOT -2.96E+00 tst  0.00E+00 PX  1.00E+00
……
 == SCF ENDED - CONVERGENCE ON ENERGY      E(AU) -4.5705661190498E+03 CYCLES   9
```

Old code: replicated data in memory

…

MPP

```
[st53@hapu33 test_Si]$ grep CYC sym/out.dat
 INFORMATION **** MAXCYCLE **** MAX NUMBER OF SCF CYCLES SET TO       200
 MAX NUMBER OF SCF CYCLES    200  CONVERGENCE ON DELTAP       10**-17
 CYC   0 ETOT(AU) -4.567509511167E+03 DETOT -4.57E+03 tst  0.00E+00 PX  0.00E+00
 CYC   1 ETOT(AU) -4.570469348171E+03 DETOT -2.96E+00 tst  0.00E+00 PX  0.00E+00
 …
 == SCF ENDED - CONVERGENCE ON ENERGY      E(AU) -4.5705661190498E+03 CYCLES   9
[
```

New code: irreducible representation

…

MPP

LOWMEM

```
[st53@hapu33 test_Si]$ grep CYC sym_nolm/out.dat
 INFORMATION **** MAXCYCLE **** MAX NUMBER OF SCF CYCLES SET TO       200
 MAX NUMBER OF SCF CYCLES    200  CONVERGENCE ON DELTAP       10**-17
 CYC   0 ETOT(AU) -4.567509511167E+03 DETOT -4.57E+03 tst  0.00E+00 PX  0.00E+00
 CYC   1 ETOT(AU) -4.570469348171E+03 DETOT -2.96E+00 tst  0.00E+00 PX  0.00E+00
 …
 == SCF ENDED - CONVERGENCE ON ENERGY      E(AU) -4.5705661190498E+03 CYCLES   9
```

New code: irreducible representation

…

MPP

NOLOWMEM

CRYSTAL memory economization on HECToR

HECToR phase 1	HECToR phase 2
1 core/node = 6 GB/core	1 core/node = 8 GB/core
2 core/node = 3 GB/core	2 core/node = 4 GB/core
	4 core/node = 2 GB/core

Memory economisation achieved: from more then 3 GB/core required to less the 2 GB/core, i.e., the whole node can be used efficiently.

Outline

Motivation

CRYSTAL code: Theoretical background

CRYSTAL code: Numerical implementation

CRYSTAL code: Parallelisation

CRYSTAL code: Memory economisation

Concussions

Conclusions

1) Memory bottleneck in the CRYSTAL code caused by replicated data structure of the Fock and density matrix is removed by replacing those data structures with its irreducible representations [saving: problem size/(2 x symmetry of the system)].

2) For big-ish systems the optimal load balance between complex and real k-points is achieved with CMPLXFAC=3, that provide for best scalability of the CRYSTAL code up to 4864 cores on HECToR.

3) Implementation of the new command MPP_BLOCK provides for better control of the block size of distributed data. Reducing block size from default value of 96 to 64 or 32 it can be achieved more than 10% speed-up in diagonalisation part and about 20% speed-up in back and similarity transform for the system size rank=19837 run on 3584 cores.

4) For systems when integral calculations time prevails (MONMO3, SHELLX) over diagonalisation time (MPP_DIAG) better synchronisation of integral calculations is needed: possible global or shared counter implementations investigation.

5) For systems where diagonalisation (MPP_DIAG) dominates over integral calculations (MONMO3, SHELLX) careful examination of blocking is needed