CentraleSupélec

SG6 - HPC

TD2-3/Lab-2 – Part 1: Deployment of an MPI application on a PC cluster

Stéphane Vialle

CentraleSupélec universite PARIS-SACLAY
ÉCOLE DOCTORALE Sciences et technologies de l'information et de la communication (STIC)
RISEGrid Research Institute for Smarter Electric Grids
Grand Est ALSACE CHAMPAGNE-ARDENNE LORRAINE

Deployment of an MPI application on a PC cluster

Deployment rules & communication scheme

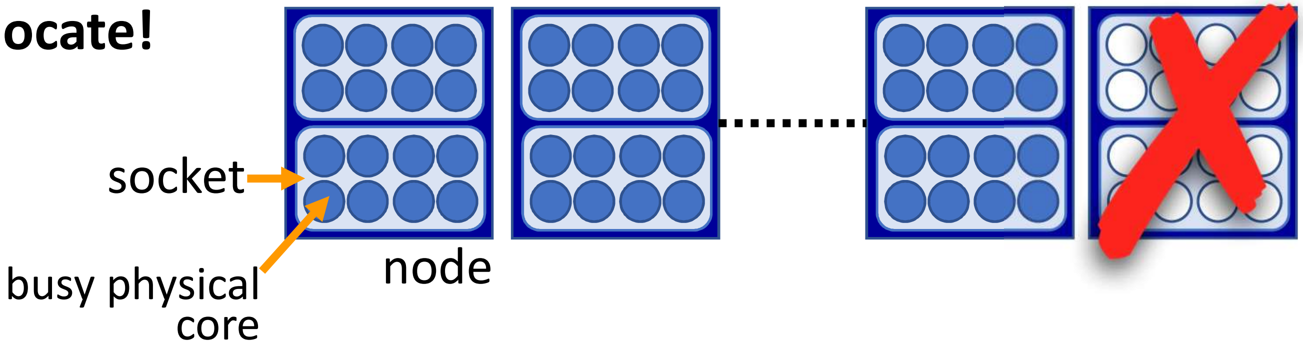
1st deployment: using processes and threads

2nd deployment: using only processes

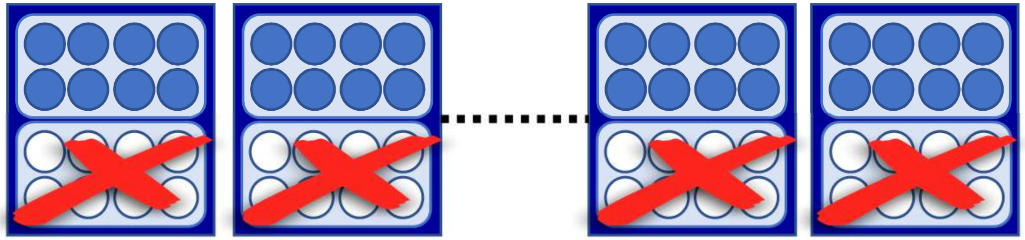
Deployment rules & communication scheme

Do not waste resources!

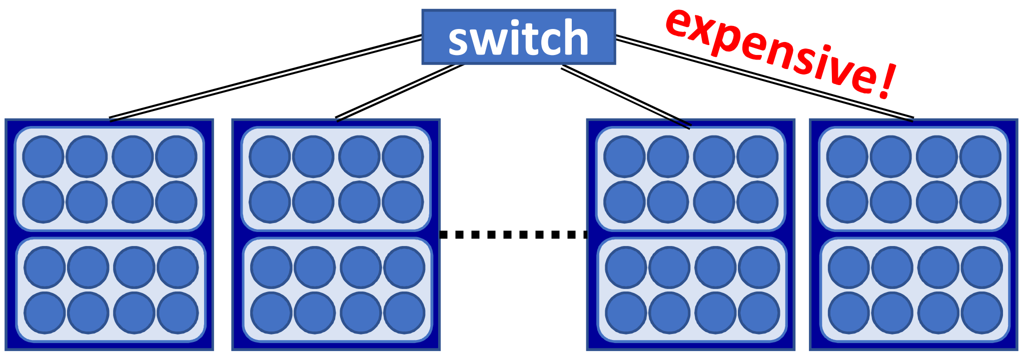
Use ALL nodes you allocate!



Use ALL physical cores of your nodes!
(with processes or threads)



Minimize the communication cost across the interconnection network (maximizing comm. inside each node)

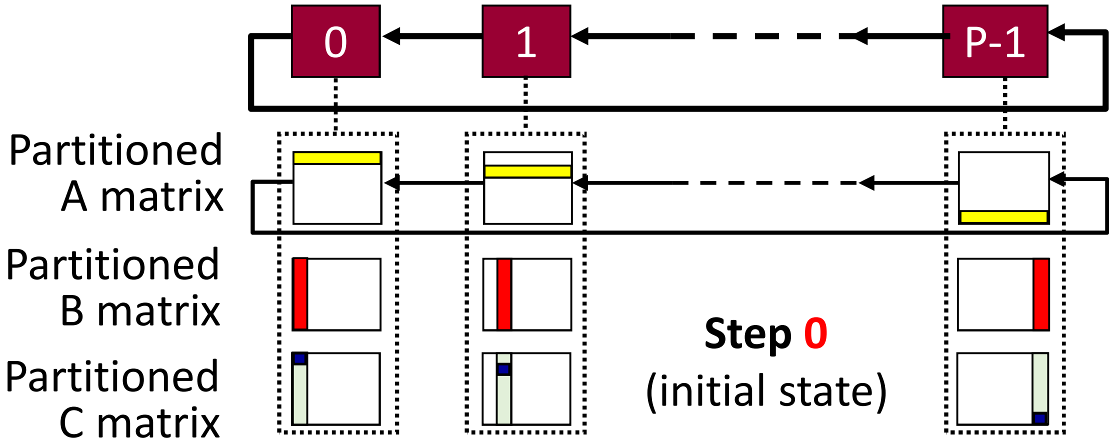


Deployment rules & communication scheme

Virtual ring of processes

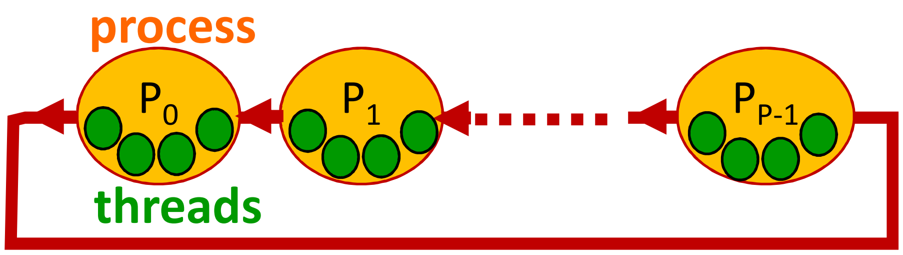
Distributed *Matrix Product* algorithm:

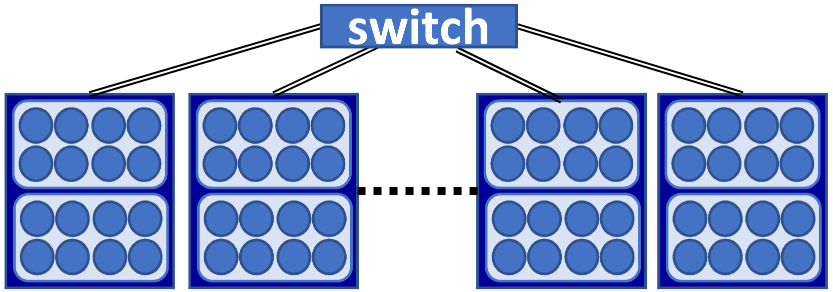
- ring comm. scheme
- P_i communicates only with P_{i-1} and P_{i+1}



Distributed & multithreaded implementation:

- MPI + OpenMP
- OpenBLAS





For a given nb of allocated nodes (N_n):

→ **Find 2 relevant *mpirun* commands**

- Not wasting any resource
- Minimizing the comm. cost

Deployment of an MPI application on a PC cluster

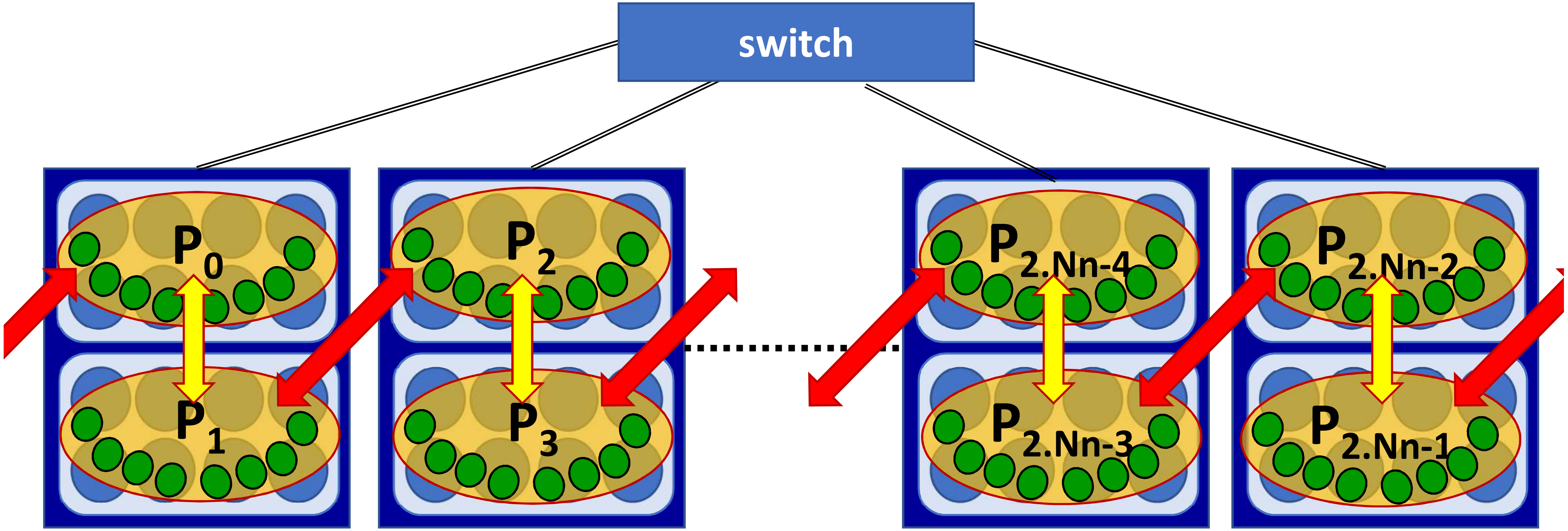
Deployment rules & communication scheme

1st deployment: using processes and threads

2nd deployment: using only processes

1st deployment: using processes and threads

Deployment strategy



```
mpirun –np  XX=2×Nn –machinefile machines.txt
            –map-by ppr:1:socket
            –rank-by socket
             –bind-to socket
             ./MatrixProduct –klc YY –k 1 –nt 8
```

comms: only 50% are expensive!

1st deployment: using processes and threads

TO DO (1)

Questions:

1. **Measure performances (Gflops) on 4, 8 and 16 nodes, with –k 1 –klc 16**

   → Use OAR « *batch mode* » with « *myrun* » shell script:

   - **Ex: oarsub -p "cluster='kyle'" -l nodes=4 *'./myrun 8 16'***
   - after unzipping the archive, don't forget:

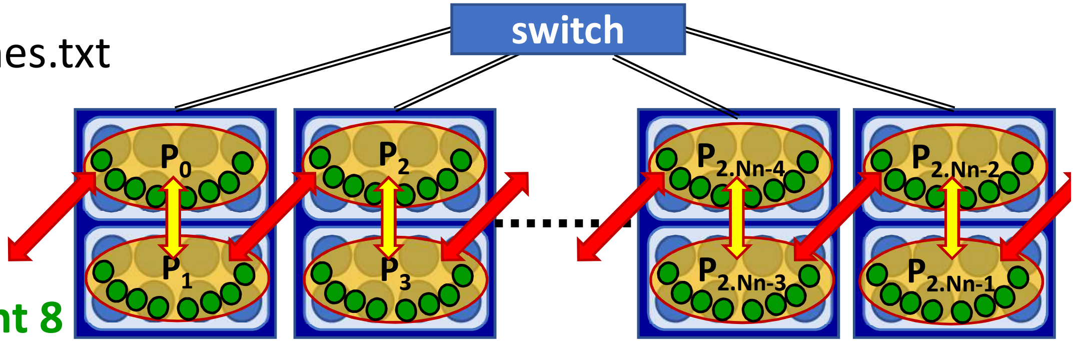
     dos2linux myrun and *chmod 700 myrun*

2. **Compare to previous measurements on Kyle cluster** (S. Vialle – 27/12/2019):

Nb of nodes	1	2	4	8	16	32
Gflops	369	529	819	1103	1359	1387

3. **Draw performance curves**
4. **Analyse the performance curves**

```
mpirun –np  XX=2×Nn –machinefile machines.txt
        –map-by ppr:1:socket
        –rank-by socket
        –bind-to socket
        ./MatrixProduct –klc YY=16 –k 1 –nt 8
```



Deployment of an MPI application on a PC cluster

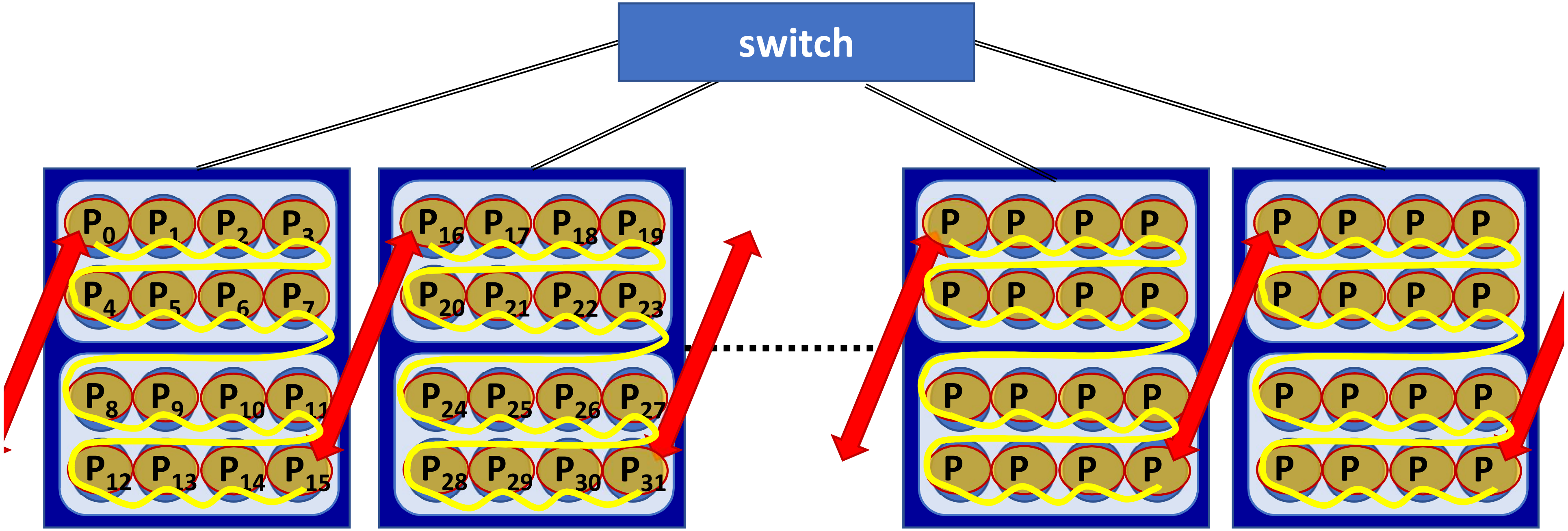
Deployment rules & communication scheme

1st deployment: using processes and threads

2nd deployment: using only processes

2nd deployment: using only processes

Deployment strategy


switch
P_0 P_1 P_2 P_3
P_4 P_5 P_6 P_7
P_8 P_9 P_{10} P_{11}
P_{12} P_{13} P_{14} P_{15}
P_{16} P_{17} P_{18} P_{19}
P_{20} P_{21} P_{22} P_{23}
P_{24} P_{25} P_{26} P_{27}
P_{28} P_{29} P_{30} P_{31}
P P P P
P P P P

```
mpirun –np  XX=?? –machinefile machines.txt
            –map-by ppr:?:???
            –rank-by ?
            –bind-to ?
            ./MatrixProduct –klc YY –k 1 –nt ??
```

Comms: only 1/16 are expensive!

But 8x more steps.

→ Same total volume on the interconnect (see course slides)

CentraleSupélec

2nd deployment: using only processes

TO DO (2)

Questions:

1. **Measure performances (Gflops) on 4, 8, 16 and 32 nodes, with –k 1 –klc 16**

   → Use OAR « *batch mode* » with « *myrun* » shell script

   → **MODIFY *myrun* shell script and adapt oarsub command**

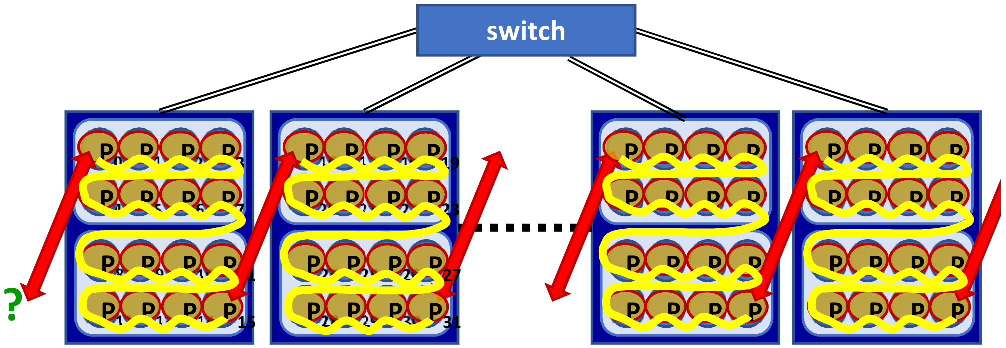
2. **Compare to previous measurements on Kyle cluster:**

Nb of nodes	1	2	4	8	16	32
Gflops	367	562	862	1290	…	…

3. **Draw performance curves (superpose with 1st deployement curves)**

4. **Analyse the performance curves**

```
mpirun –np  XX=?? –machinefile machines.txt
        –map-by ppr:?:???
        –rank-by ?
         –bind-to ?
         ./MatrixProduct –klc YY=16 –k 1 –nt ??
```



CentraleSupélec

TD2-3/Lab-2 – Part 1:
Deployment of an MPI application on a PC cluster

End