Message Passing Interface

January 19, 2021

Message Passing Interface (MPI)

- Efforts began in 1991 by Jack Dongarra, Tony Hey, and David W. Walker.
- Standard for message passing in a distributed memory environment
- MPI Forum in 1993
  - Version 1.0: 1994
  - Version 2.0: 1997
  - Version 3.0: 2012
  - Version 4.0: 2020 (draft)

MPI Implementations

"The MPI standard includes point-to-point message-passing, collective communications, group and communicator concepts, process topologies, environmental management, process creation and management, one-sided communications, extended collective operations, external interfaces, I/O, some miscellaneous topics, and a profiling interface." – MPI report

- MPICH (ANL)
- MVAPICH (OSU)
- OpenMPI
- Intel MPI
- Cray MPI

CSE Lab cluster

- ~ 60 nodes connected via Ethernet
- Each node has 12 cores
- Intel(R) Core(TM) i7-8700 CPU @ 3.20GHz
- NFS filesystem
  - Your home directories are NFS-mounted on all nodes
- Login with CSE login credentials to any machine (IP address range 172.27.19.1 – 172.27.19.40)
  - Its possible that some machines are not reachable/usable, try some other IP!

CSE Lab Cluster

- Enable passwordless ssh (ssh-keygen)
- ssh csewsX (from any csews*) passwordlessly
- for i in `seq 1 40`; do ssh csews$i uptime ; done
- “Are you sure you want to continue connecting?” yes

```
 10:54:26 up  1:28,  0 users,  load average: 0.01, 0.05, 0.01
 10:54:26 up 12 days, 22:19,  0 users,  load average: 0.01, 0.04, 0.08
 10:54:27 up 16 days, 10:02,  4 users,  load average: 0.01, 0.02, 0.00
ssh: connect to host csews33 port 22: No route to host
 10:54:30 up 12 days, 23:54,  5 users,  load average: 0.06, 0.04, 0.09
The authenticity of host 'csews35 (172.27.19.35)' can't be established.
ECDSA key fingerprint is SHA256:caQgqH23dg1oKLtKAM9ffsJlsNnOPLYl2ZuckbxhEMM.
Are you sure you want to continue connecting (yes/no)? yes
```

Programming

- Shell scripts (e.g. bash)
- ssh basics
  - E.g. ssh –X
  - …
- Mostly in C
- Compilation, Makefiles, ...
- Linux environment variables
  - PATH
  - LD_LIBRARY_PATH
  - …

MPI Installation – Cluster

Install MPICH 3.3 (https://www.mpich.org/static/downloads/3.3/) in your home directory (from any node)

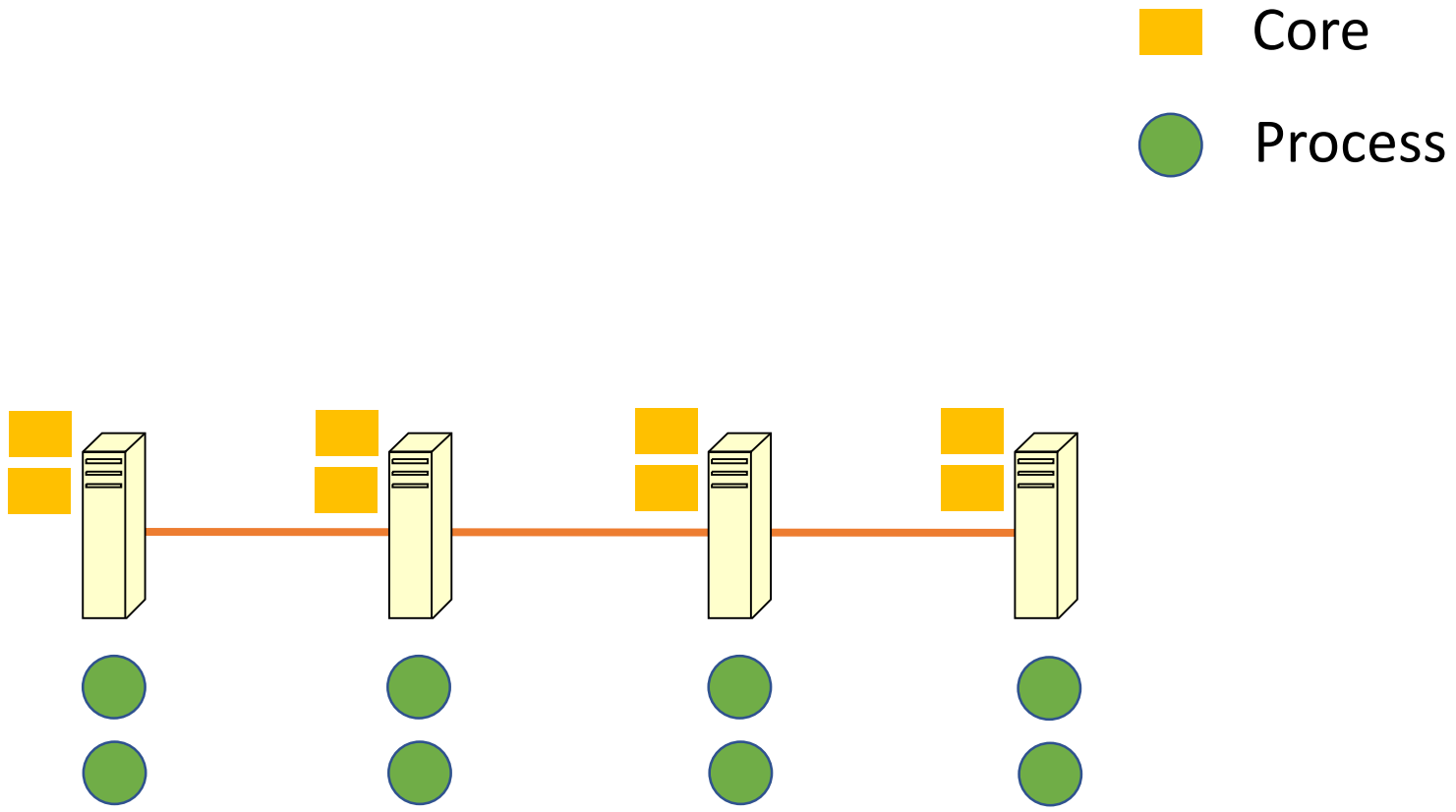
- Download mpich-3.3.tar.gz
- Follow installation instructions from https://www.mpich.org/static/downloads/3.3/mpich-3.3-installguide.pdf
- DO NOT use /tmp
- If mpirun is already installed locally on the system, do not use that node to install
- Verify after installation that `which mpirun` from any node points to your installation

MPI Installation – Laptop

- Linux or Linux VM on Windows
  - apt/yum/brew
- Windows
  - No support

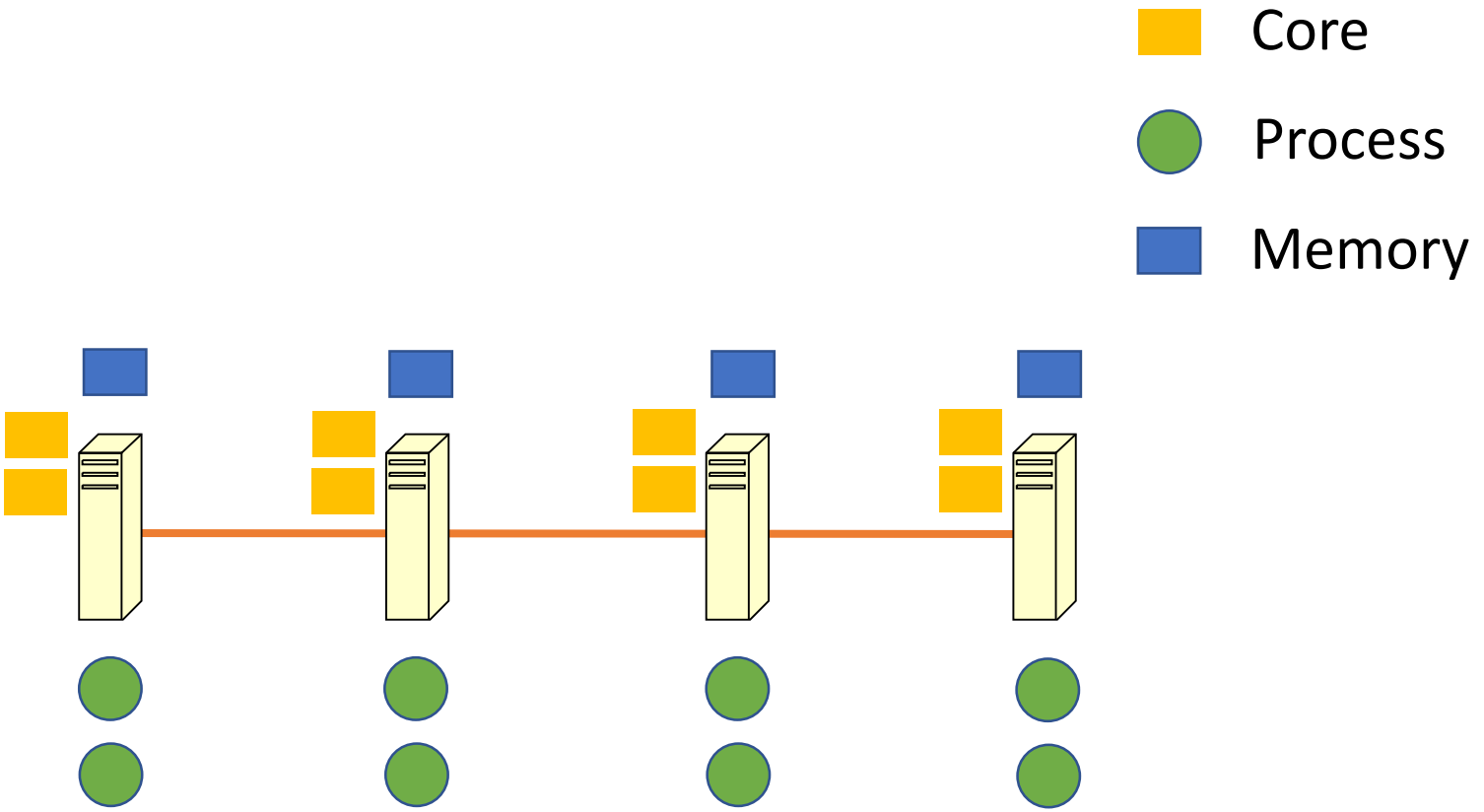
- https://www.mpich.org/documentation/guides/

Our Parallel World



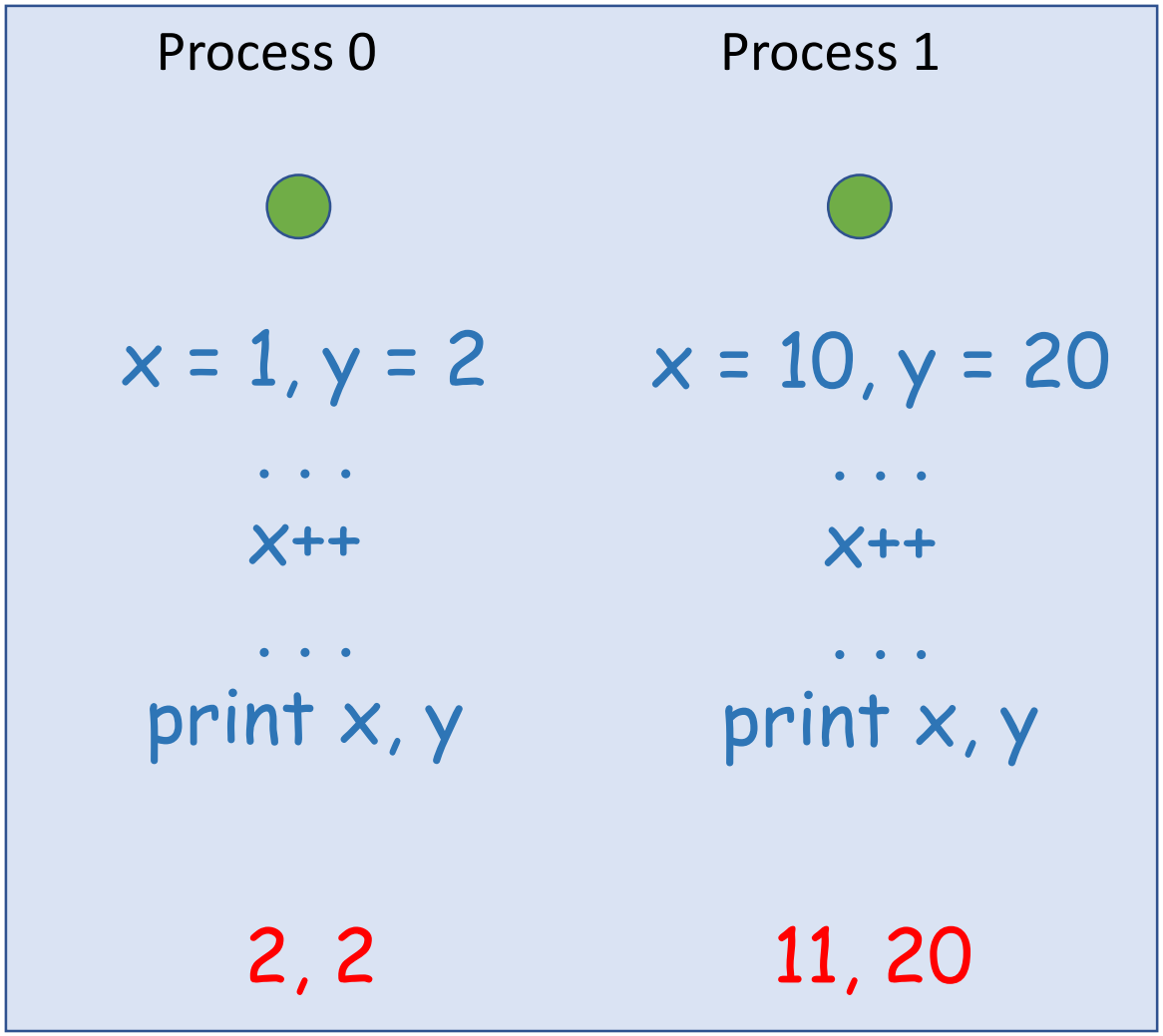
NO centralized server/master

Our Parallel World

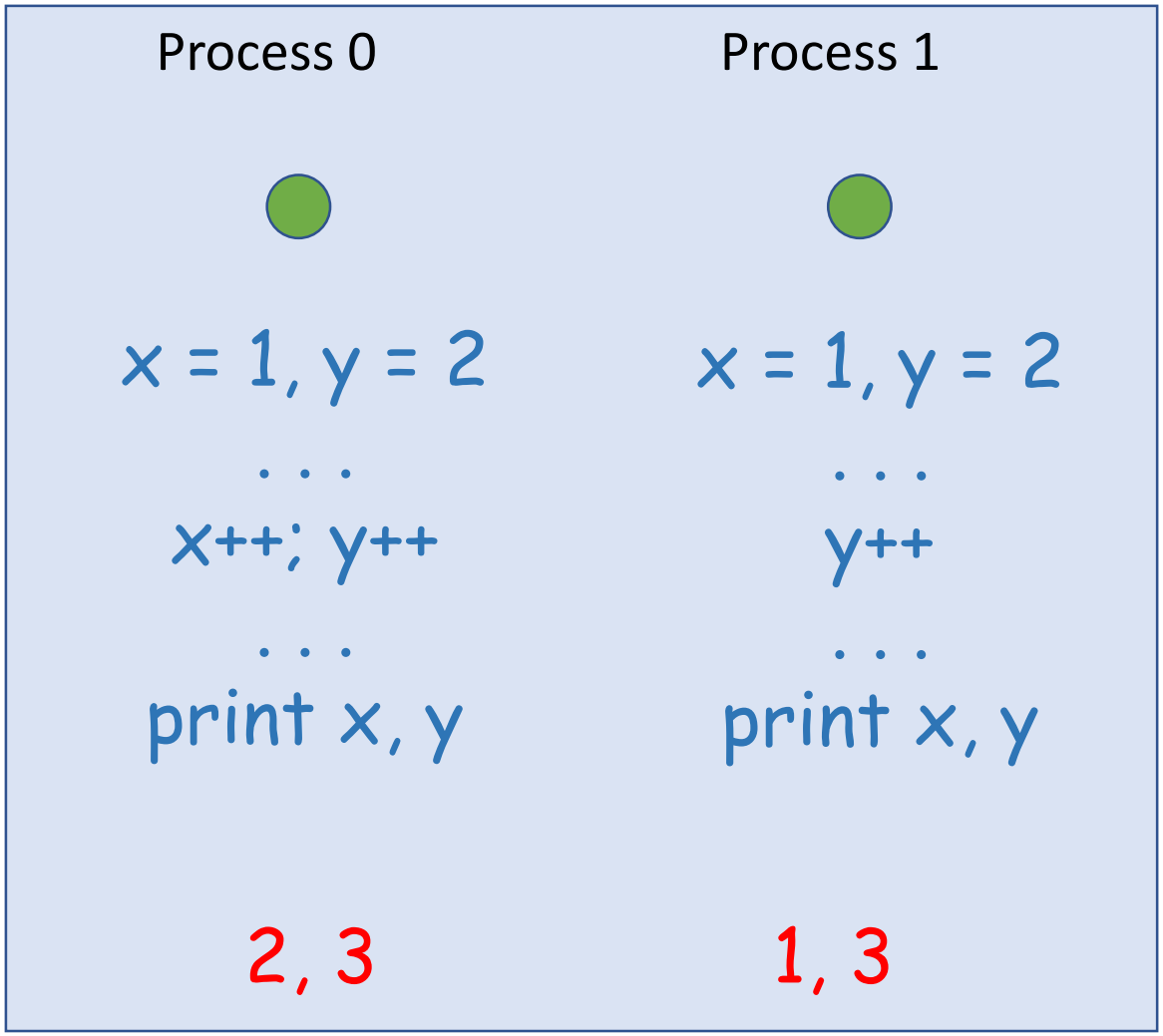


Distributed memory (each process has its own address space)

Distinct Process Address Space



Distinct Process Address Space



MPI

- Standard for message passing
- Explicit communications
- High programming complexity
- Requires communication scope

Simple MPI Code

```
#include <stdio.h>
#include "mpi.h"

int main(int argc, char *argv[])
{
  // initialize MPI
  MPI_Init (&argc, &argv);

  printf ("Hello, world!\n");

  // done with MPI
  MPI_Finalize();
}

~
~
```

MPI Code Execution Steps

- Compile
  - mpicc -o program.x program.c
- Execute
  - mpirun -np 1 ./program.x (mpiexec -np 1 ./program.x)
    - Runs 1 process on the launch node
  - mpirun -np 6 ./program.x
    - Runs 6 processes on the launch node

Execute on Multiple Hosts

MPI process MPI process MPI process MPI process
MPI process MPI process MPI process MPI process

4 nodes, processes per node (ppn)=2

mpiexec –n <number of processes> -f <hostfile> ./exe

Run examples/cpi (mpirun –np 2 <path to examples/cpi>)

<hostfile>

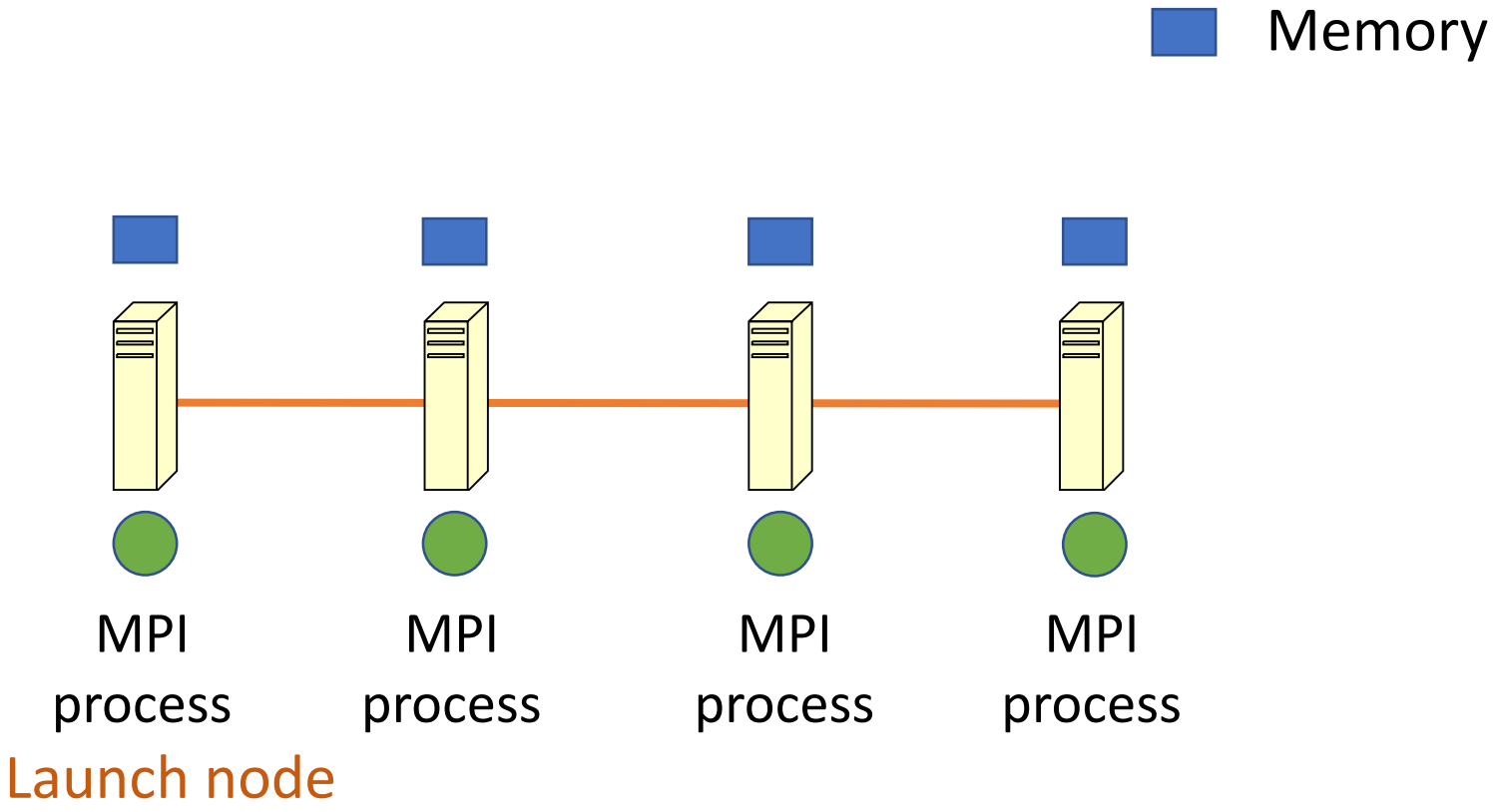
Host1:2
Host2:2
…

DEMO – Hello World

```
pmalakar@csews2:~/class/2020-21-II$ mpirun -np 1 ./1.helloworld
Hello, world!
pmalakar@csews2:~/class/2020-21-II$ mpirun -np 2 ./1.helloworld
Hello, world!
Hello, world!
pmalakar@csews2:~/class/2020-21-II$ mpirun -np 20 ./1.helloworld
Hello, world!
Hello, world!
Hello, world!
Hello, world!
Hello, world!
Hello, world!
Hello, world!
Hello, world!Hello, world!
Hello, world!
Hello, world!
Hello, world!

Hello, world!
Hello, world!
Hello, world!
Hello, world!
Hello, world!
Hello, world!
Hello, world!
Hello, world!
```

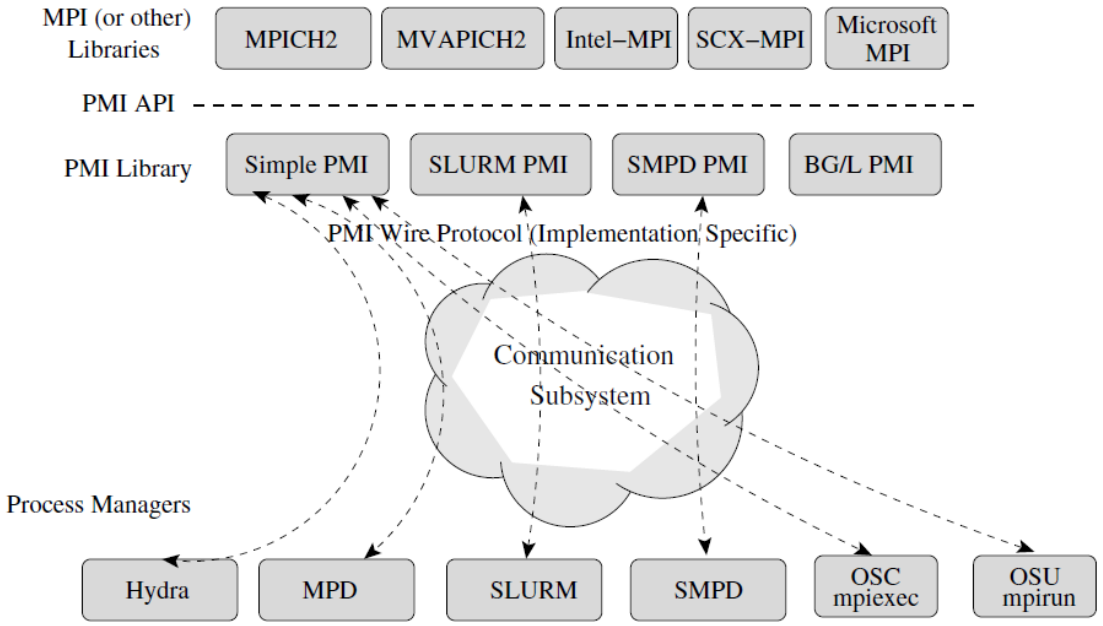
Process Launch



Reading (optional):

A Scalable Process-Management Environment for Parallel Programs

Process Management Setup



Parallel program library (e.g. MPI)

Process management interface (PMI)

Resource manager/ Job scheduler/ Process Manager

Reading (optional):

[PMI: A Scalable Parallel Process-Management Interface for Extreme-Scale Systems]

MPI Internals

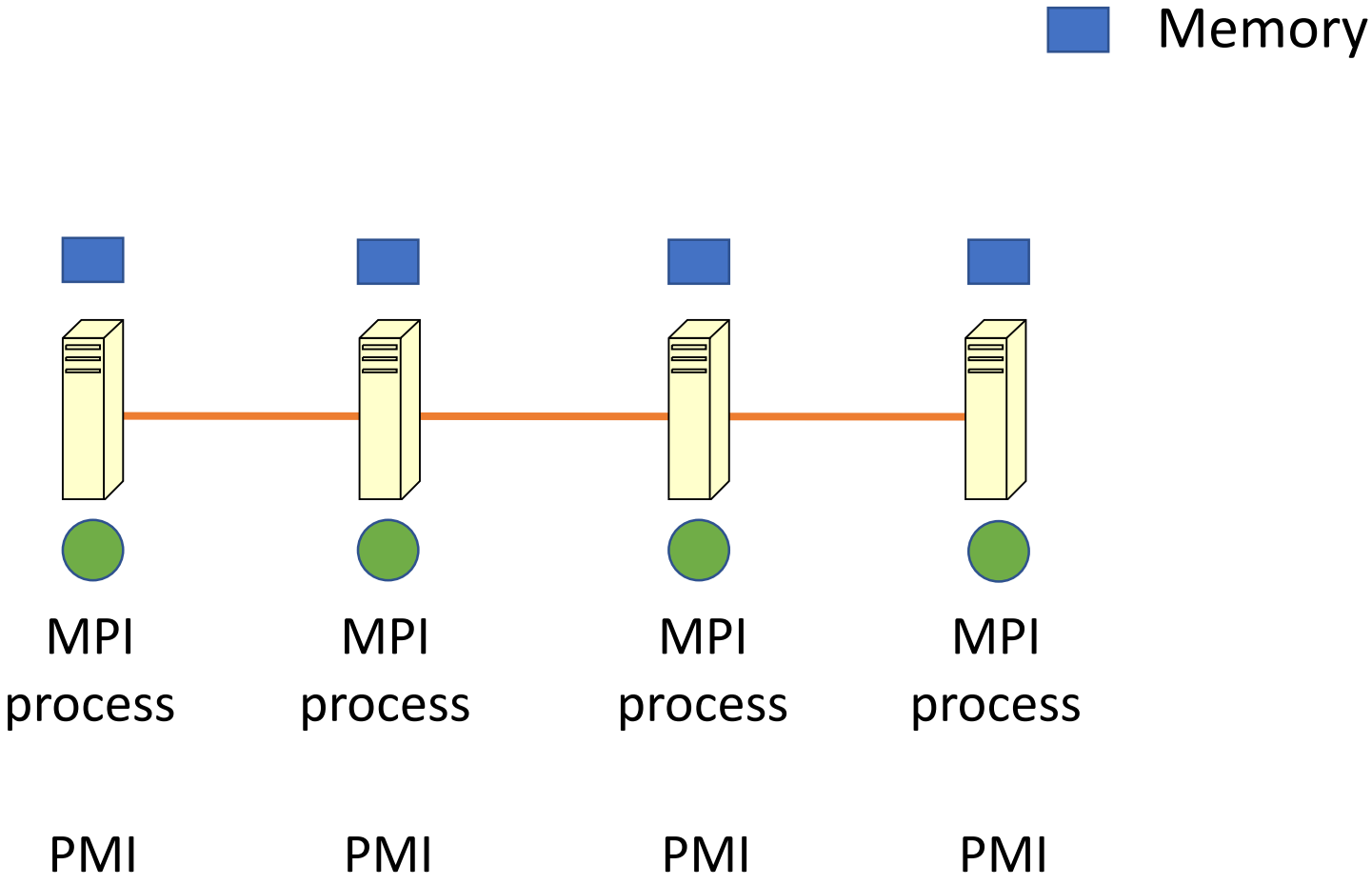
Process Manager

- Start and stop processes in a scalable way
- Setup communication channels for parallel processes

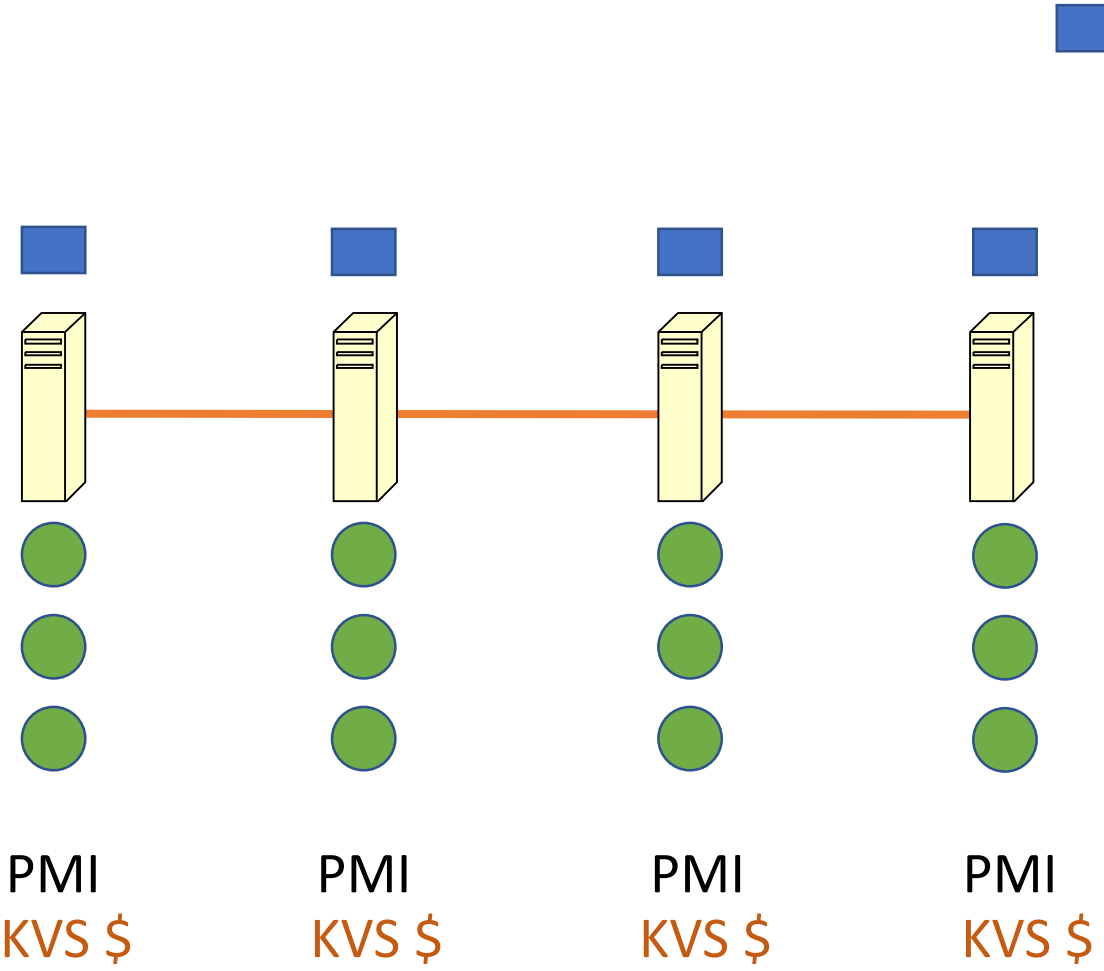
Process Management Interface

- Processes can exchange information about peers by querying PMI
- Provides a logically centralized service for all processes in an MPI job
- Uses key-value store for process-related data

Process Launch



Process Launch

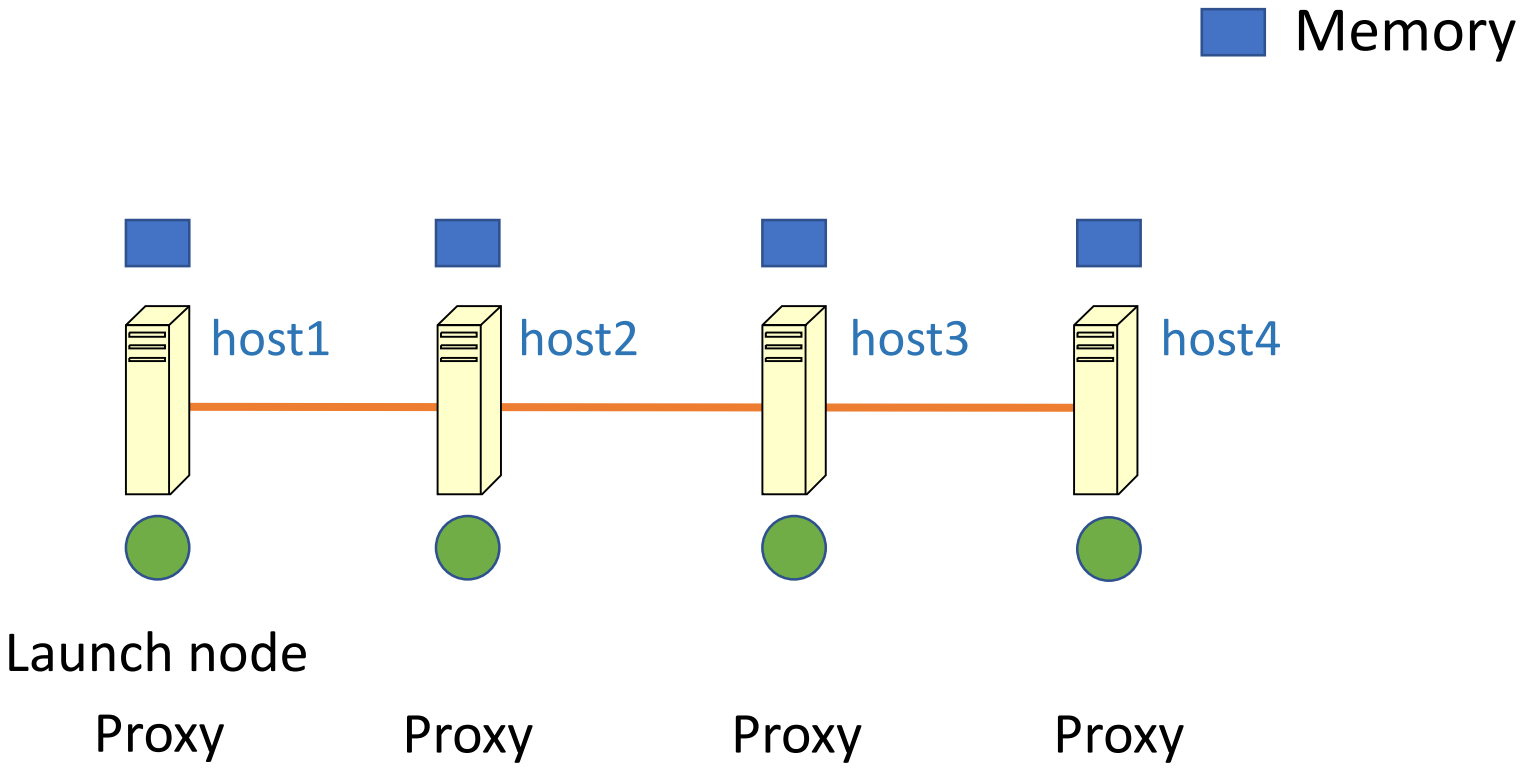


Hydra Process Manager

- A process management system for starting parallel jobs
- Uses existing daemons (viz. ssh) to start MPI processes
- Automatically detects resource managers and interacts with them
- $ mpiexec ./app
  - Hydra gets information about allocated resources and launches processes
- Passes environment variables from the shell on which mpiexec is launched to the launched processes

There are others – gforker, slurm, etc.

mpiexec



mpiexec –n 4 –hosts host1,host2,host3,host4 ./exe

Launch Node

mpiexec -np 8 -hosts host1:3,host2:3,host3:3 ./exe

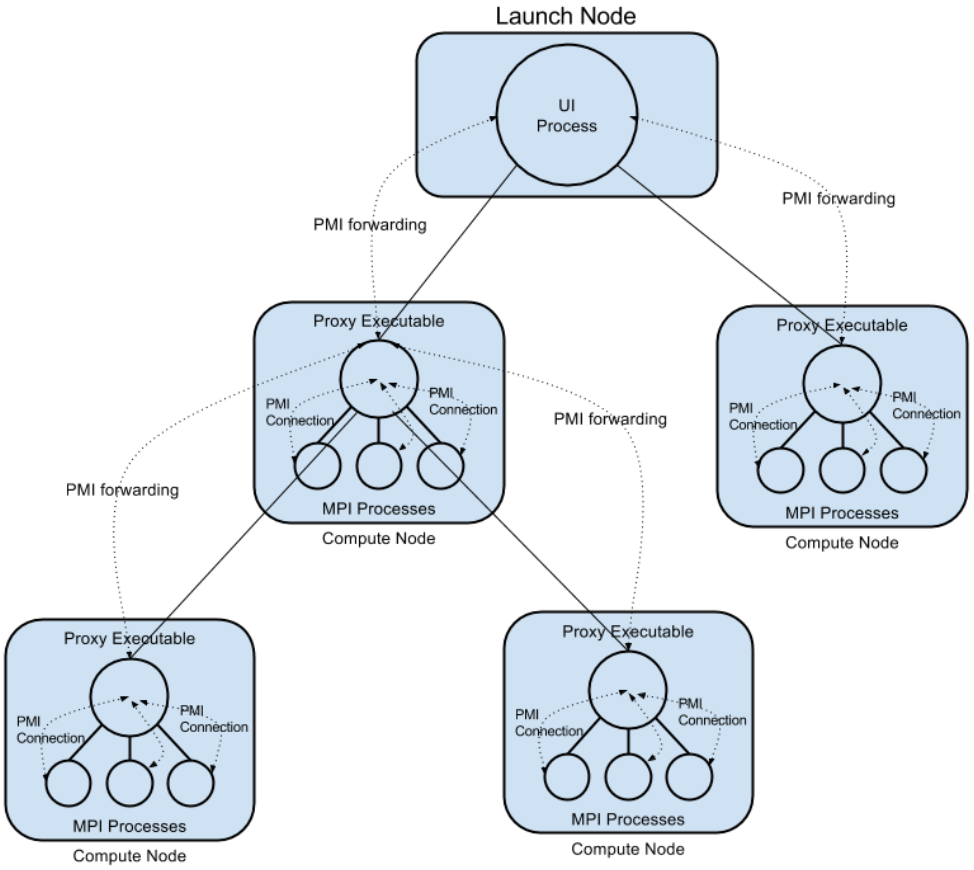
```
pmalakar 17952 17943  0 09:41 ?        00:00:00 /usr/lib/openssh/sftp-server
pmalakar 20853 16203  0 10:20 pts/1    00:00:00 mpiexec -np 8 -hosts 172.27.19.2 3 172.27.19.3 3 172.27
.19.4 3 ./IMB-MPI1 AllReduce
pmalakar 20854 20853  0 10:20 ?        00:00:00 /users/faculty/pmalakar/mpich-3.2.1-install/bin/hydra_p
mi_proxy --control-port 172.27.19.2:46385 --rmk user --launcher ssh --demux poll --pgid 0 --retries 10
--usize -2 --proxy-id 0
pmalakar 20855 20853  0 10:20 ?        00:00:00 /usr/bin/ssh -x 172.27.19.3 "/users/faculty/pmalakar/mp
ich-3.2.1-install/bin/hydra_pmi_proxy" --control-port 172.27.19.2:46385 --rmk user --launcher ssh --dem
ux poll --pgid 0 --retries 10 --usize -2 --proxy-id 1
pmalakar 20856 20853  0 10:20 ?        00:00:00 /usr/bin/ssh -x 172.27.19.4 "/users/faculty/pmalakar/mp
ich-3.2.1-install/bin/hydra_pmi_proxy" --control-port 172.27.19.2:46385 --rmk user --launcher ssh --dem
ux poll --pgid 0 --retries 10 --usize -2 --proxy-id 2
pmalakar 20857 20854 76 10:20 ?        00:00:03 ./IMB-MPI1 AllReduce
pmalakar 20858 20854 76 10:20 ?        00:00:03 ./IMB-MPI1 AllReduce
pmalakar 20859 20854 76 10:20 ?        00:00:03 ./IMB-MPI1 AllReduce
pmalakar 20861 17877  0 10:20 pts/4    00:00:00 ps -aef
```

Compute Node Processes

```
pmalakar  8756  8728  0 10:18 pts/0    00:00:00 -bash
pmalakar  8759  8755  0 10:18 ?        00:00:00 /usr/lib/openssh/sftp-server
root      8781  1123  0 10:20 ?        00:00:00 sshd: pmalakar [priv]
pmalakar  8845  8781  0 10:20 ?        00:00:00 sshd: pmalakar@notty
pmalakar  8846  8845  0 10:20 ?        00:00:00 /users/faculty/pmalakar/mpich-3.2.1-install/bin/hydra_pmi_prox
y --control-port 172.27.19.2:46385 --rmk user --launcher ssh --demux poll --pgid 0 --retries 10 --usize -2 --p
roxy-id 1
pmalakar  8847  8846 99 10:20 ?        00:00:12 ./IMB-MPI1 AllReduce
pmalakar  8848  8846 99 10:20 ?        00:00:12 ./IMB-MPI1 AllReduce
pmalakar  8849  8846 99 10:20 ?        00:00:12 ./IMB-MPI1 AllReduce
```

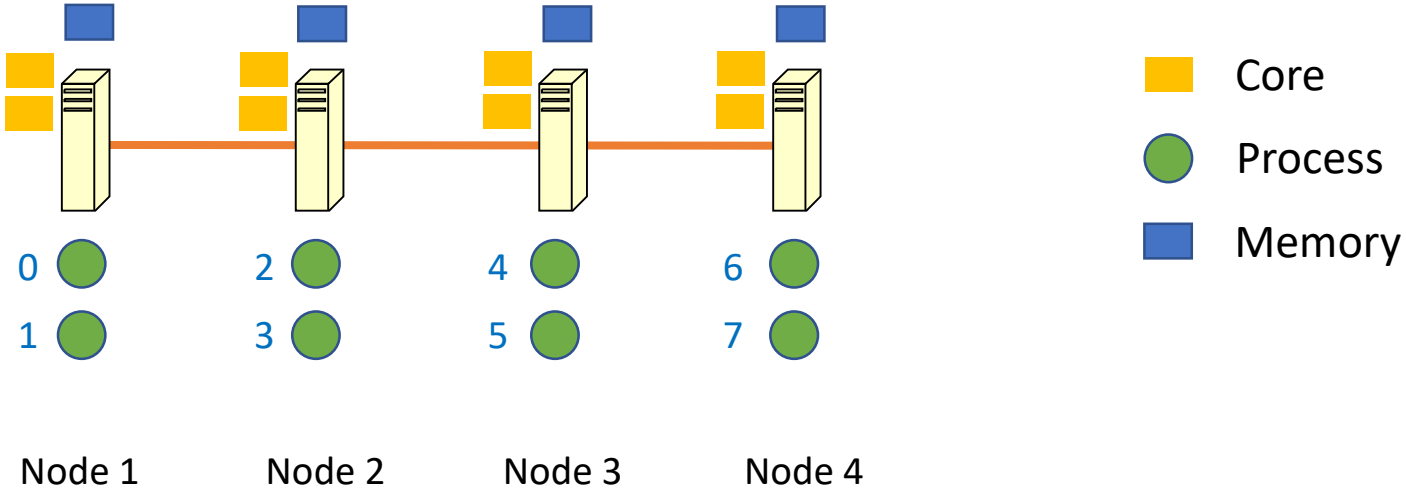
```
pmalakar  8838  8774  0 10:20 pts/1    00:00:00 -bash
pmalakar  8841  8837  0 10:20 ?        00:00:00 /usr/lib/openssh/sftp-server
root      8851  1250  0 10:20 ?        00:00:00 sshd: pmalakar [priv]
pmalakar  8915  8851  0 10:20 ?        00:00:00 sshd: pmalakar@notty
pmalakar  8916  8915  0 10:20 ?        00:00:00 /users/faculty/pmalakar/mpich-3.2.1-install/bin/hydra_p
mi_proxy --control-port 172.27.19.2:46385 --rmk user --launcher ssh --demux poll --pgid 0 --retries 10
--usize -2 --proxy-id 2
pmalakar  8917  8916 99 10:20 ?        00:00:14 ./IMB-MPI1 AllReduce
pmalakar  8918  8916 99 10:20 ?        00:00:14 ./IMB-MPI1 AllReduce
```

Hydra and mpiexec



Source: wiki.mpich.org

Multiple Processes



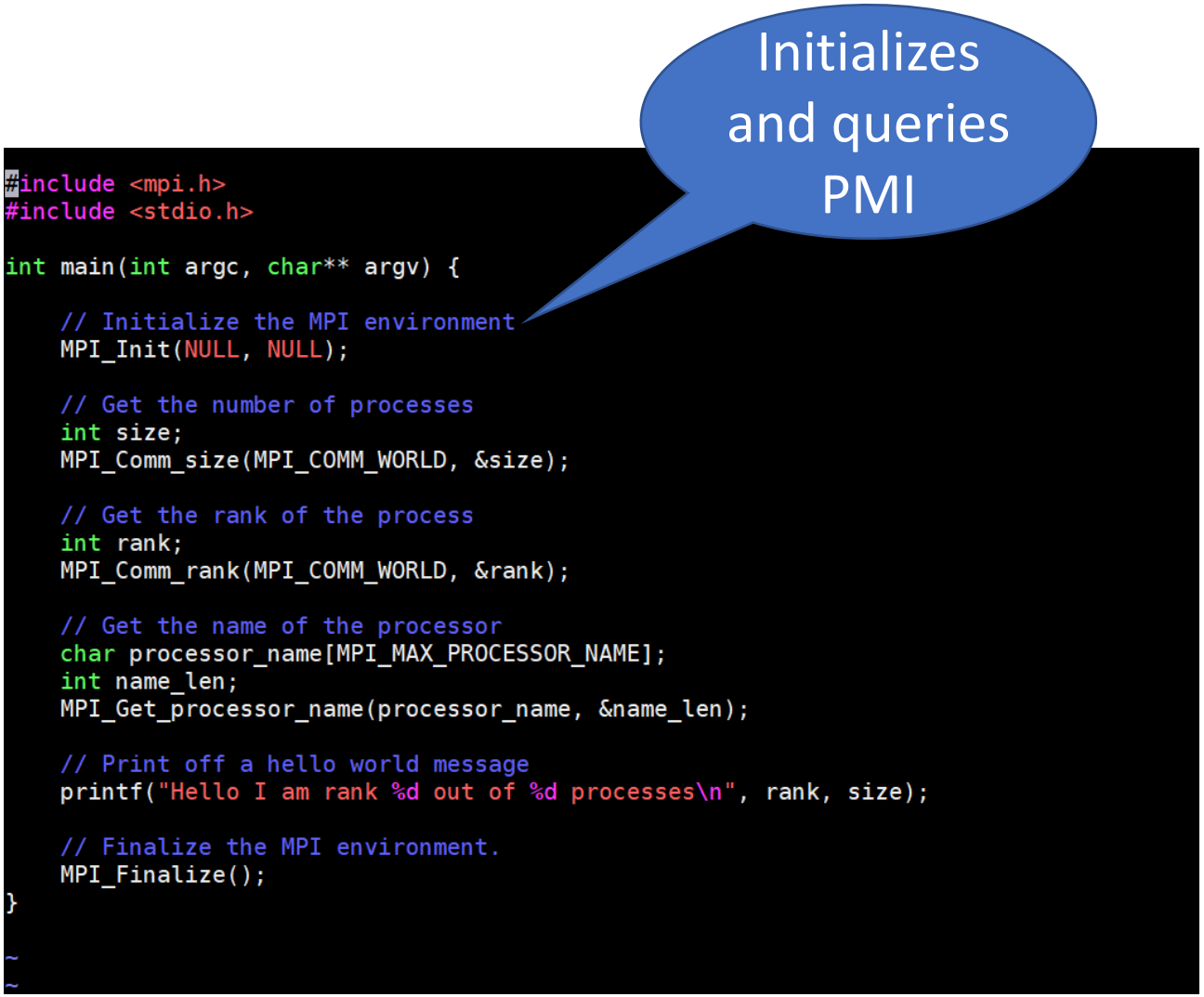
mpiexec –np 8 –f hostfile ./program.x

Communication Channels

- Sockets for network I/O (wire protocol in PMI)
- PMI is responsible for creating/initializing/cleanup
- MPI handles communications, progress etc.

Reading (optional): Design and Evaluation of Nemesis, a Scalable, Low-Latency, Message-Passing Communication Subsystem

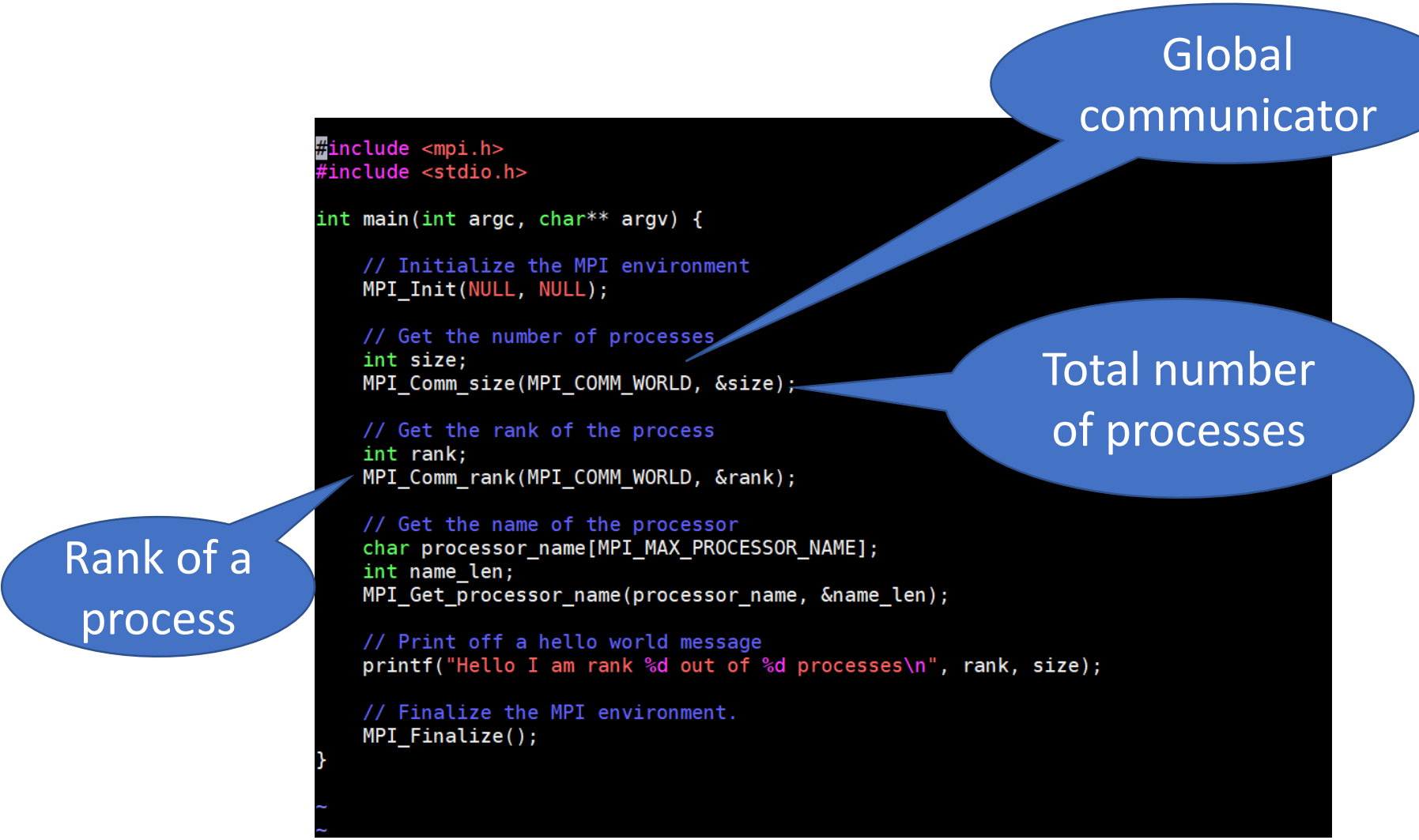
MPI Process Identification



MPI_Init

- gather information about the parallel job
- set up internal library state
- prepare for communication

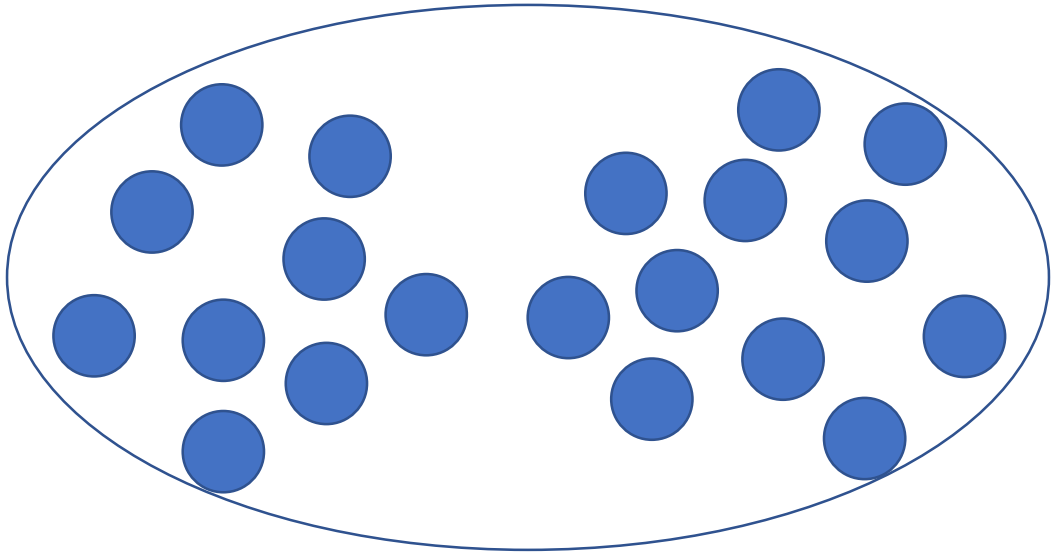
MPI Process Identification



Communicator

- Communication handle among a group/collection of processes
- Representative of communication domain
- Associated with a context ID (in MPICH)
- Predefined:
  - MPI_COMM_WORLD
  - MPI_COMM_SELF

MPI_COMM_WORLD



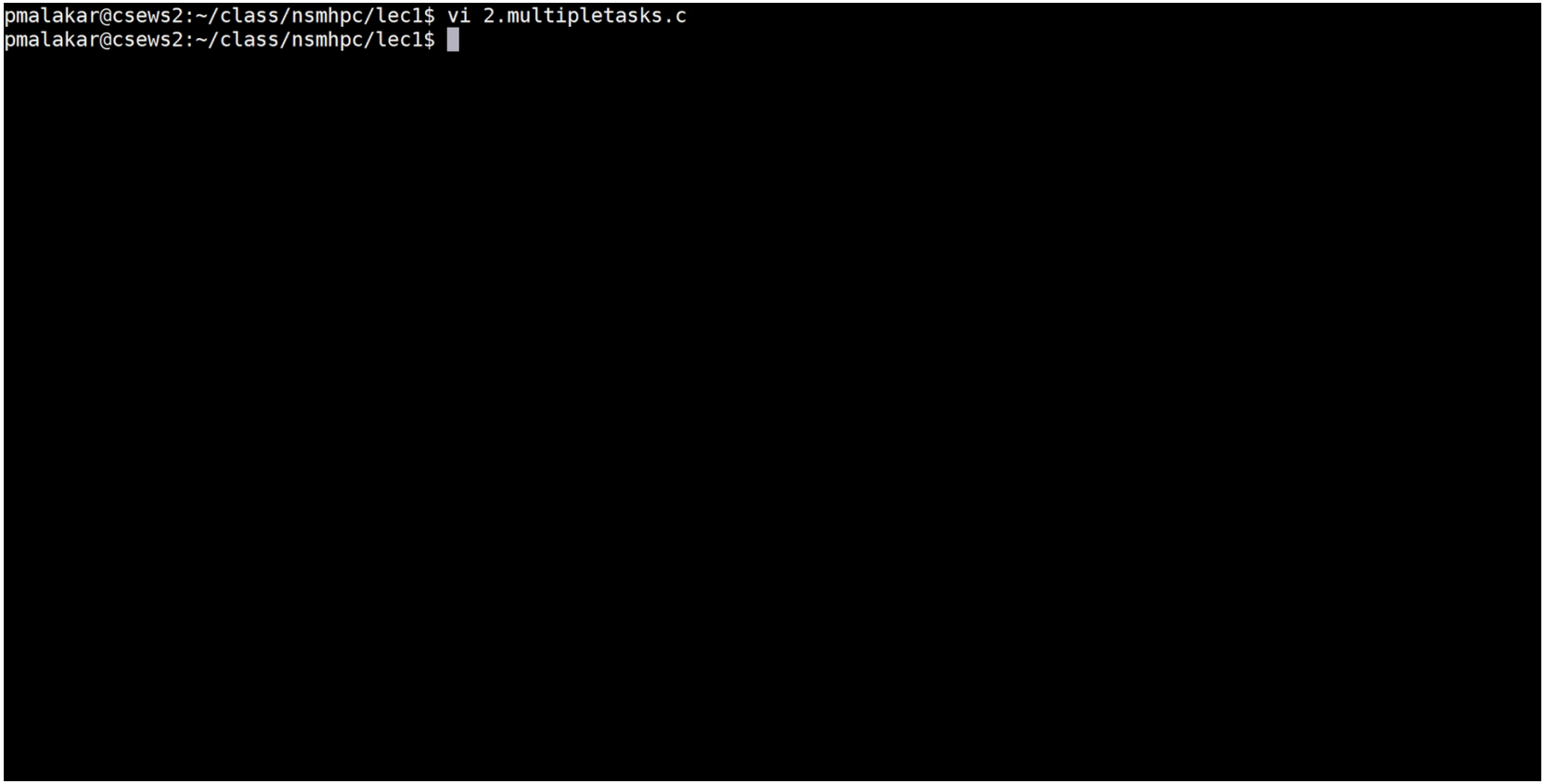
Required in every MPI communication

csews*

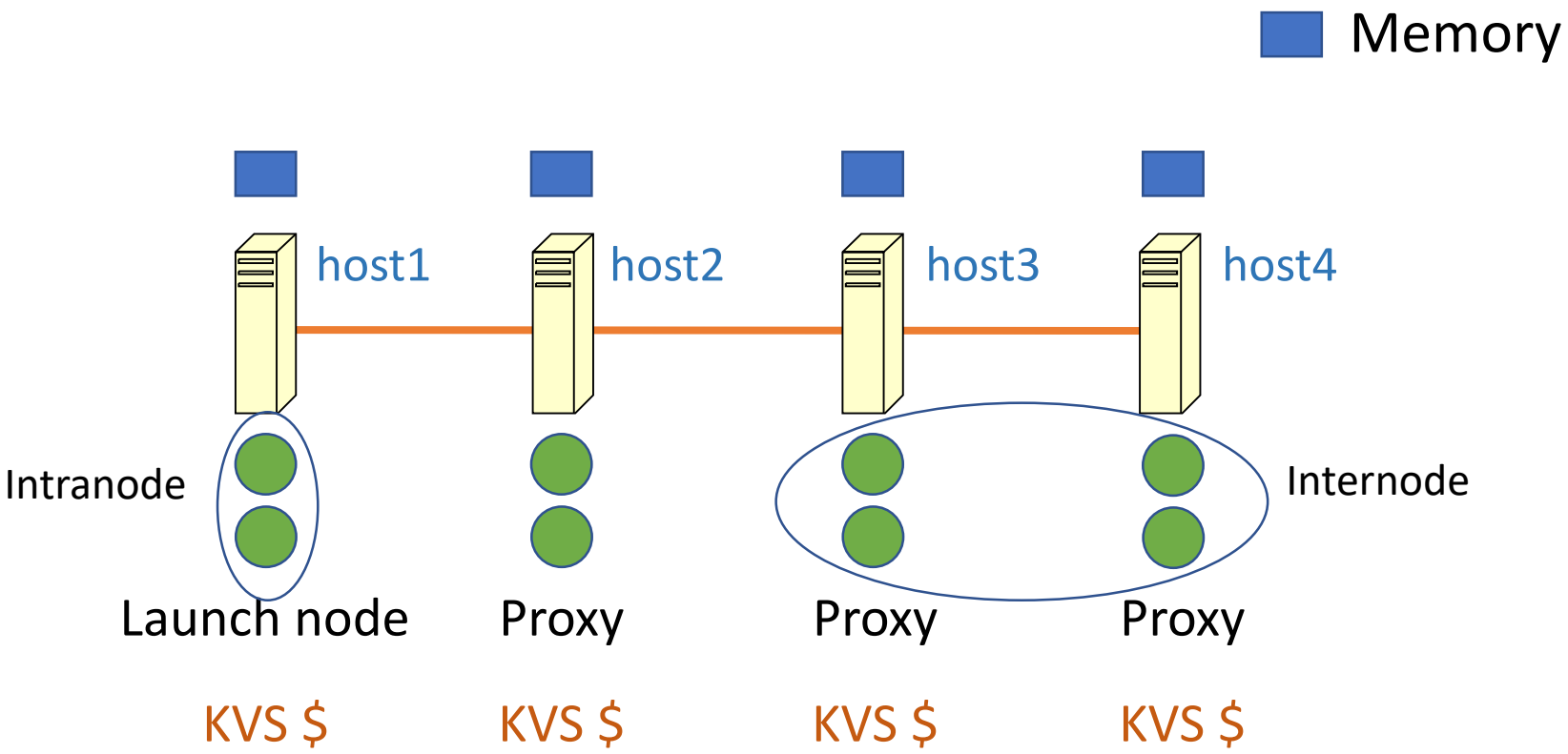
```
pmalakar@csews1:~/class/Aug7$ lscpu
Architecture:        x86_64
CPU op-mode(s):      32-bit, 64-bit
Byte Order:          Little Endian
CPU(s):              12
On-line CPU(s) list: 0-11
Thread(s) per core:  2
Core(s) per socket:  6
Socket(s):           1
NUMA node(s):        1
Vendor ID:           GenuineIntel
CPU family:          6
Model:               158
Model name:          Intel(R) Core(TM) i7-8700 CPU @ 3.20GHz
Stepping:            10
CPU MHz:             900.353
CPU max MHz:         4600.0000
CPU min MHz:         800.0000
BogoMIPS:            6384.00
Virtualization:      VT-x
L1d cache:           32K
L1i cache:           32K
L2 cache:            256K
L3 cache:            12288K
NUMA node0 CPU(s):   0-11
```

```
pmalakar@csews1:~/class/Aug7$ grep processor /proc/cpuinfo
processor       : 0
processor       : 1
processor       : 2
processor       : 3
processor       : 4
processor       : 5
processor       : 6
processor       : 7
processor       : 8
processor       : 9
processor       : 10
processor       : 11
pmalakar@csews1:~/class/Aug7$
```

DEMO – Process Rank



Hydra Process Manager



mpiexec –n 4 –hosts host1,host2,host3,host4 ./exe

MPI Code Execution Options

Same host

- mpirun -np 6 ./program.x

Multiple hosts

- mpirun -np 6 –f hostfile ./program.x
- Round robin process placement
  - mpirun -np 6 –hosts csews1,csews2 ./program.x
- Host-wise placement
  - mpirun -np 6 –hosts csews1:3,csews2:3 ./program.x

DEMO – Process Placement

```
pmalakar@csews2:~/class/2020-21-II$ mpirun -np 4 -hosts csews1,csews2 ./3.multipletaskscoreID | sort -k1n
0 of 4 Running on csews1:2
1 of 4 Running on csews2:6
2 of 4 Running on csews1:10
3 of 4 Running on csews2:10
pmalakar@csews2:~/class/2020-21-II$ mpirun -np 4 -hosts csews1,csews2 ./3.multipletaskscoreID | sort -k1n
0 of 4 Running on csews1:2
1 of 4 Running on csews2:3
2 of 4 Running on csews1:7
3 of 4 Running on csews2:5
pmalakar@csews2:~/class/2020-21-II$ mpirun -np 4 -hosts csews1:2,csews2:2 ./3.multipletaskscoreID | sort -k1n
0 of 4 Running on csews1:8
1 of 4 Running on csews1:10
2 of 4 Running on csews2:2
3 of 4 Running on csews2:5
pmalakar@csews2:~/class/2020-21-II$ mpirun -np 4 -hosts csews1:2,csews2:2 ./3.multipletaskscoreID | sort -k1n
0 of 4 Running on csews1:4
1 of 4 Running on csews1:2
2 of 4 Running on csews2:3
3 of 4 Running on csews2:6
```