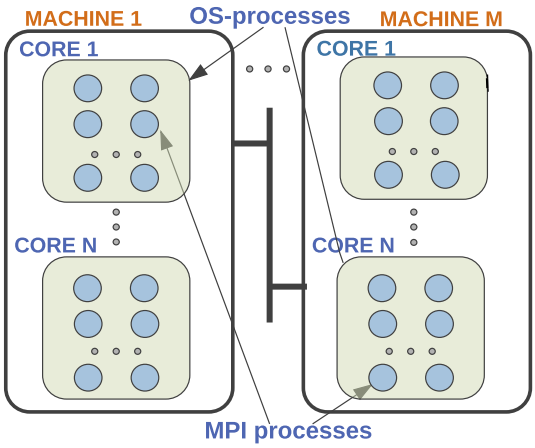
FG-MPI: Fine-Grain MPI

Humaira Kamal and Alan Wagner <kamal, wagner>@cs.ubc.ca
The University of British Columbia
http://www.cs.ubc.ca/~humaira/fgmpi.html

NEW VERSION!

derived from MPICH 3.2

FG-MPI extends the execution model of the Message Passing Interface (MPI) to expose **large-scale, fine-grain concurrency**. **FG-MPI** is integrated into the **MPICH** middleware, a widely used, **production quality** implementation of the MPI standard from the Argonne National Laboratory.



FG-MPI adds an **'-nfg'** flag to mpiexec to specify the number of fine-grain MPI processes inside an OS-process.

Added concurrency is easy to express

```
mpiexec -nfg 1000 -n 4 myprog
```

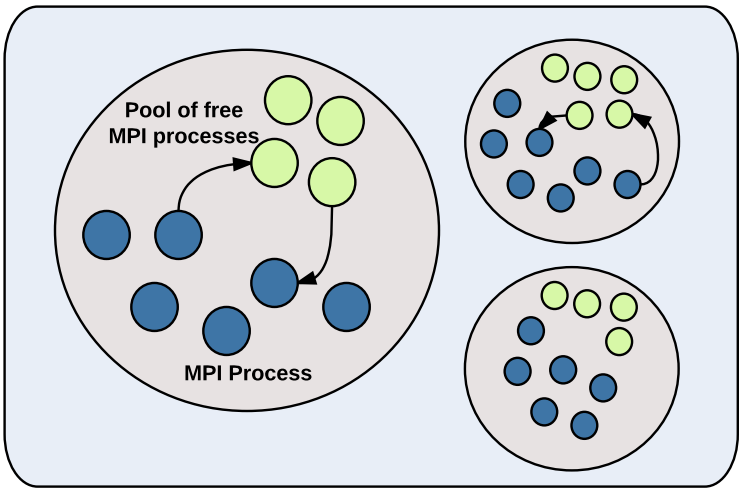
```
mpiexec -nfg 500 -n 8 myprog
```

```
mpiexec -nfg 500 -n 3 myprog:-nfg 250 -n 2 myprog:
        -nfg 1000 -n 2 myprog
```

4000 MPI processes

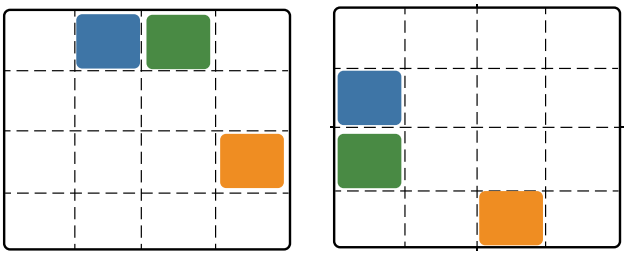
- MPI processes can be mapped to cores, machines, and **now can be mapped to OS processes**.
- Different mappings can be executed **without recompilation.**
- **Seamless execution of hundreds and thousands of MPI processes** on a notebook or cluster.
- **Backwards compatibility** with the existing mpiexec command.
- **Free to mix processes;** mix existing MPI processes with FG-MPI processes.
- **Runs on commodity systems at scale.**

A simpler alternative to dynamic processes



- FG-MPI uses coroutines to implement light-weight MPI processes with **fast context-switching time** and **low communication and synchronization overhead.**
- Can allocate **a pool of light-weight MPI processes** that remain dormant (free) until activated by a message.
- Promotes **dynamic load-balancing** by allowing computation to move to free processes.

Fit working set to cache through '-nfg' flag.



- Achieve **better memory locality** and cache hit ratios simply by using the '-nfg' flag on the command-line for block-structured algorithms.
- Use the '-nfg' flag to tune the working set size to fit the cache **without manually optimizing the code.**
- Gives **portability** without modifying the source code.

Related publications:
Added Concurrency to Improve MPI Performance on Multicore. ICPP 2012.
An Integrated Fine-Grain Runtime System for MPI. *Journal of Computing,* 2014.

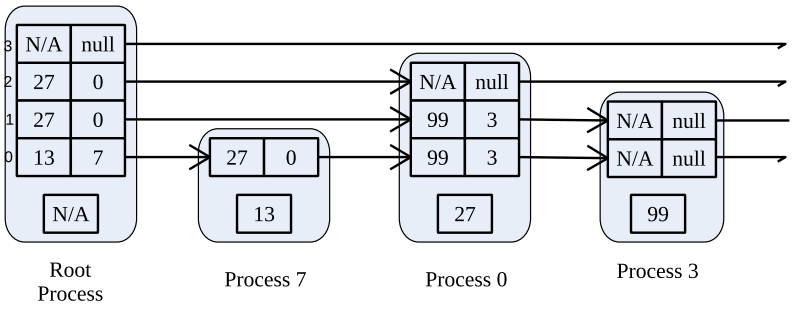
FG-MPI: Fine-Grain MPI

http://www.cs.ubc.ca/~humaira/fgmpi.html

Kamal, Wagner

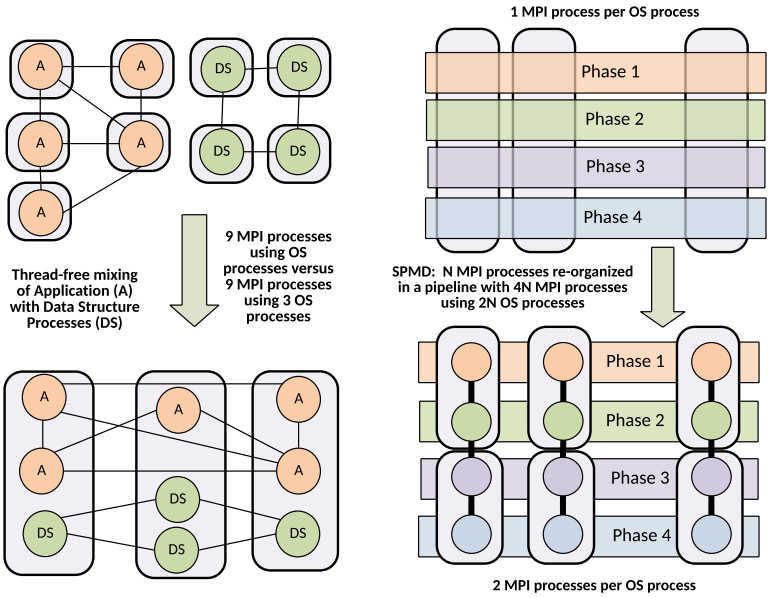
Actor-like process to process pointer structures

- Design novel, scalable **distributed data structures** where each data element is modeled as a process.



Related publications: A Scalable Distributed Skip list for Range Queries, HPDC 2014, A Service-oriented Scalable Dictionary in MPI. CPA 2014.

Use finer grain, to expose more structure, provide more ways to compose processes and allow for more flexible mapping to cores and machines.

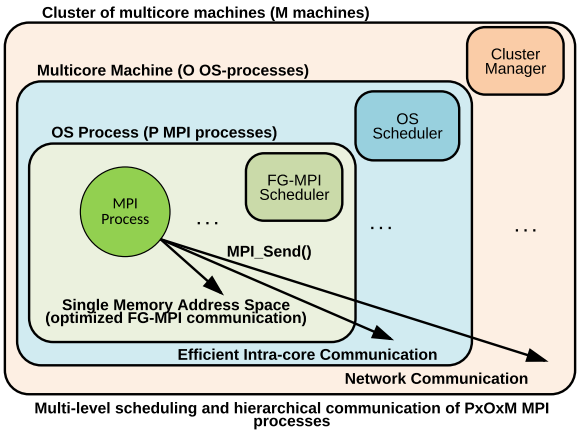


- Enable task-oriented concurrency with simpler support for **MPMD** (Multiple Program Multiple Data) programs.
- Allow the MPI processes inside an OS process to execute **functions instead of main programs**.
- Allow a **process oriented programming approach** that allows one to fit the processes to the problem rather than the machine.
- Provide **a simple API** for the user to specify **more flexible mappings** of MPI processes to functions, and MPI processes to OS processes, cores and machines.

Development features of FG-MPI

- **Develop MPI programs that scale** to hundreds and thousands on their notebooks and workstations.
- Use the integrated runtime scheduler to react to events inside the MPI middleware. There is also the flexibility to **select different runtime schedulers** on the command line.
- Use a deterministic process scheduler (e.g. round robin) **to debug and test** programs.
- Run all MPI processes inside a single OS-process **to detect program safety issues** like deadlock.
- **Easy porting** of many MPI programs to FG-MPI through the addition of a small bit of boiler-plate code.

Hierarchical, Location-aware Communication



- Exploits locality of MPI processes for optimized communication within the same OS-process.
- Leverages and extends MPICH hierarchical communication algorithms within a node and across nodes.
- FG-MPI has been used on a compute-cluster to execute exascale number of processes.

https://www.westgrid.ca/westgrid_news/2013-01-14/ubc_researchers_use_westgrid_explore_exascale_computing

We acknowledge the support of Intel Inc., Mitacs Canada and NSERC for the FG-MPI project.