multi-processor jobs in DIRAC WMS

P. Gay (U. Bordeaux), A. Tsaregorodtsev (CNRS-IN2P3-CPPM)

Introduction

- DIRAC terminology:
  - Taken from F. Stagni https://docs.google.com/presentation/d/19FmrwDNStQmdNaQGhMt2f_8uyHZMFjUJK5lSNoOXkb0
  - “payload job”: user application job
  - “multi-processor job”: payload application will try to use multiple cores on the same node
  - “computing slot”: resource allocated by a provider where a pilot wrapper is running (batch job)
  - “multi-processor slot”: allocated resource has more than one OS CPU core available in the same slot

What we needed

- Users need to run *multi-threaded / multi-process* applications in *multi-core* worker nodes through DIRAC WMS
  - Multi-processor enabled payload jobs that need to break the 24 hours elapsed limit of the computing slot
  - High memory usage (> 1GB) can also lead to the need of multi-processor computing slots or even “whole-node” allocation
- How to tell DIRAC to submit jobs to the underlying (grid CE) middleware with correct requirements ?
  - We need the resource CE to allocate the correct number of queue slots on the same node
  - This feature was available on CREAM-CE (and others) for some time

What we did

- Matcher (A. Tsaregorodtsev)
  - Handle number of processors requirement as mandatory tag in TaskQueueDB, in the same way as RAM requirements (Github PR #2529)
  - This is where all the magic is done
- JobAgent
  - Fix Timeleft utility to be able to understand multi-processor environment (Github PR #2680)
  - JobAgent needs to match *only* multi-processor job payloads when inside a multi-processor computing slot (Github PR #2694)
- SiteDirector
  - New agent: MultiProcessorSiteDirector (Github PR #2823)
  - Restricted to handle multi-processor task queues in order not to conflict with existing SiteDirector agent (can be easily changed if new agent is widely adopted)
  - Groups task queues with same multi-processor requirements
  - Submits pilots to multi-processor enabled CEs with correct requirements for each different group
  - JobAgents running in these computing slots are configured with an option instructing on their multi-processor context

How to use it

- Get a v6r15 DIRAC WMS
- Configure some CEs to accept multi-processor jobs in the DIRAC CS
  - /Resources/.../CEs/<ce-hostname>/MaxProcessors = <integer>
  - /Resources/.../CEs/<ce-hostname>/WholeNode = <True | False>
  - Can be set at queue level too
- Run a MultiProcessorSiteDirector agent (don't touch your already running SiteDirectors)
- Add requirements in the “Tags” field of your job JDLs
  - <X> processors: `Tags = {“<X>Processors”}; #(e.g. “12Processors”)`
  - Whole node: `Tags = {"WholeNode"};`
- Inside payload jobs, you can access to multi-processor capabilities through environment variables
  - DIRAC_PROCESSORS (integer)
  - DIRAC_WHOLENODE (“True” | ”False”)

What we could do in the future

- Multi-processor pilot submission is for CREAM-CE only → extend to other compatible CE types
- Apply this machinery to Clouds/VM
- Add useful multi-processor information in accounting system (number of processors, whole-node, // efficiency)
- Automatic multi-processor CS configuration with detection on CEs capabilities?
- Merge back SiteDirector and MultiProcessorSiteDirector codes
- Think about how to extend this mechanism to *multiple nodes* jobs to expose HPC resources (similar but not identical to DIRAC MPI service)
- Document it, add tests...

What (I think) we shouldn't do

- Try to mix different payload requirements in the same walltime limited MP computing slot
  - This would require implementing a complex scheduling algorithm
  - Could dramatically reduce computing slot efficiency even if done carefully
  - Cloud based computing slots may be different on this aspect?

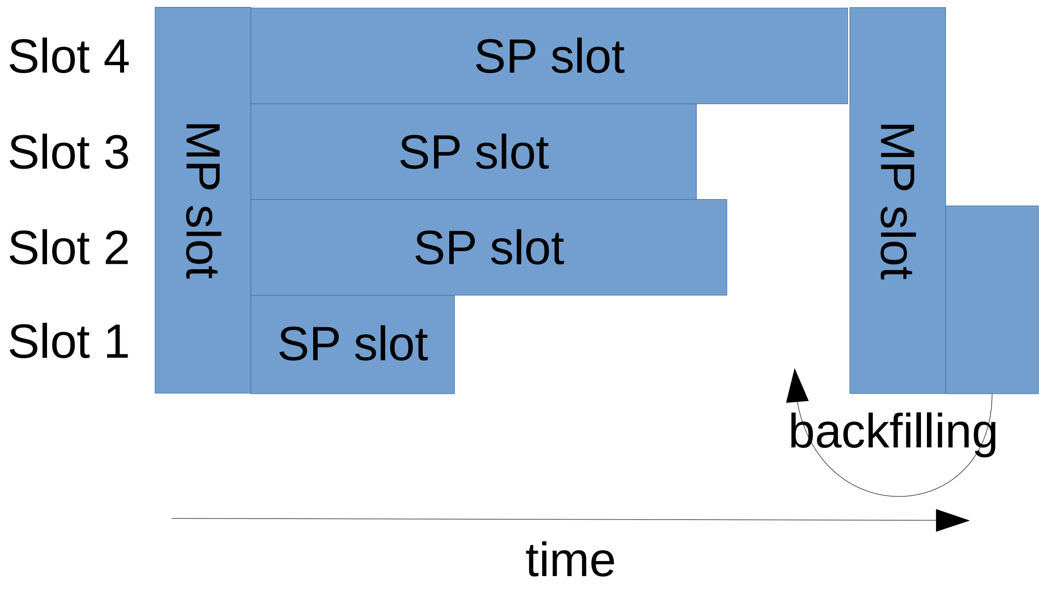
Conclusion

- A basic multi-processor support is available in DIRAC v6r15 for CREAM-CE resources
  - Requires running a separate MultiProcessorSiteDirector agent
  - Needs feedback
- Thanks to:
  - France Grilles' FG-DIRAC for development / testing setup
  - A. Tsaregorodtsev and V. Hamar for being so patient with me

Questions?

Backup slides

MP scheduling efficiency, backfilling


Slot 4
Slot 3
Slot 2
Slot 1
MP slot
SP slot
SP slot
SP slot
SP slot
MP slot
backfilling
time