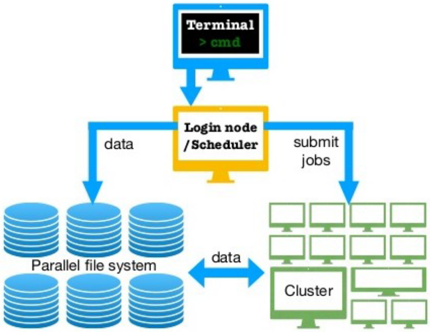
SCALABLE COMPUTING LABORATORY
VANDERBILT UNIVERSITY

SC3260 / SC5260

Batch Scheduler

Lecture by: Ana Gainaru

Today



- HPC system middleware
  - Distributed operating system
    - Memory management, processes and communication management
  - Parallel file system
    - Access performance, resiliency, security
  - Scheduler
  - Daemons on compute nodes
    - Performance monitoring, fault tolerance

VANDERBILT
UNIVERSITY

Today

Batch Scheduling

- From the **user's perspective**
  - Submission principles
  - Performance
- From the **system's perspective**
  - Principles
  - Brief theoretical resutls
  - Currently used schedulers
  - How good is a schedule?

VANDERBILT UNIVERSITY

Recap

Why are schedulers needed?

Recap

Why are schedulers needed?

- Performance
- Fairness (every user wants to be on a dedicated machine)

Performance / Fairness

From the system's perspective

Administrators want to keep the system utilized

- Utilization (max) : percentage of the CPU time that is spent computing
- Power consumption (min)
- User fairness : give space on the machine to all users

VANDERBILT UNIVERSITY

Performance / Fairness

From the system's perspective

Administrators want to keep the system utilized

- Utilization (max) : percentage of the CPU time that is spent computing
- Power consumption (min)
- User fairness : give space on the machine to all users

From the user's perspective

Users want their job to compute as fast as possible

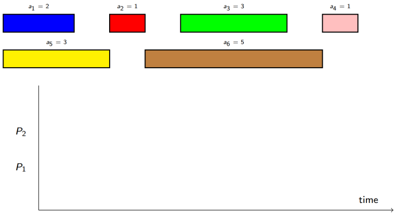
- Makespan (min) : time to complete the job from start to end
- Response time (min) : time to complete the job from submission to end
- Stretch (min) : ration between the response time and the ideal execution time

VANDERBILT UNIVERSITY

Scheduling policies

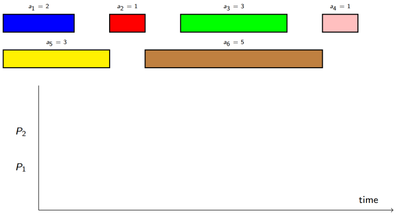
The scheduler can be used to balance all the metrics

- User fairness
- System utilization
- Application response time

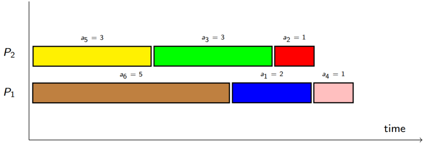


The scheduler can be used to balance all the metrics

- User fairness
- System utilization
- Application response time



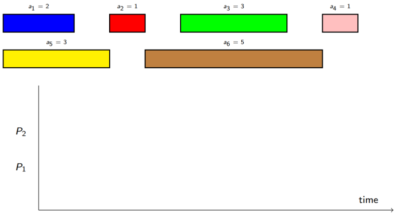
Longest job first



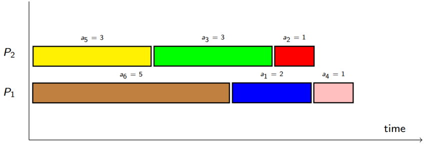
Scheduling policies

The scheduler can be used to balance all the metrics

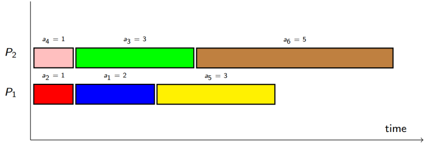
- User fairness
- System utilization
- Application response time



Longest job first



Shortest job first



VANDERBILT UNIVERSITY

Batch Schedulers

- Applications in HPC systems are run as batch jobs, i.e. time-limited requests for resources to run the application binaries.
- Once an application is submitted on a cluster it becomes a job
- Each job is defined as a **Number of nodes** (p_i) and a **Time** (t_i)
  `I want 6 nodes for 1h`

Typically users are charged against an allocation: e.g. "You only get 100 CPU hours per week"

A batch scheduler is a central middleware to manage resources (e.g. processors)

- accept jobs (computing tasks) submitted by users
- decide when and where jobs are executed
- start jobs execution

VANDERBILT UNIVERSITY

Batch Schedulers

Schedulers take into account:

- unavailability of some nodes
- users jobs mutual exclusion
- specific needs for jobs (memory, network, ...)

While trying to :

- maximize resources usage
- be fair among users

To run multiple applications concurrently, **HPC schedulers order the execution of batch jobs** to achieve high utilization while controlling their turnaround times

VANDERBILT UNIVERSITY

Batch Schedulers

Typical wanted features:

- Interactive mode
- Batch mode
- Parallel jobs support
- Multi-queues with priorities
- Reservations
- Admission policies (limit on usage, notions of user groups)
- Resources matching
- File staging
- Jobs dependences
- Backfilling
- Environment reconfiguration

There are many existing batch schedulers: Slurm, LSF, Moab, PBS/Torque, EASY, OAR, ...

These are complex systems with many config options !

VANDERBILT UNIVERSITY

Batch Scheduler Components

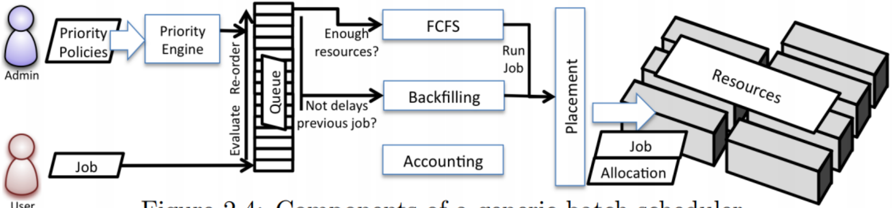


Figure 2.4: Components of a generic batch scheduler.

Life-cycle of a batch job

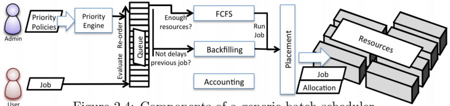


Figure 2.4: Components of a generic batch scheduler.

1) Job submission to the system

The job submission must provide

- Detailed specification of the requested resources (e.g. the number of cores, minimum RAM per core, or specific compute nodes to run on)
- An estimate of the job's runtime
- A priority request (expressing the job's importance)
- Optionally, a list of dependencies on other jobs (e.g. statements that the job should not start until some set of conditions is met)

VANDERBILT UNIVERSITY

Life-cycle of a batch job

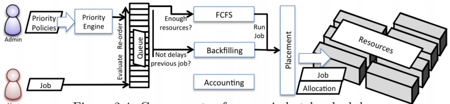


Figure 2.4: Components of a generic batch scheduler.

2) The scheduler contacts the resource manager

- If no other jobs are waiting and there are enough resources available, the scheduler runs the job immediately
- If there are holes in a schedule that would fit the current job, the job is ran immediately (backfilling)
- Otherwise, the job is appended to a job waiting queue

VANDERBILT UNIVERSITY

Life-cycle of a batch job

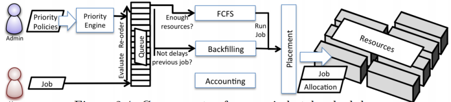


Figure 2.4: Components of a generic batch scheduler.

3) Jobs placed in the waiting queue

- Jobs in the waiting queue are initially ordered by arrival time
- Jobs are ranked and re-ordered by a priority engine
- Different ranking policies define priorities, based on
  - job size (e.g. smaller jobs should run sooner)
  - priority class (e.g. jobs in the real time class should run before any other job)
  - fairness (i.e. priorities dictated by system quotas)
  - wait time in the queue (jobs that have been waiting a long time to start)
  - other administrator-defined criteria

Life-cycle of a batch job

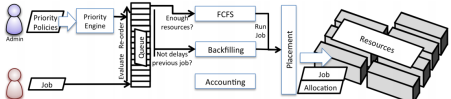


Figure 2.4: Components of a generic batch scheduler.

4) Jobs are allocated on the compute nodes

- Jobs progress towards the top area of the waiting queue until they are extracted by the scheduling algorithms and then executed
- Online or reservation-based placement decision
- Scheduler informs the resource manager about the new placement

Most HPC batch schedulers include the FCFS (First Come First Served) and backfilling scheduling algorithms with different priority re-ordering

Online / Reservation-based

There are usually many jobs in the queue waiting for resources to become available

Online Scheduler

- When a job finishes, the scheduler chooses the first job in the queue to execute that fits the available resources
- To make sure that large jobs do not starve, the scheduler divided all jobs in the queue in batches
- **Advantage** Easy to implement, fast, the resource requests of jobs don't need to be accurate
- **Disadvantage** Local optimal execution, not the best utilization nor makespan

VANDERBILT UNIVERSITY

Online / Reservation-based

There are usually many jobs in the queue waiting for resources to become available

Reservation-based Scheduler

- On job arrival and when a job finishes, the scheduler computes tentative start times for all (most of) the jobs in the queue in order to maximize utilization. **These start times are called reservations**
- Jobs start within their assigned reservations
- **Advantage** Gives the best job placements, fair and starvation free algorithm
- **Disadvantage** More complex and slower (cut of in the waiting queue), resource requests must reflect resource usage

VANDERBILT UNIVERSITY

There are usually many jobs in the queue waiting for resources to become available

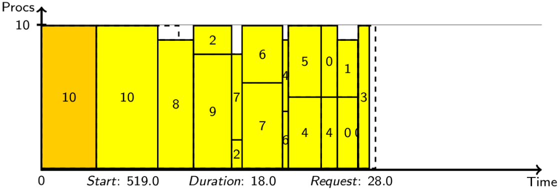
Reservation-based Scheduler

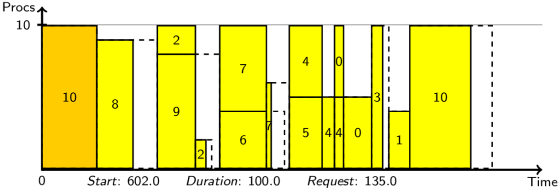
- On job arrival and when a job finishes, the scheduler computes tentative start times for all (most of) the jobs in the queue in order to maximize utilization. **These start times are called reservations**
- Jobs start within their assigned reservations
- **Advantage** Gives the best job placements, fair and starvation free algorithm
- **Disadvantage** More complex and slower (cut of in the waiting queue), resource requests must reflect resource usage

Most schedulers are reservation-based using priority queues and backfilling

VANDERBILT UNIVERSITY

Online / Reservation-based





Placement of 11 jobs using online or reservation-based strategies.

- The reservations are computed only during job arrival and not job ending

Stochastic jobs will get better results from online schedulers

VANDERBILT UNIVERSITY

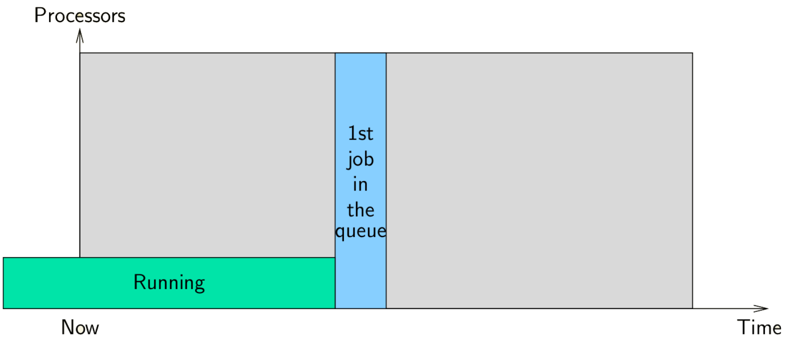
Online / Reservation-based

- FCFS = simplest scheduling option
- **Fragmentation** = need for backfilling

VANDERBILT UNIVERSITY

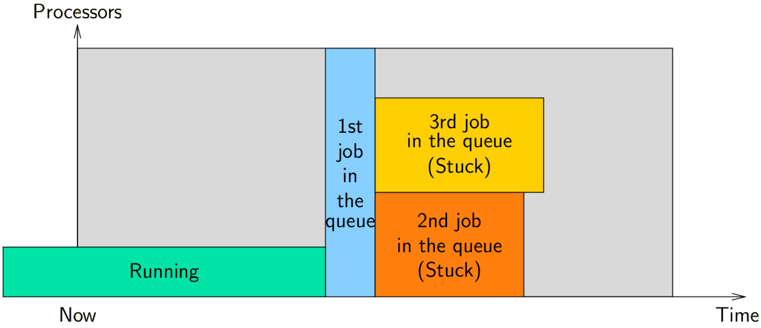
Online / Reservation-based

- FCFS = simplest scheduling option
- **Fragmentation** = need for backfilling



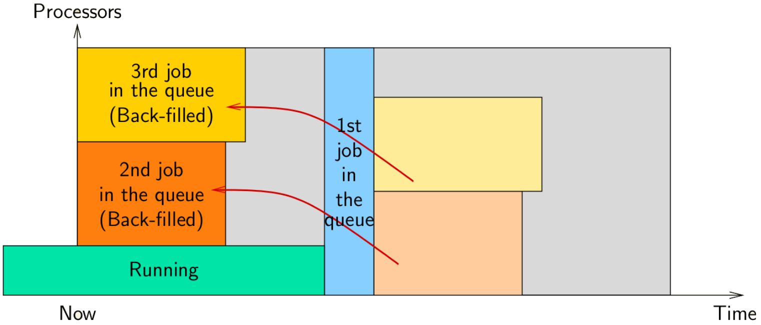
Online / Reservation-based

- FCFS = simplest scheduling option
- Fragmentation = need for backfilling



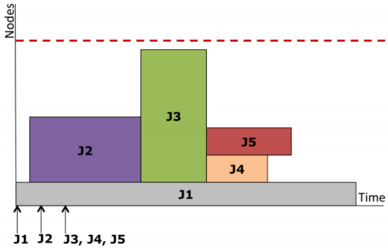
Online / Reservation-based

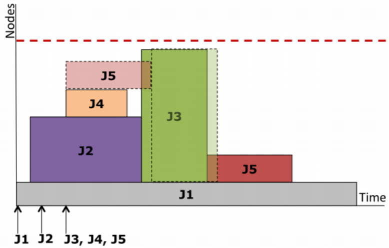
- FCFS = simplest scheduling option
- **Fragmentation** = need for backfilling



Online / Reservation-based

- FCFS = simplest scheduling option
- **Fragmentation** = need for backfilling





Additional functions

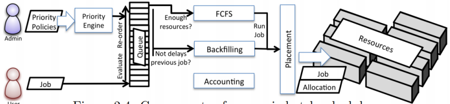


Figure 2.4: Components of a generic batch scheduler.

Mechanisms that are needed to manage an HPC system

- Placement systems that calculate which resources should be used for specific jobs.
  - These decisions take into account the network topology or special job requirements
  - Example: in a system with a fat-tree interconnect topology, a tightly coupled application will run faster if all its assigned nodes are leaves pending from same network switch
- Workload managers include functions to handle the basic operations to run an HPC system, such as managing the compute resources, staging-in jobs, controlling their execution, and staging-out resources

VANDERBILT UNIVERSITY

Additional functions

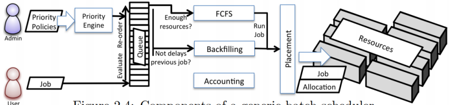


Figure 2.4: Components of a generic batch scheduler.

HPC Accounting for registering the use of compute hours and resources by user jobs

- Prevent users from utilizing the system beyond their assigned quota (e.g. by de-prioritizing their jobs)
- Encourage those who have not used it (e.g. by elevating the priority of users with little quota usage)

Backfilling policies

Which job(s) should be picked for promotion through the queue?

Backfilling policies

Which job(s) should be picked for promotion through the queue?

- Many heuristics are possible
- Two have been studied in detail
  - EASY
  - Conservative Back Filling (CBF)
- In practice EASY is used in almost all current schedulers
- The OAR scheduler (used by french clusters) uses CBF

EASY Backfilling

Extensible Argonne Scheduling System

Maintain only one reservation, for the first job in the queue.

Definitions:

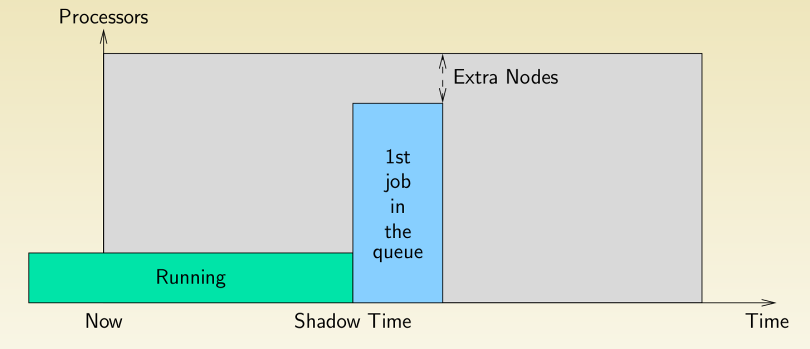
- **Shadow time** time at which the first job in the queue starts execution
- **Extra nodes** number of nodes idle when the first job in the queue starts execution

Policy

1. Go through the queue in order starting with the 2nd job.
2. Backfill a job if it will terminate by the shadow time, or it needs less than the extra nodes.

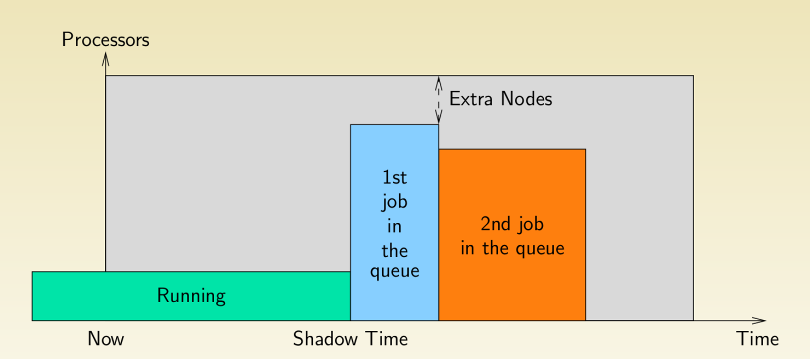
VANDERBILT UNIVERSITY

EASY



Property

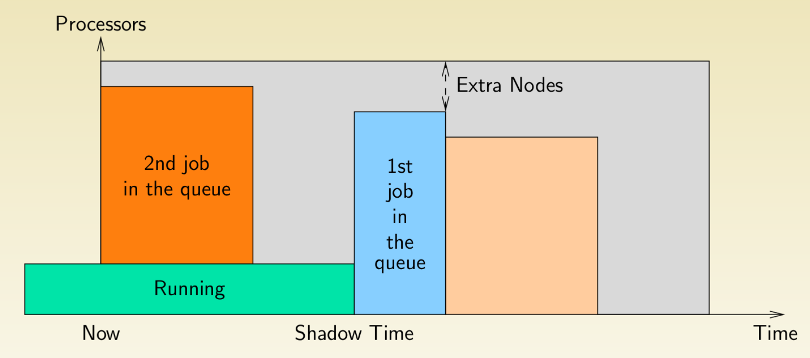
- The first job in the queue will never be delayed by backfilled jobs



Property

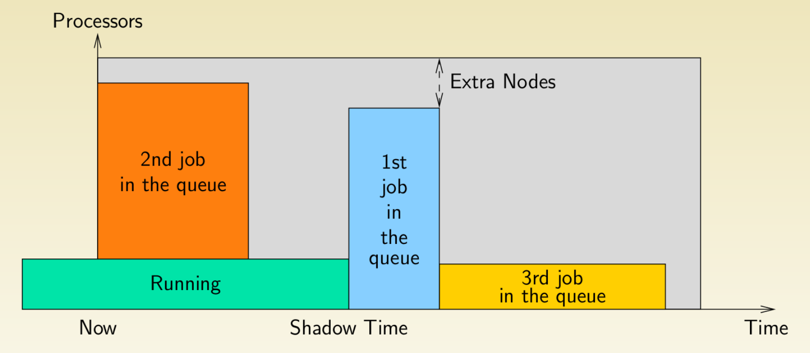
- The first job in the queue will never be delayed by backfilled jobs

EASY



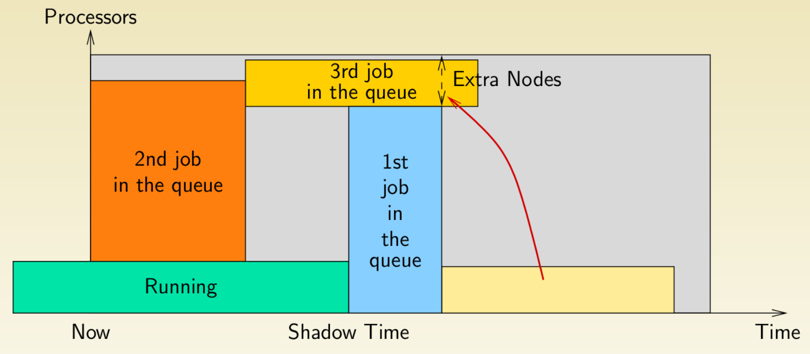
Property

- The first job in the queue will never be delayed by backfilled jobs



Property

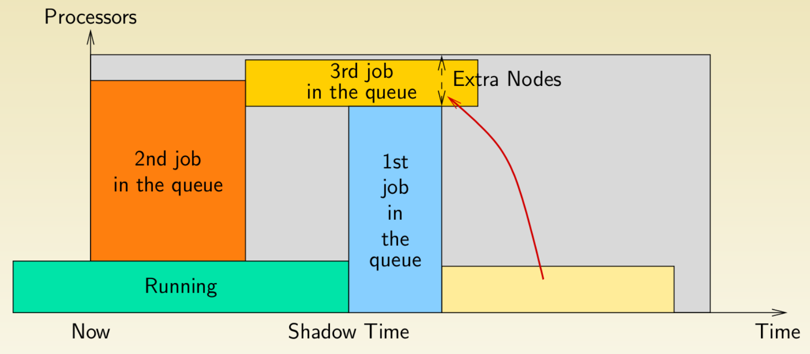
- The first job in the queue will never be delayed by backfilled jobs



Property

- The first job in the queue will never be delayed by backfilled jobs

EASY

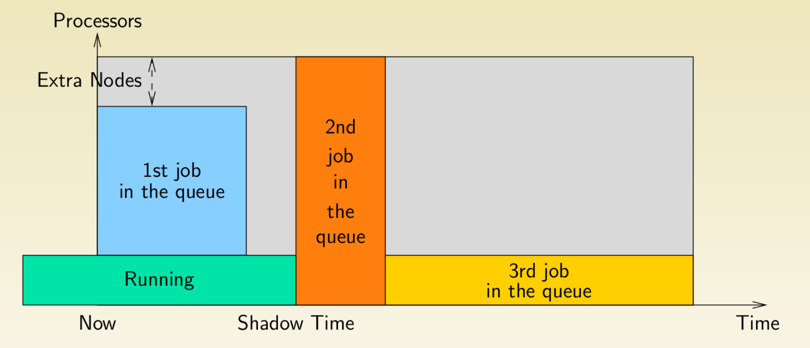


Property

- The first job in the queue will never be delayed by backfilled jobs
- BUT, other jobs may be delayed infinitely!

VANDERBILT UNIVERSITY

EASY

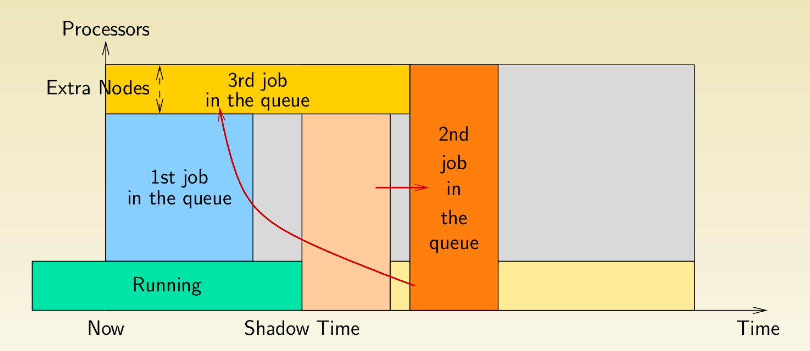


Property

- The first job in the queue will never be delayed by backfilled jobs
- BUT, other jobs may be delayed infinitely!

VANDERBILT UNIVERSITY

EASY

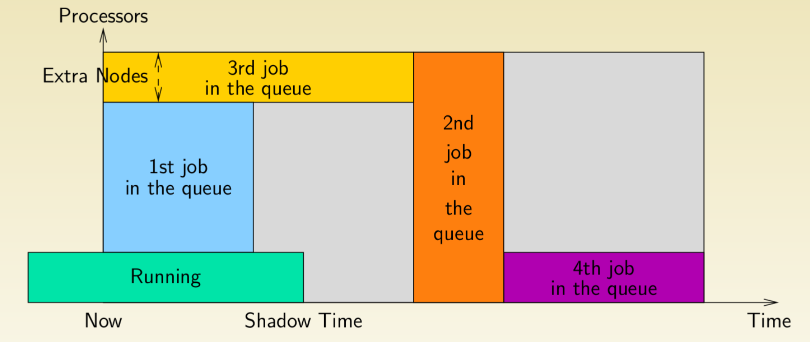


Property

- The first job in the queue will never be delayed by backfilled jobs
- BUT, other jobs may be delayed infinitely!

VANDERBILT UNIVERSITY

EASY

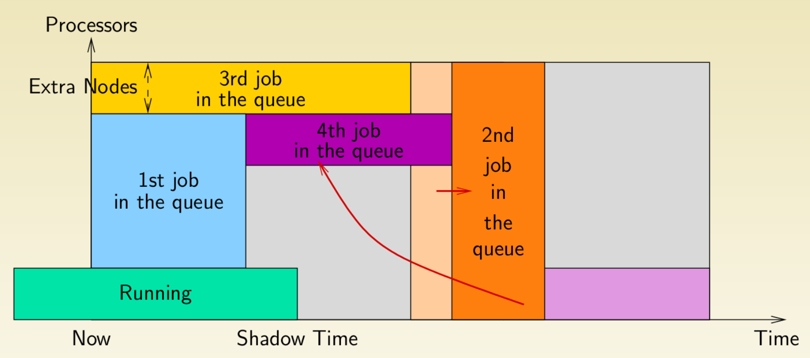


Property

- The first job in the queue will never be delayed by backfilled jobs
- BUT, other jobs may be delayed infinitely!

VANDERBILT UNIVERSITY

EASY



Property

- The first job in the queue will never be delayed by backfilled jobs
- BUT, other jobs may be delayed infinitely!

VANDERBILT UNIVERSITY

EASY Properties

Unbounded Delay

- The first job in the queue will never be de- layed by backfilled jobs
- BUT, other jobs may be delayed infinitely!

No starvation

- Delay of first job is bounded by runtime of current jobs
- When the first job finishes, the second job becomes the first job in the queue
- Once it is the first job, it cannot be delayed further

VANDERBILT UNIVERSITY

Other backfilling approach

Conservative Backfilling

- EVERY job has a reservation. A job may be backfilled only if it does not delay any other job ahead of it in the queue
- Fixes the unbounded delay problem that EASY has. More complicated to implement (The algorithm must find holes in the schedule) though.
- EASY favors small long jobs and harms large short jobs.

VANDERBILT UNIVERSITY

How Good is the Schedule?

When does backfilling happen?

Possibly when

- A new job arrives
- The first job in the queue starts
- When a job finishes early

When does backfilling happen?

Possibly when

- A new job arrives
- The first job in the queue starts
- When a job finishes early

Users provide job **runtime estimates** (Jobs are killed if they go over). Trade-off:

- provide a **conservative estimate**: goes through the queue faster (may be backfilled)
- provide a **loose estimate**: your job will not be killed

VANDERBILT UNIVERSITY

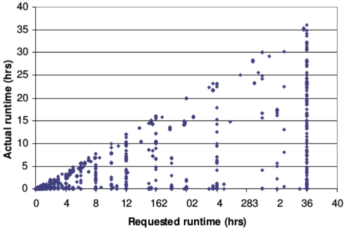
When does backfilling happen?

Possibly when

- A new job arrives
- The first job in the queue starts
- When a job finishes early

Users provide job runtime estimates (Jobs are killed if they go over). Trade-off:

- provide a **conservative estimate**: goes through the queue faster (may be backfilled)
- provide a **loose estimate**: your job will not be killed



Measure performance

... but how do we know what a "good" schedule is? FCFS, EASY, CFB, Random?

What we need are metrics to quantify how good a schedule is. It has to be an aggregate metric over all jobs

Measure performance

... **but how do we know what a "good" schedule is?** FCFS, EASY, CFB, Random?

What we need are metrics to quantify how good a schedule is. It has to be an aggregate metric over all jobs

1. **Turn-around time or flow (Wait time + Run time)**
   Job 1 needs 1h of compute time and waits 1s
   Job 2 needs 1s of compute time and waits 1h
   Clearly Job 1 is really happy, and Job 2 is not happy at all

VANDERBILT UNIVERSITY

Measure performance

... but how do we know what a "good" schedule is? FCFS, EASY, CFB, Random?

What we need are metrics to quantify how good a schedule is. It has to be an aggregate metric over all jobs

1. **Turn-around time or flow (Wait time + Run time)**
   Job 1 needs 1h of compute time and waits 1s
   Job 2 needs 1s of compute time and waits 1h
   Clearly Job 1 is really happy, and Job 2 is not happy at all

2. **Wait time** (equivalent to "user happiness")
   Job 1 asks for 1 nodes and waits 1 h
   Job 2 asks for 512 nodes and waits 1h
   Again, Job 1 is unhappy while Job 2 is probably sort of happy.

We need a metric that represents happiness for small, large, short, long jobs

VANDERBILT UNIVERSITY

Measure performance

- **Slowdown or Stretch** (turn-around time divided by turn-around time if alone in the system)
  Doesn't really take care of the small/large problem.
  Could think of some scaling, but unclear !

For now this is all we have

VANDERBILT UNIVERSITY

Measure performance

- **Slowdown or Stretch** (turn-around time divided by turn-around time if alone in the system)
  Doesn't really take care of the small/large problem.
  Could think of some scaling, but unclear !

For now this is all we have We can run simulations of the scheduling algorithms, and see how they fare. We need to test these algorithms in representative scenarios Supercomputer/cluster traces. Collect the following for long periods of time:

- Time of submission
- How many nodes asked
- How much time asked
- How much time was actually used
- How much time spent in the queue

Measure performance

Example experiment: replace user estimate by f times the actual run time Possible to improve performance by multiplying user estimates by 2!

	EASY	CBF
Mean Slowdown		
KTH	-4.8%	-23.0%
CTC	-7.9%	-18.0%
SDSC	+4.6%	-14.2%
Mean Response time		
KTH	-3.3%	-7.0%
CTC	-0.9%	-1.6%
SDSC	-1.6%	-10.9%

VANDERBILT UNIVERSITY

Performance of Schedulers

- All the schedulers presented are all heuristics
  - They are not specifically designed to optimize the metrics we have designed
- It is difficult to truly understand the reasons for the results.
- But one can derive some empirical wisdom.

- One of the reasons why one is stuck with possibly obscure heuristics is that we're dealing with an on-line problem
- We cannot wait for all jobs to be submitted to make a decision
- But we can wait for a while, accumulate jobs, and schedule them together.

VANDERBILT UNIVERSITY

Summary

Batch Schedulers are what we're stuck with at the moment

They are often hated by users

- I submit to the queue asking for 10 nodes for 1 hour.
- I wait for two days.
- My code finally starts, but doesn't finish within 1 hour and gets killed!!

VANDERBILT UNIVERSITY

Summary

Batch Schedulers are what we're stuck with at the moment

They are often hated by users

- I submit to the queue asking for 10 nodes for 1 hour.
- I wait for two days.
- My code finally starts, but doesn't finish within 1 hour and gets killed!!

A lot of research (and theoretical results), a few things happening "in the field".

When you go to a company that has clusters (like most of them), they typically have a job scheduler, so it's good to have some idea of what it is.

Next

1. SLURM and how to use it
2. A few promising directions for the future
   - Gang scheduling
   - Task based scheduler (work stealing)

SLURM

Simple Linux Utility for Resource Management

- Development started in 2002 @ Lawrence Livermore National Lab as a resource manager for Linux clusters
  - Sophisticated scheduling plugins added in 2008
- About 550,000 lines of C code today
- Supports Linux and limited support for other Unix variants
- Used on many of the world's largest computers
- Active global user community

SLURM design

- Highly scalable
  - Managing 3.1 million core Tianhe-2
  - Tested to much larger systems using emulation
- Open source GPLv2, available on Github: `https://github.com/SchedMD/`
- Fault-tolerant (no single point of failure)

- Dynamically linked objects loaded at run time based upon configuration file and/or user options
- Multiple plugins in 30 classes currently available
  - Network topology: 3D-torus, tree, etc
  - MPI: OpenMPI, PMI2, PMIx
  - Process tracking: cgroup, linuxproc, pgid, ipmi, etc.

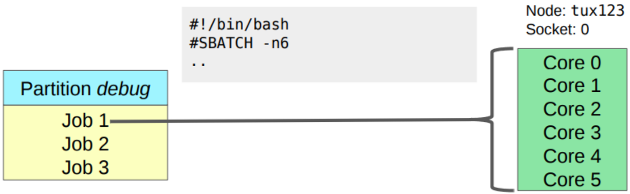
VANDERBILT UNIVERSITY

- Users submit jobs to partitions (queue)

Priority ordered queue of jobs

Partition *debug*
Job 1
Job 2
Job 3

- Jobs are allocated resources



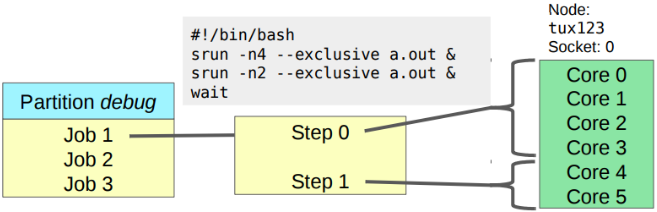
SLURM Usage

- Users submit jobs to partitions (queue)

  Priority ordered queue of jobs

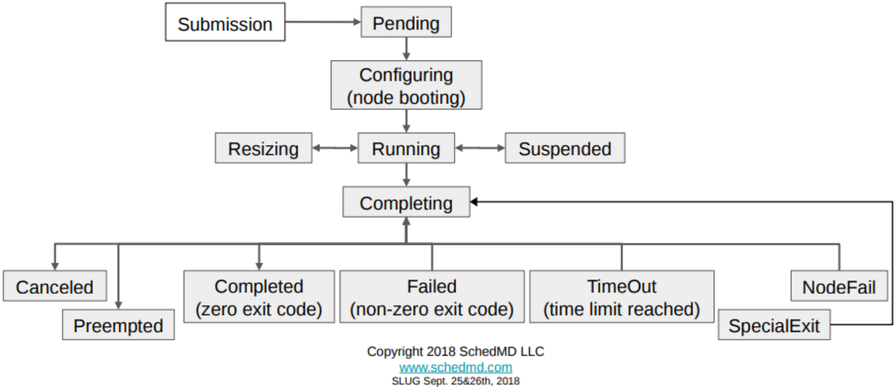
Partition *debug*
Job 1
Job 2
Job 3

- Jobs spawn steps, which are allocated resources from within the job's allocation



VANDERBILT UNIVERSITY

Job States



VANDERBILT UNIVERSITY

SLURM implementation

SLURM can be configured by system administrators in different formats

Frequently using a reservation-based configuration that assigns reservations on job arrival, job finish and at priority change.

- Can use one or multiple priority queues
- FCFS policy of ordering the queue
- Job priority is increased when waiting long time in the queue
- EASY Backfilling
- Node hours are required, job is killed in case of under-estimation
- Supports interactive sessions (`salloc`)
- Supports non-interactive sessions (`srun`, `sbatch`)

VANDERBILT UNIVERSITY

Summary

Documentation for SLURM available at `https://slurm.schedmd.com`

What you used for the homeworks should be enough for most projects

VANDERBILT UNIVERSITY

Summary

Documentation for SLURM available at `https://slurm.schedmd.com`

What you used for the homeworks should be enough for most projects

Next

- Gang scheduling
- Task based scheduler (work stealing)

Gang Scheduling: Basis

- All processes belonging to a job run at the same time (the term **gang** denotes all processors within a job).
- Each process runs alone on each processor.
- BUT: there is rapid **coordinated** context switching.
- It is possible to **suspend/preempt** jobs arbitrarily

May allow more flexibility to improve some metrics

Gang Scheduling: Basis

- All processes belonging to a job run at the same time (the term **gang** denotes all processors within a job).
- Each process runs alone on each processor.
- BUT: there is rapid **coordinated** context switching.
- It is possible to **suspend/preempt** jobs arbitrarily

May allow more flexibility to improve some metrics

- If processing times are not known in advance (or grossly erroneous), preemption can help short jobs that would be "stuck" behind a long job.
- Should improve machine utilization

VANDERBILT UNIVERSITY

Gang Scheduling: Example

- A 128 node cluster.
- A running 64-node job.
- A 32-node job and a 128-node job are queued.

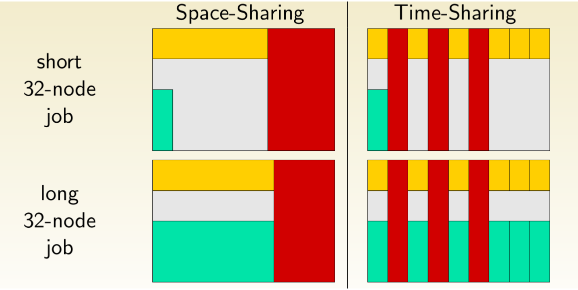
Should the 32-node job be started?

VANDERBILT UNIVERSITY

Gang Scheduling: Example

- A 128 node cluster.
- A running 64-node job.
- A 32-node job and a 128-node job are queued.

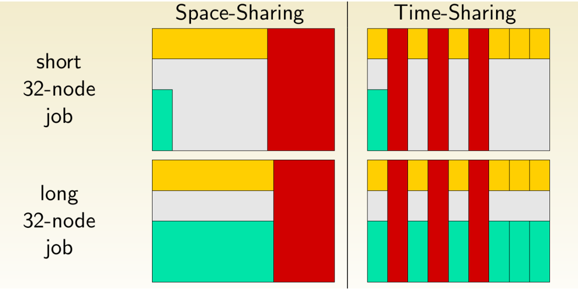
Should the 32-node job be started?



Gang Scheduling: Example

- A 128 node cluster.
- A running 64-node job.
- A 32-node job and a 128-node job are queued.

Should the 32-node job be started?



More uniform slowdown, better resource usage.

Drawbacks

- Overhead for context switching (trade-off between overhead and fine grain)
- Overhead for coordinating context switching across multiple processors

Drawbacks

- Overhead for context switching (trade-off between overhead and fine grain)
- Overhead for coordinating context switching across multiple processors
- Reduced cache efficiency(Frequent cache flushing)
- RAM Pressure (more jobs must fit in memory, swapping to disk causes unacceptable overhead)

VANDERBILT UNIVERSITY

Drawbacks

- Overhead for context switching (trade-off between overhead and fine grain)
- Overhead for coordinating context switching across multiple processors
- Reduced cache efficiency(Frequent cache flushing)
- RAM Pressure (more jobs must fit in memory, swapping to disk causes unacceptable overhead)

Typically not used in production HPC systems (batch scheduling is preferred)

Some implementations (MOSIX, Kerighed)

Work stealing

- A task-graph G needs to be executed on p processors.
- Non-clairvoyant setting : the structure of G and/or the execution times of its constitutive tasks are discovered online

VANDERBILT UNIVERSITY

Sharing vs Stealing

Batch scheduling

- Centralized scheduling
- A single list stores all ready tasks
- All processors retrieve work from that list
- Advantage(s)
  - Global view and knowledge
- Drawback(s)
  - Does not scale (contentions, etc.)

Work stealing

- Distributed scheduling
- Each processor owns a list of 'its' ready tasks
- Advantage(s)
  - No contention problem
  - Scalable solution
- Drawback(s)
  - Processors with empty lists do not know where to retrieve work from

VANDERBILT UNIVERSITY

Policies

Global round-robin

- A global variable holds the identity of the next processor to steal from
- Variable incremented after each steal (successful or not)
- Advantage : eventual progress
- Drawback : centralized solution...

Local round-robin

- Each processor has its own variable indicating the next processor it should try to steal from
- Variable incremented after each steal (successful or not)
- Advantage : eventual progress ; solution is scalable
- Drawback : all stealing processors may attempt to steal from the same processor; arbitrary notion of "distance" between processors

VANDERBILT UNIVERSITY

For now we're stuck with batch scheduling

Why don't we like Batch Scheduling?

For now we're stuck with batch scheduling

Why don't we like Batch Scheduling? Because queue waiting times are difficult to predict

- depends on the status of the queue
- depends on the scheduling algorithm used
- depends on all sorts of configuration parameters set by system administrator
- depends on future job completions!
- etc.

There is more and more demand for reservation support

VANDERBILT UNIVERSITY

Conclusion

Batch schedulers are complex pieces of software that are used in practice

- A lot of experience on how they work and how to use them
- But ultimately everybody knows they are an imperfect solution
- Many view the lack of theoretical foundations as a big problem

You need to know about them since every cluster uses them

VANDERBILT UNIVERSITY

Further Reading

Book on the theory of scheduling

D.B. Shmoys, J. Wein, and D.P. Williamson. *Scheduling parallel machines on-line* Symposium on Foundations of Computer Science, 0:131-140, 1991.

SLURM slides from

`https://slurm.schedmd.com/SLUG18/slurm_overview.pdf`

List of SLURM commands

`https://hpc-carpentry.github.io/hpc-intro/13-scheduler/index.html`

Figures from today's slides courtesy of Arnaud Legrand and Guillaume Pallez
`http://people.bordeaux.inria.fr/gaupy/ressources/teachings/2019/algo_hpc/`

VANDERBILT UNIVERSITY