Slurm

Scheduling on ARCHER2

Adrian Jackson, Iakovos Panourgias

EPCC, The University of Edinburgh

a.jackson@epcc.ed.ac.uk

@adrianjhpc

Reusing this material

ARCHER2 vs ARCHER

- ARCHER:
  - PBS:
    - `qsub`: submit a job
    - `qsub –I`: submit an interactive job
    - `qstat`: query the status of a job
    - `qdel`: delete a job
    - `qstat –q`: query the status of the system
  - Job launcher:
    - `aprun`
- ARCHER2:
  - Slurm:
    - `sbatch`: submit a job
    - `salloc`: submit an interactive job
    - `squeue`: query the status of a job
    - `scancel`: delete a job
    - `sinfo`: query the status of the system
  - Job launcher:
    - `srun`

ARCHER2 vs ARCHER

```
adrianj@uan01~> sinfo

PARTITION AVAIL  TIMELIMIT  NODES  STATE NODELIST

standard     up 1-00:00:00     45  down*
nid[001045,001047,001061,001068…]

standard     up 1-00:00:00     10  drain
nid[001016,001069,001468,…]

standard     up 1-00:00:00      5   resv nid[001000-
001003,001021]

standard     up 1-00:00:00    513  alloc nid[001004-
001015,001017,001022-001044,…]

standard     up 1-00:00:00    447   idle nid[001018-
001020,001046,001062-001067,…]

standard     up 1-00:00:00      1   down nid001138
```

```
adrianj@eslogin004:~> qstat -q

server: sdb

Queue            Memory CPU Time Walltime Node  Run   Que   Lm  State
---------------- ------ -------- -------- ---- ----- ----- ----  -----
parallel           --      --    24:00:00  --     0     0   --   D S
phase2             --      --    24:00:00  --     0     0   --   D S
ppn                --      --    24:00:00  --     0     0   --   E R
high               --      --    24:00:00  --     0     0   --   E R
weekend            --      --    24:00:00  --     0    48   --   E S
standard           --      --    24:00:00  --   281   125   --   E R
long               --      --    48:00:00  --    75     4   --   E R
short              --      --       --     --     0     0   --   E R
serial             --      --    24:00:00  --    24    12   --   E R
largemem           --      --    48:00:00  --     4     7   --   E S
low                --      --    03:00:00  --     0    31   --   E S
R7327082           --      --    00:20:00  --     3     0   --   E R
R7327794           --      --       --     --     1     0   --   E R
                                               ----- -----
                                                 388   227
```

ARCHER2 vs ARCHER

```
adrianj@uan01:~> sacctmgr show assoc where user=adrianj
format=account,user,maxtresmins

   Account       User     MaxTRESMins

---------- ---------- -------------

 cse-admin    adrianj          cpu=0

       y07    adrianj          cpu=0

       z19    adrianj
```

```
adrianj@eslogin004:~> budgets

=========================================

                 Budget       Remaining kAUs

-----------------------------------------

                    z01              89.953

                z01-cse               9.960

             z01-csetds    No resources left

               z01-test    No resources left

                z19-cse           23005.127

             z19-csetds            1999.965

=========================================
```

ARCHER2 vs ARCHER

```
adrianj@uan01~> sinfo

PARTITION AVAIL  TIMELIMIT  NODES  STATE NODELIST

standard     up 1-00:00:00     45  down*
nid[001045,001047,001061,001068…]

standard     up 1-00:00:00     10  drain
nid[001016,001069,001468,…]

standard     up 1-00:00:00      5   resv nid[001000-
001003,001021]

standard     up 1-00:00:00    513  alloc nid[001004-
001015,001017,001022-001044,…]

standard     up 1-00:00:00    447   idle nid[001018-
001020,001046,001062-001067,…]

standard     up 1-00:00:00      1   down nid001138
```

```
adrianj@eslogin004:~> qstat -q


server: sdb


Queue            Memory CPU Time Walltime Node   Run   Que   Lm  State
---------------- ------ -------- -------- ---- ----- ----- ----  -----
parallel           --      --    24:00:00  --      0     0   --   D S
phase2             --      --    24:00:00  --      0     0   --   D S
ppn                --      --    24:00:00  --      0     0   --   E R
high               --      --    24:00:00  --      0     0   --   E R
weekend            --      --    24:00:00  --      0    48   --   E S
standard           --      --    24:00:00  --    281   125   --   E R
long               --      --    48:00:00  --     75     4   --   E R
short              --      --       --     --      0     0   --   E R
serial             --      --    24:00:00  --     24    12   --   E R
largemem           --      --    48:00:00  --      4     7   --   E S
low                --      --    03:00:00  --      0    31   --   E S
R7327082           --      --    00:20:00  --      3     0   --   E R
R7327794           --      --       --     --      1     0   --   E R
                                               ----- -----
                                                 388   227
```

ARCHER2 vs ARCHER

```
adrianj@uan01:~> squeue

             JOBID PARTITION     NAME     USER ST       TIME  NODES
NODELIST(REASON)

             30752  standard     bash grte2001  R      14:31      1 nid001017

             30749  standard     hpcc  adrianj  R      15:40    512 nid[001004-
001015,…]
```

```
adrianj@eslogin004:~> qstat | more

Job id           Name             User             Time Use S Queue

---------------- ---------------- ---------------- -------- - -----

5148849.sdb      12S-2P           taibui                  0 H standard

5914122.sdb      pollutionprac    s1879801                0 H standard

6438161.sdb      PuWide_1         mzv2c                   0 H weekend

6439801.sdb      LES_narrow       mzv2c                   0 H weekend

6452569.sdb      DNS_Splitter     mzv2c                   0 H weekend

6463985.sdb      DNS_Splitter     mzv2c                   0 H weekend

6952967.sdb      tite2M-vac1.5    uccawra                 0 H weekend

7019105.sdb      postproc_atmos_  dflocco                 0 H serial

7019107.sdb      postproc_nemo_e  dflocco                 0 H serial

7019108.sdb      postproc_atmos_  dflocco                 0 H serial

7019127.sdb      postproc_cice_e  dflocco                 0 H serial

7019130.sdb      postproc_atmos_  dflocco                 0 H serial

7019131.sdb      postproc_cice_e  dflocco                 0 H serial

7019136.sdb      postproc_atmos_  dflocco                 0 H serial

7024877.sdb      vac-2            uccawra                 0 H weekend

7097417.sdb      ave4             gcastigl                0 H long

7287426.sdb      ag100            bsohail                 0 H standard

7325564.sdb      nogly_new        mohdfbs          00:00:02 R standard

7325565.sdb      nogly_new        mohdfbs                 0 H standard
```

ARCHER2 vs ARCHER

```
#!/bin/bash
#SBATCH --job-name=Example_MPI_Job
#SBATCH --time=0:20:0
#SBATCH --nodes=4
#SBATCH --tasks-per-node=128
#SBATCH --cpus-per-task=1
#SBATCH --account=[budget code]
#SBATCH --partition=standard
#SBATCH --qos=standard

srun --cpu-bind=cores ./my_mpi_executable.x
```

```
#!/bin/bash --login
#PBS -N hello_world
#PBS -l walltime=0:5:0
#PBS -l select=43
#PBS -A [budget code]

cd $PBS_O_WORKDIR
aprun -n 1024 $HOME/ my_mpi_executable.x
```

ARCHER2 vs ARCHER

- Interactive run on ARCHER:

```
qsub -IVl select=8,walltime=1:0:0 -A [project code]
qsub: waiting for job 492383.sdb to start
adrianj@mom3:~> aprun –n 192 ./my_exe
```

- Interactive run on ARCHER2:

```
salloc --nodes=8 --tasks-per-node=128 --cpus-per-task=1 --time=1:0:0 --partition=standard --
qos=standard --account=[budget code]

salloc: Granted job allocation 30751

salloc: Waiting for resource configuration

salloc: Nodes nid[001019-001020,001062-001067] are ready for job

adrianj@uan01:/work/z19/z19/adrianj/>srun –cpu-bind=cores ./my_exe
```

ARCHER2 configuration

- Partitions:
  - Standard
  - Highmem (for the full system, not currently available)
  - GPU (for the full system, not currently available)*

- QoS:
  - Standard: up to 24 hour jobs, up to the full system
    - 16 running jobs + 16 queued jobs per user
  - Long: from 24-48 hour jobs, up to 64 nodes *
    - 64 nodes in total per user
    - 512 nodes in long queue use for the 4 cabinet system in total across all users
  - Short: up to 20 minutes, up to 4 nodes*
    - 1 running job + 2 queued job per user
    - requires a reservation
      - `sbatch –reservation=shortqos …`
    - only active 08.00-20.00 Monday-Friday

*subject to change in the future

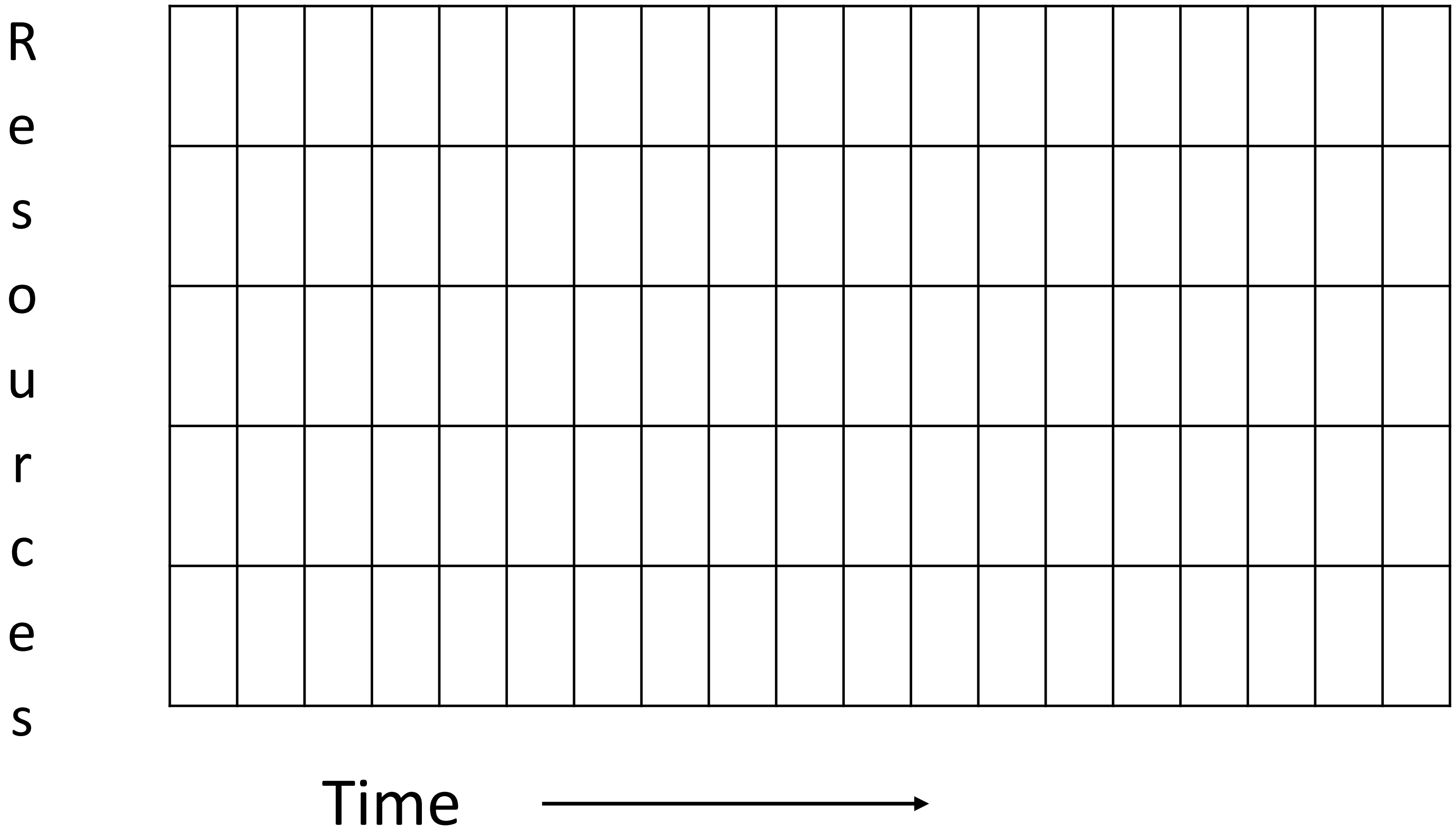
Placement and binding

- Slurm (srun) is doing our placement and binding
  - Processes and threads to cores

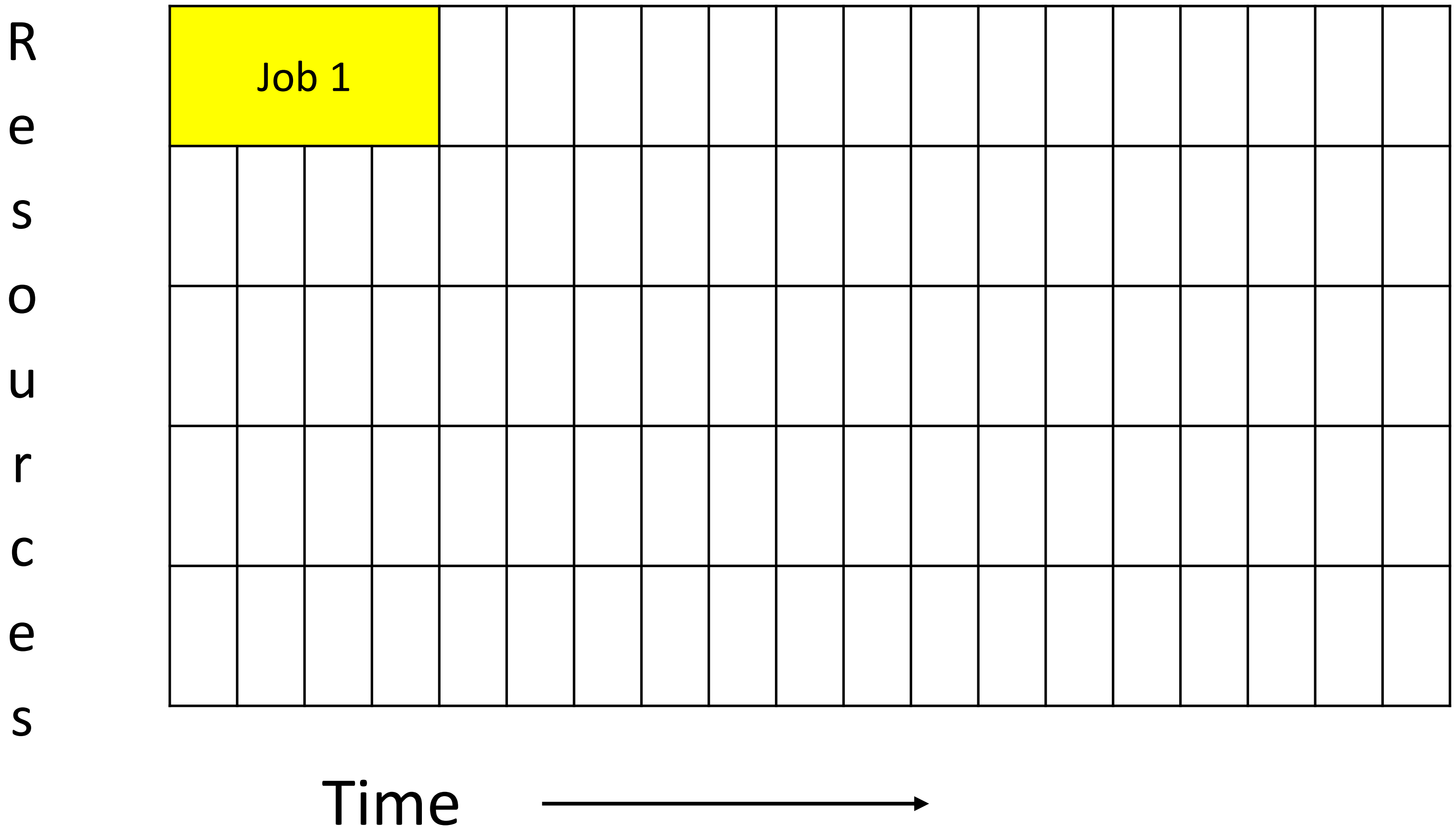
```
#!/bin/bash
#SBATCH --job-name=Example_MPI_Job
#SBATCH --time=0:20:0
#SBATCH --nodes=4
#SBATCH --ntasks=32
#SBATCH --tasks-per-node=8
#SBATCH --cpus-per-task=16
#SBATCH --account=[budget code]
#SBATCH --partition=standard
#SBATCH --qos=standard

# Set the number of threads to 16 and specify placement
#   There are 16 OpenMP threads per MPI process
#   We want one thread per physical core
export OMP_NUM_THREADS=16
export OMP_PLACES=cores
srun --hint=nomultithread --distribution=block:block
./my_mixed_executable.x arg1 arg2
```

Backfill Scheduling Example



Backfill Scheduling Example

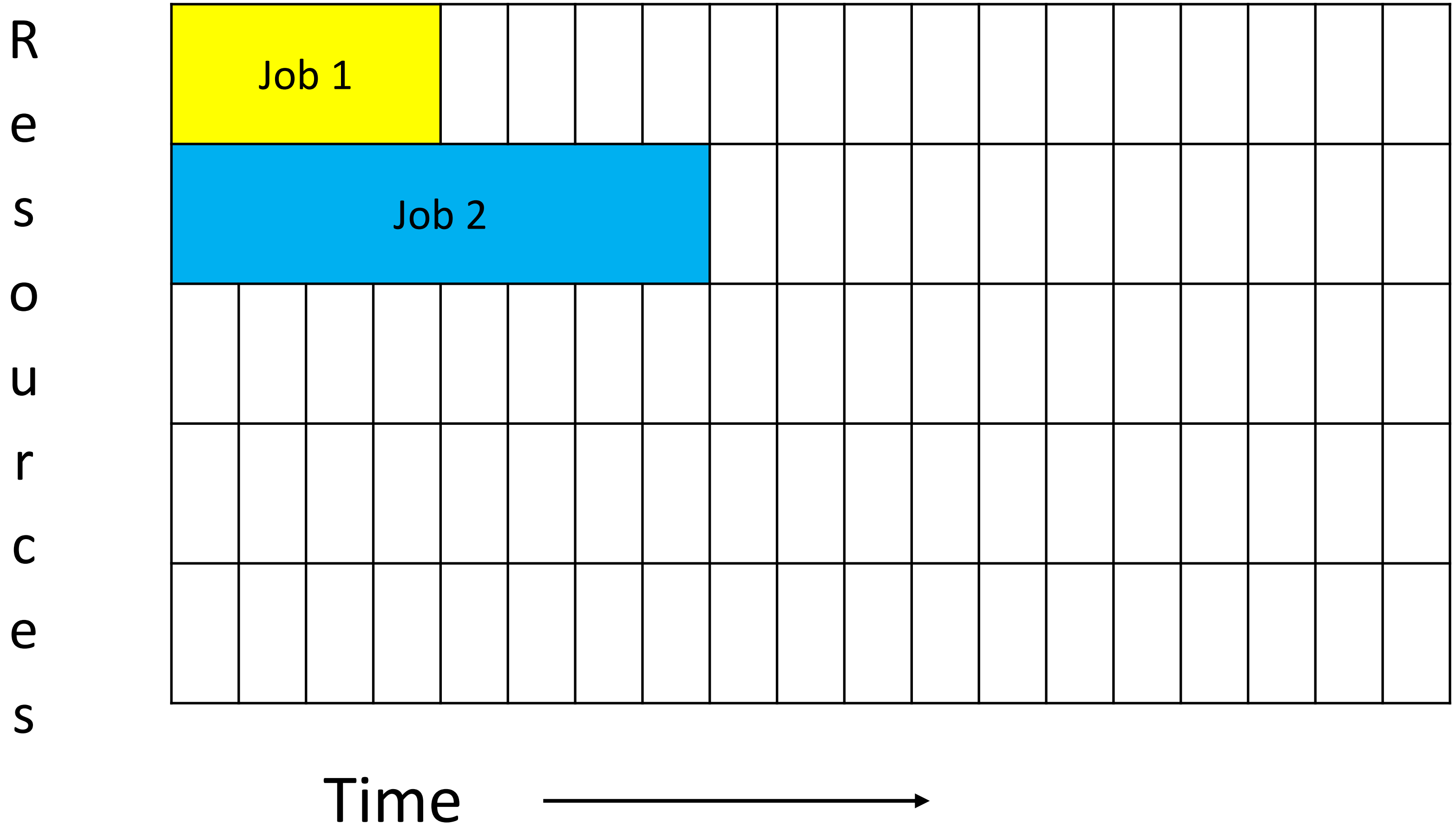


Job 1

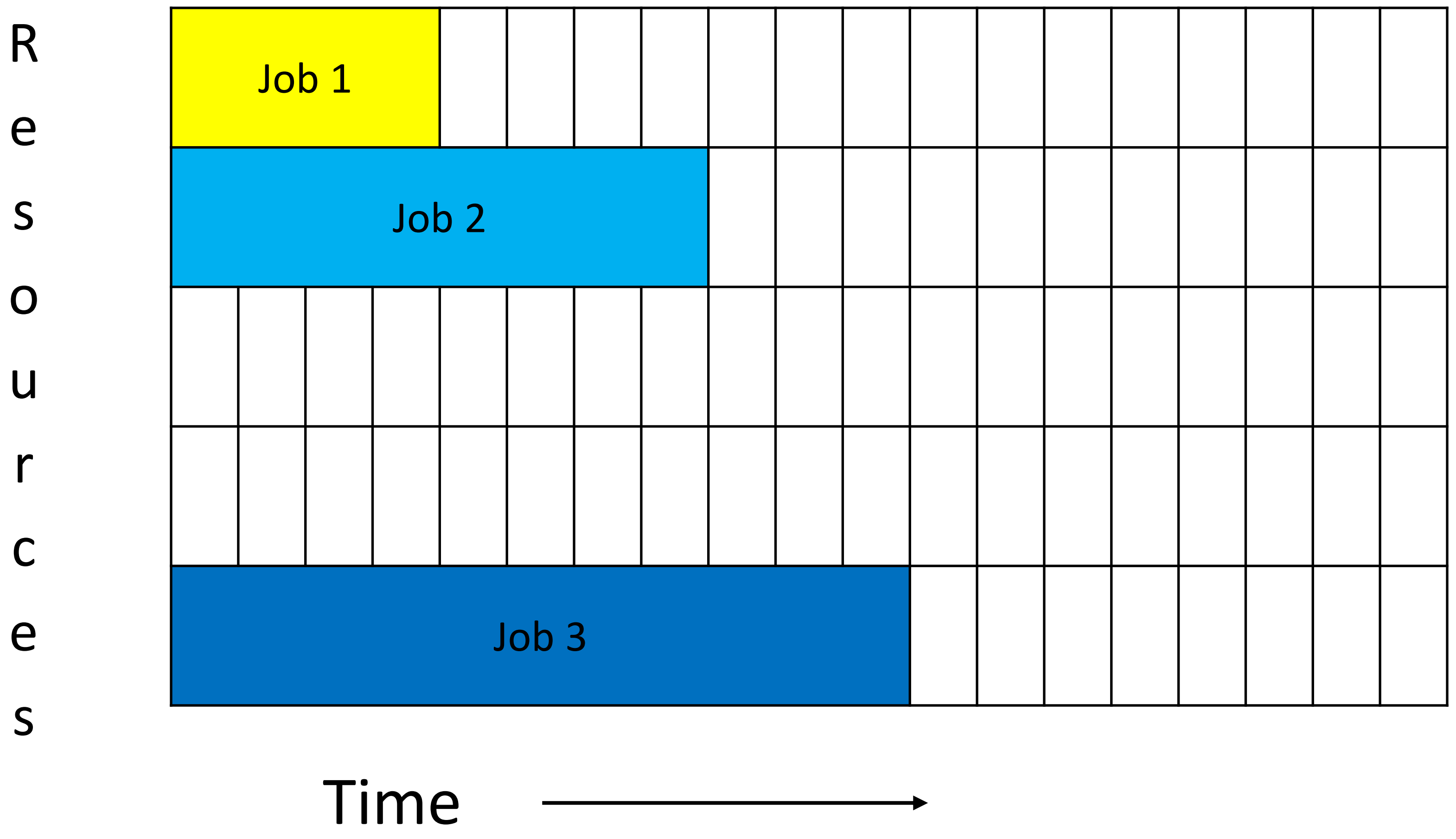
Resources

Time

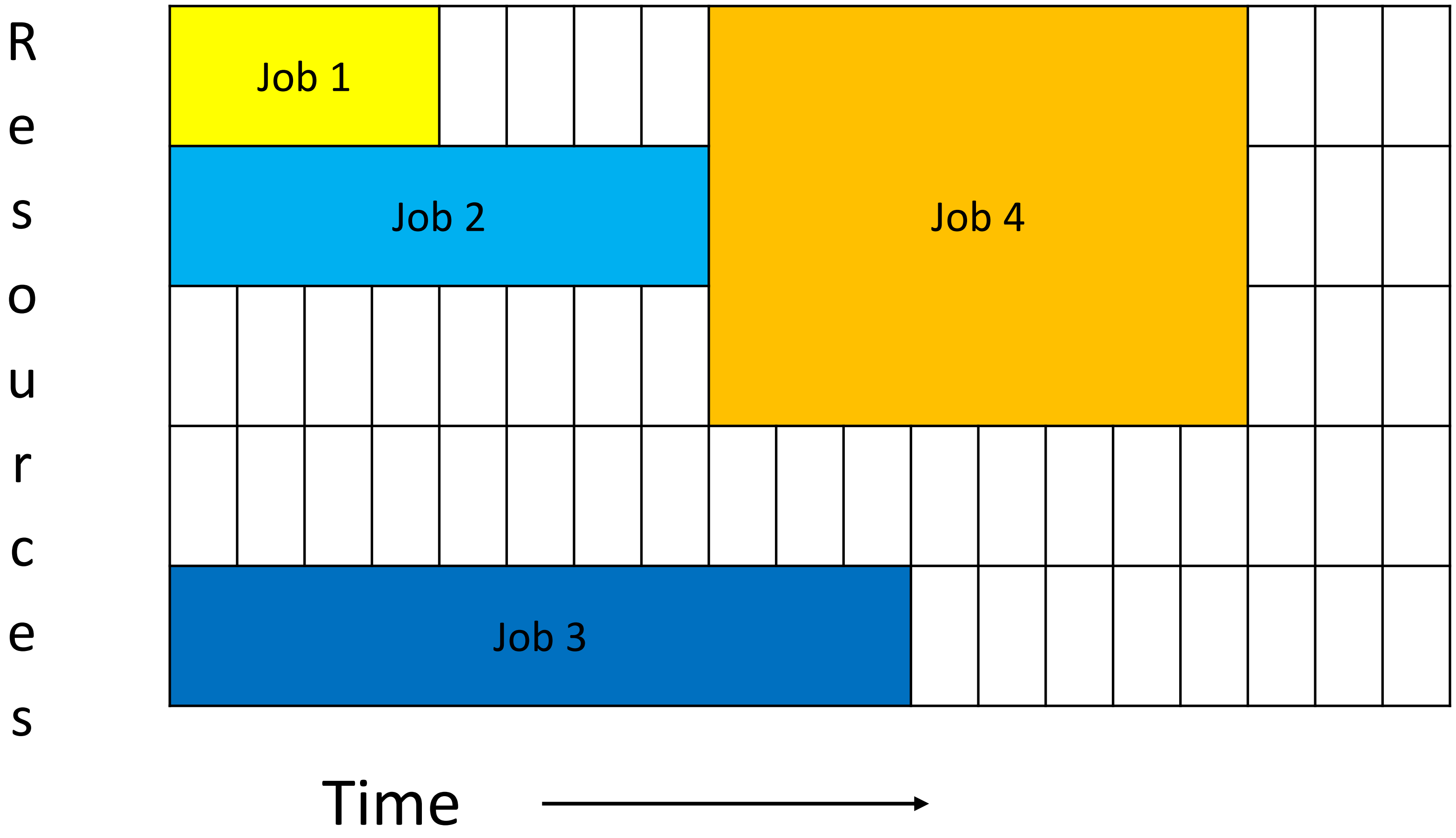
Backfill Scheduling Example



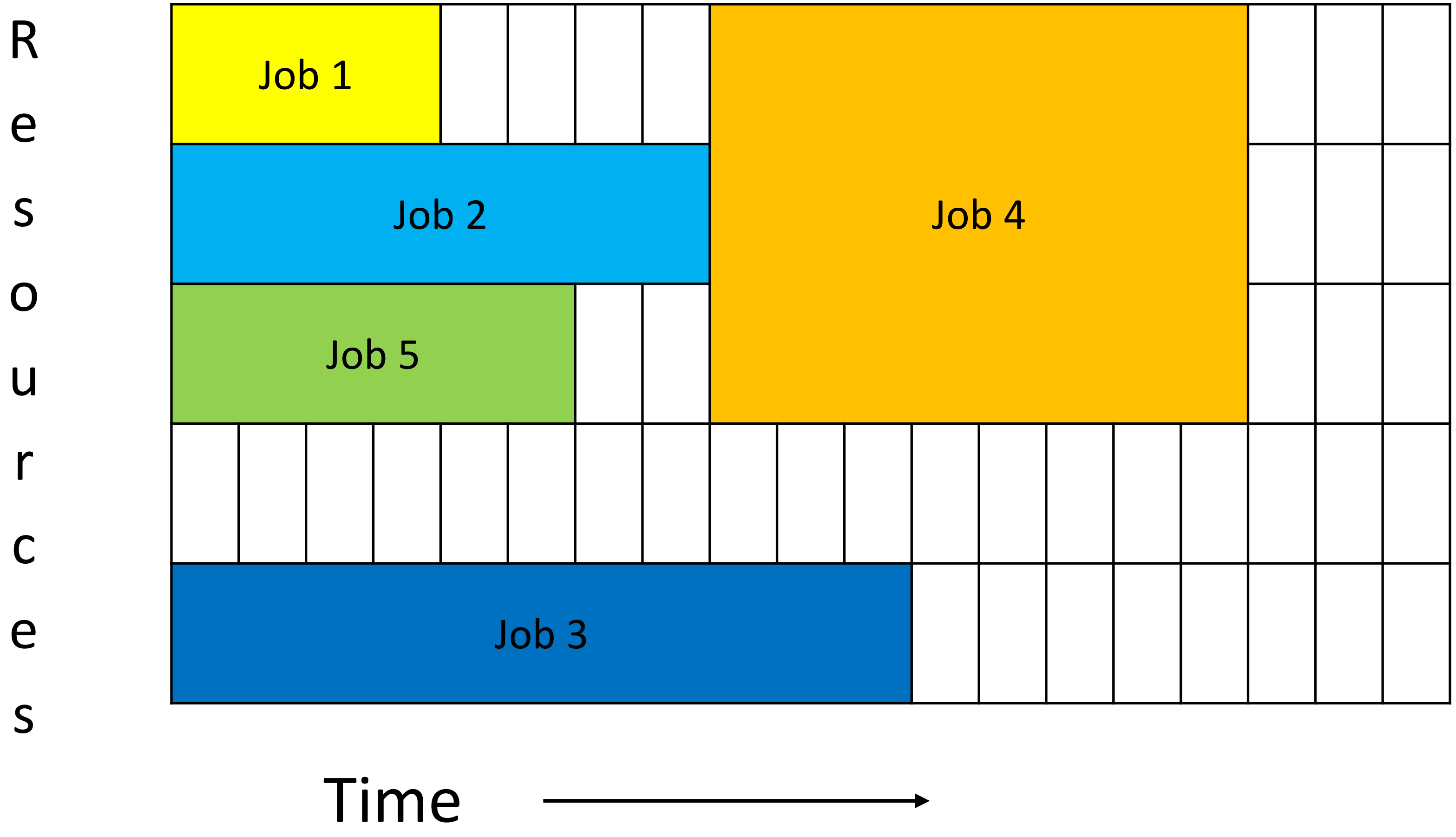
Backfill Scheduling Example



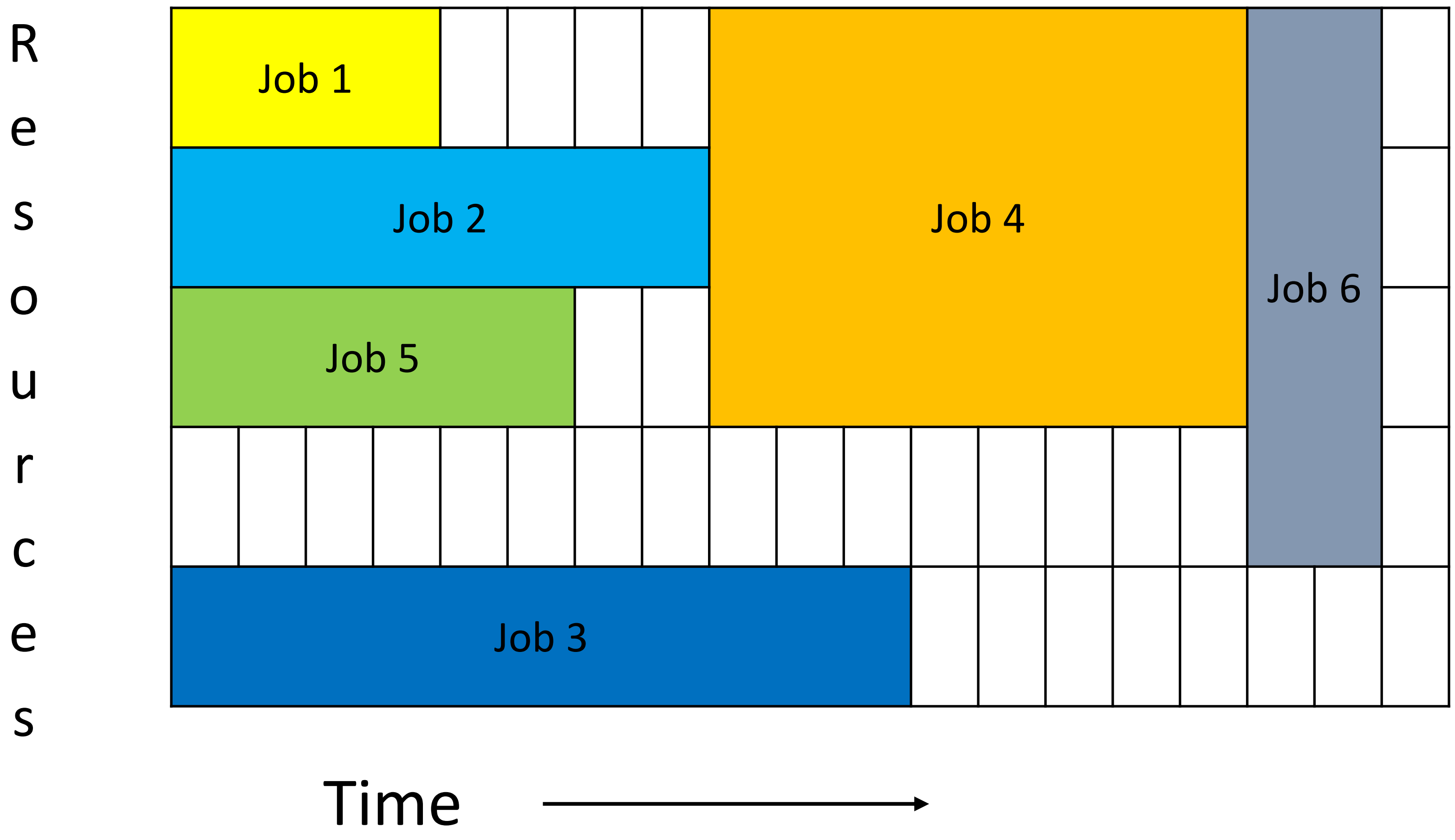
Backfill Scheduling Example



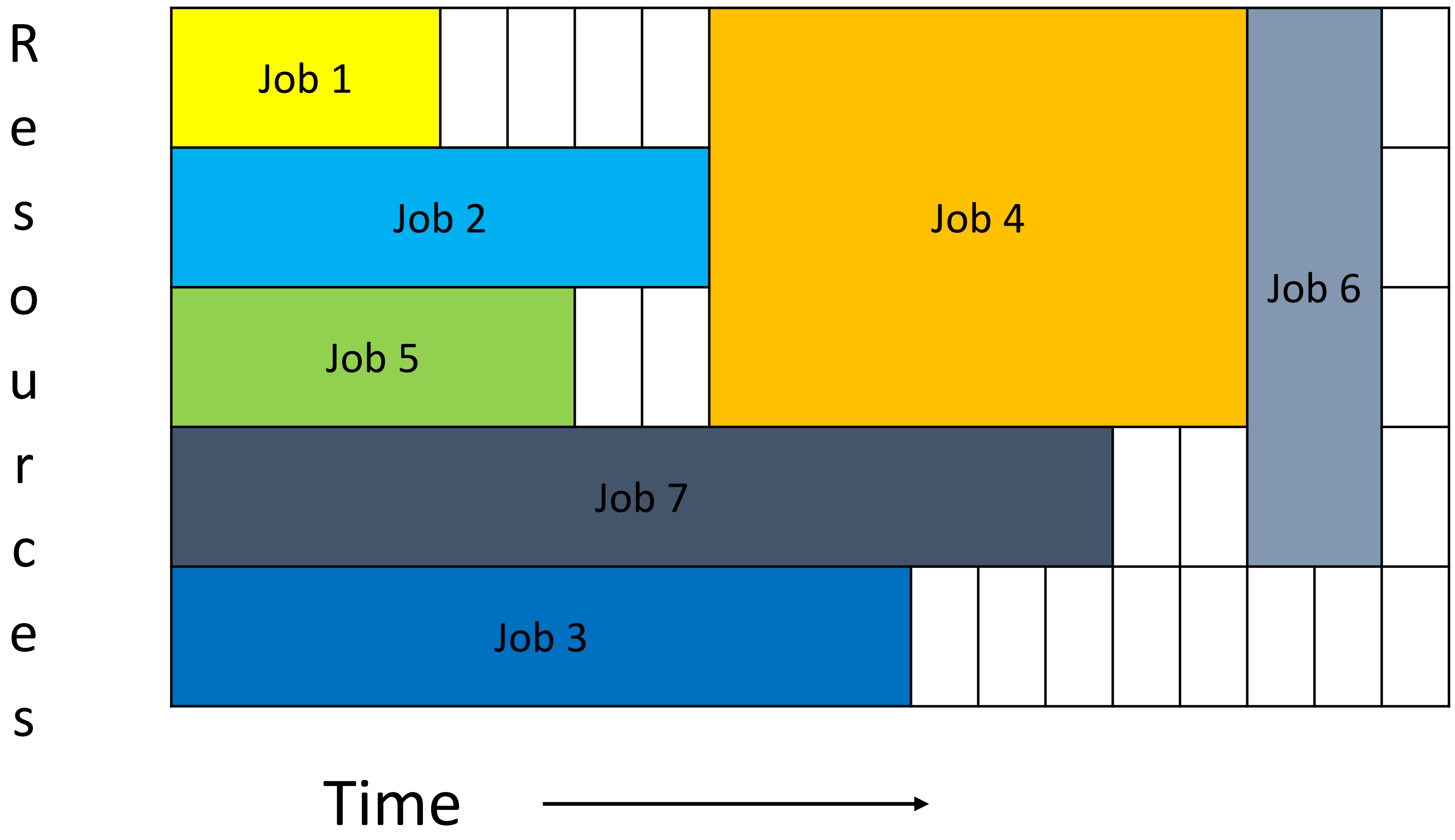
Backfill Scheduling Example



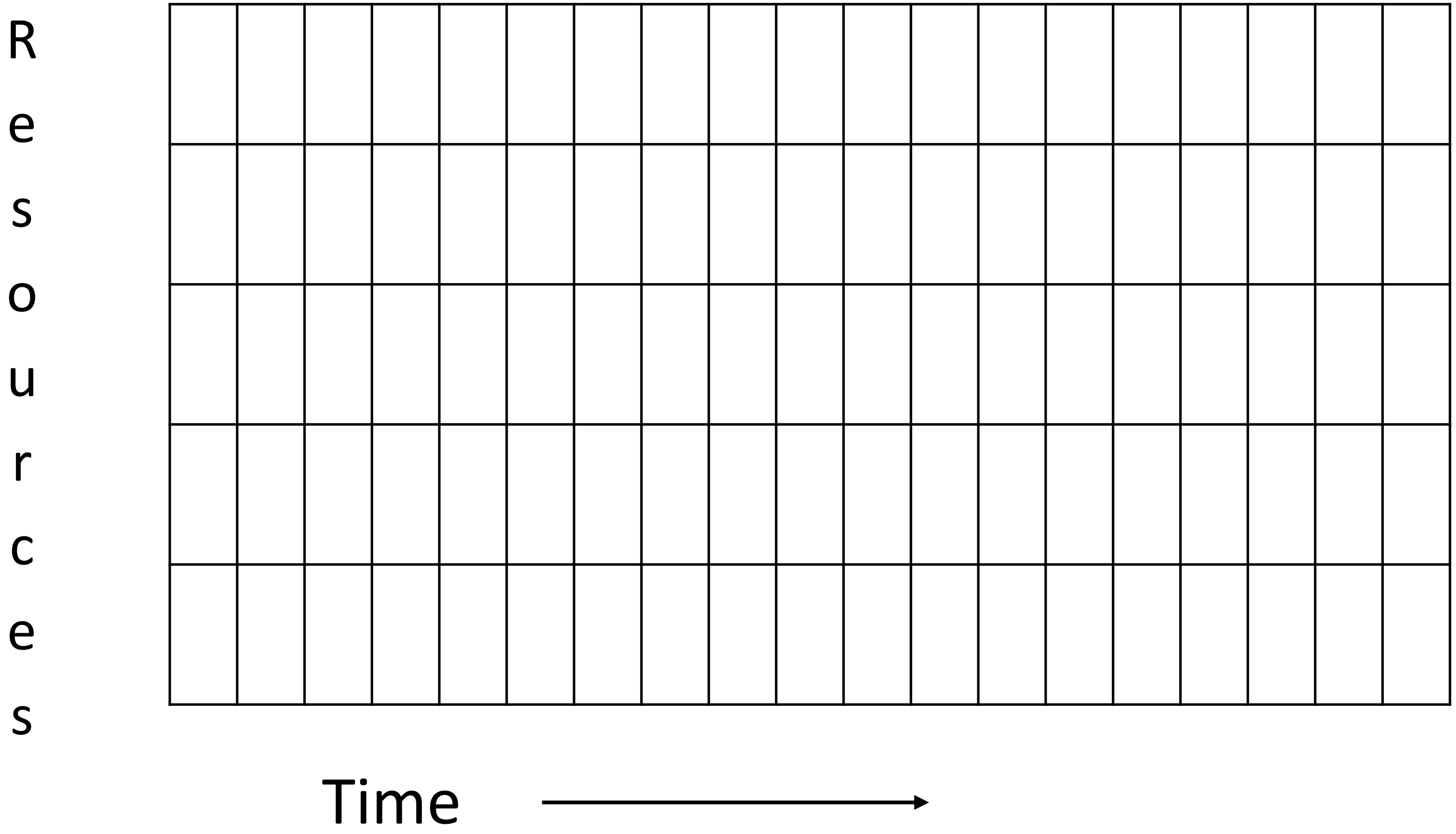
Backfill Scheduling Example



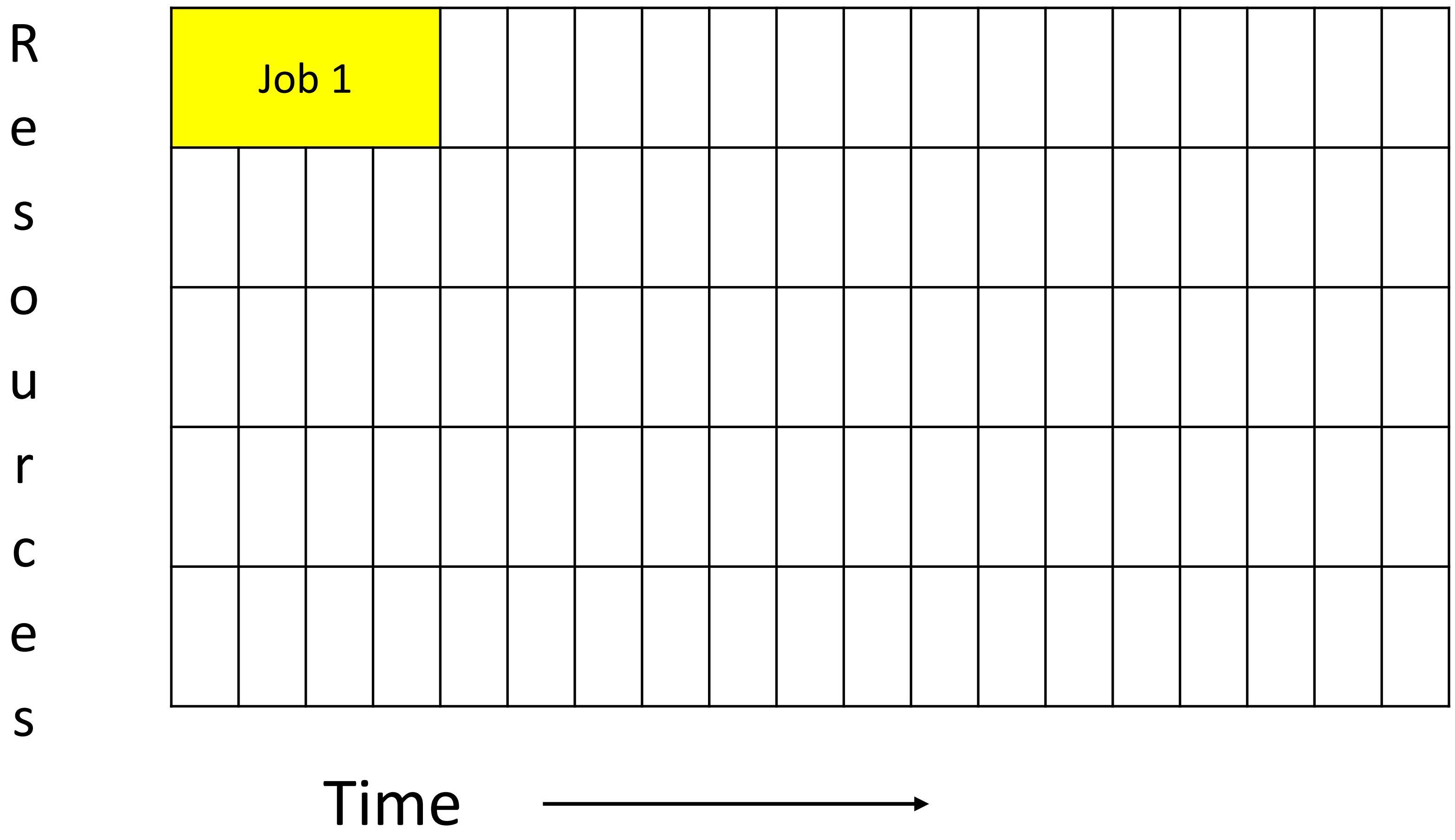
Backfill Scheduling Example



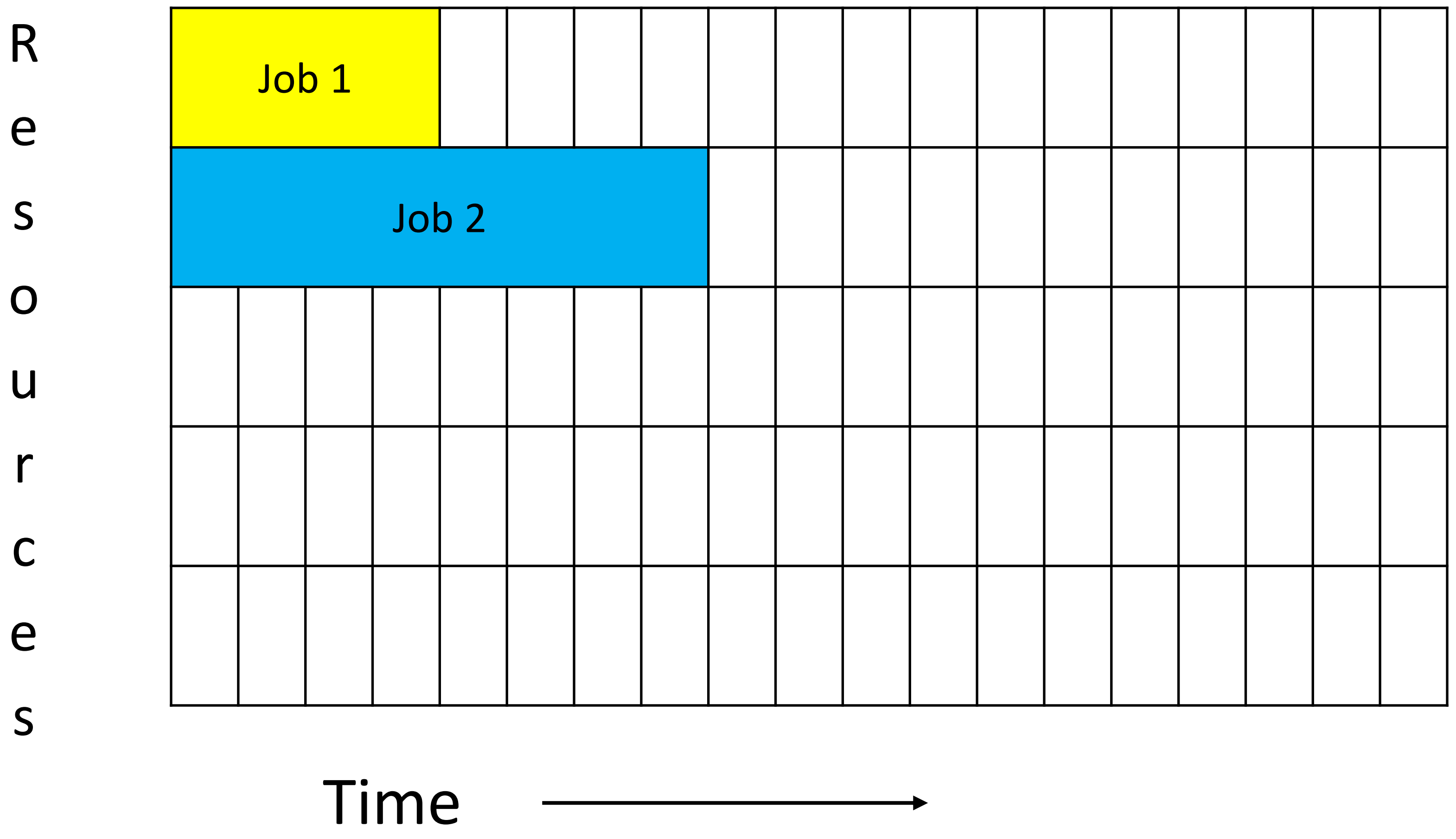
FIFO Scheduling Example



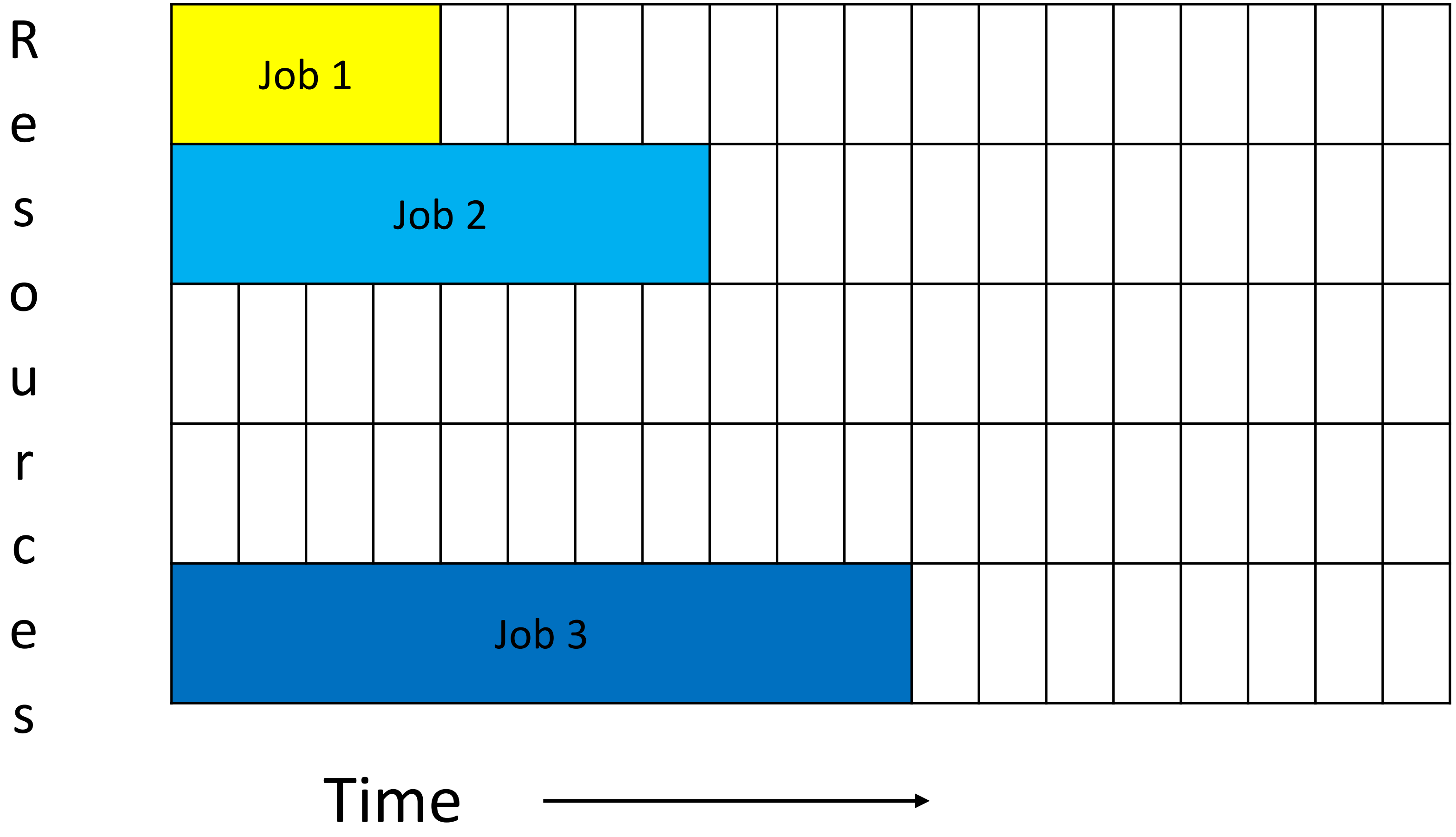
FIFO Scheduling Example



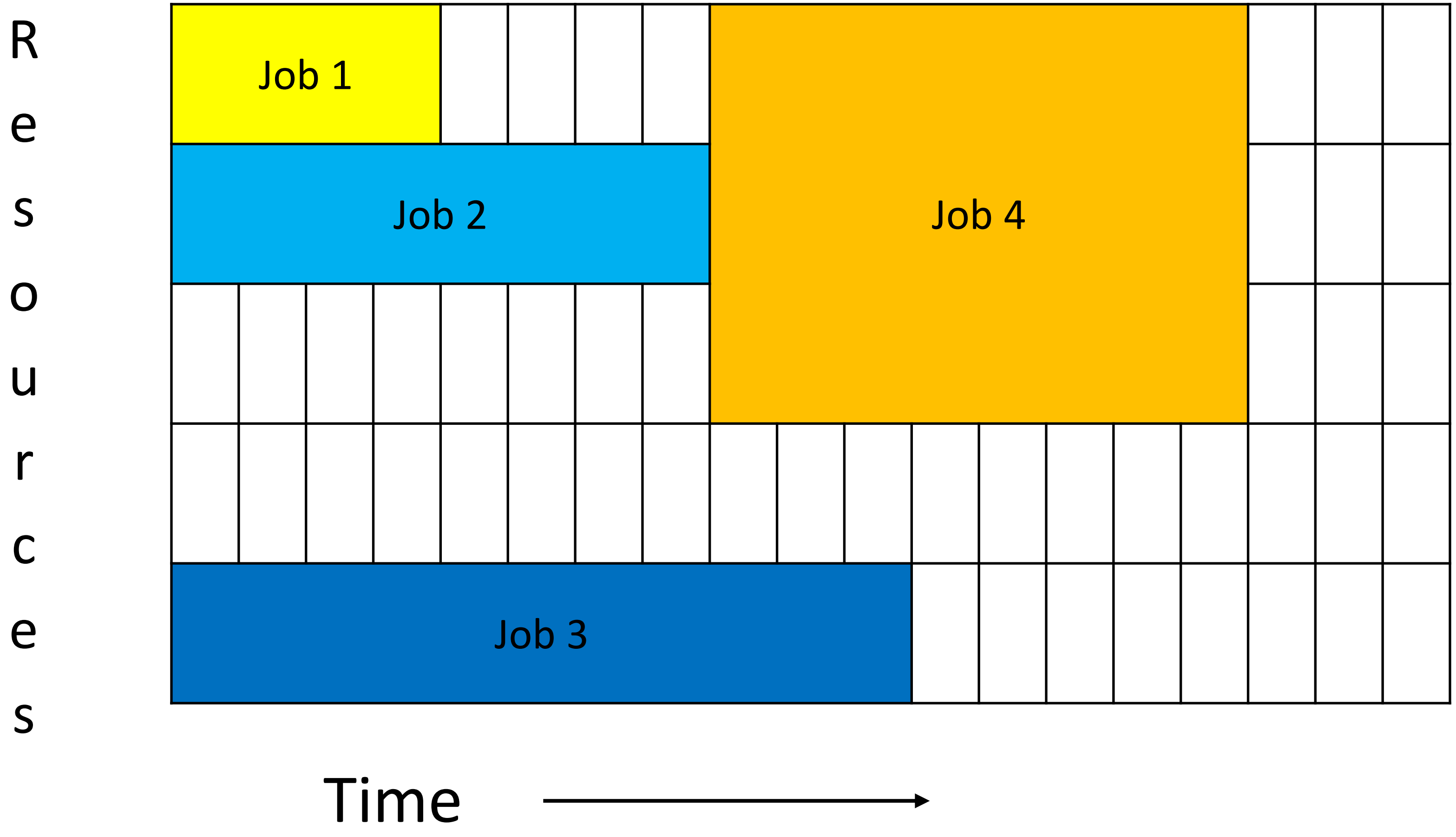
FIFO Scheduling Example



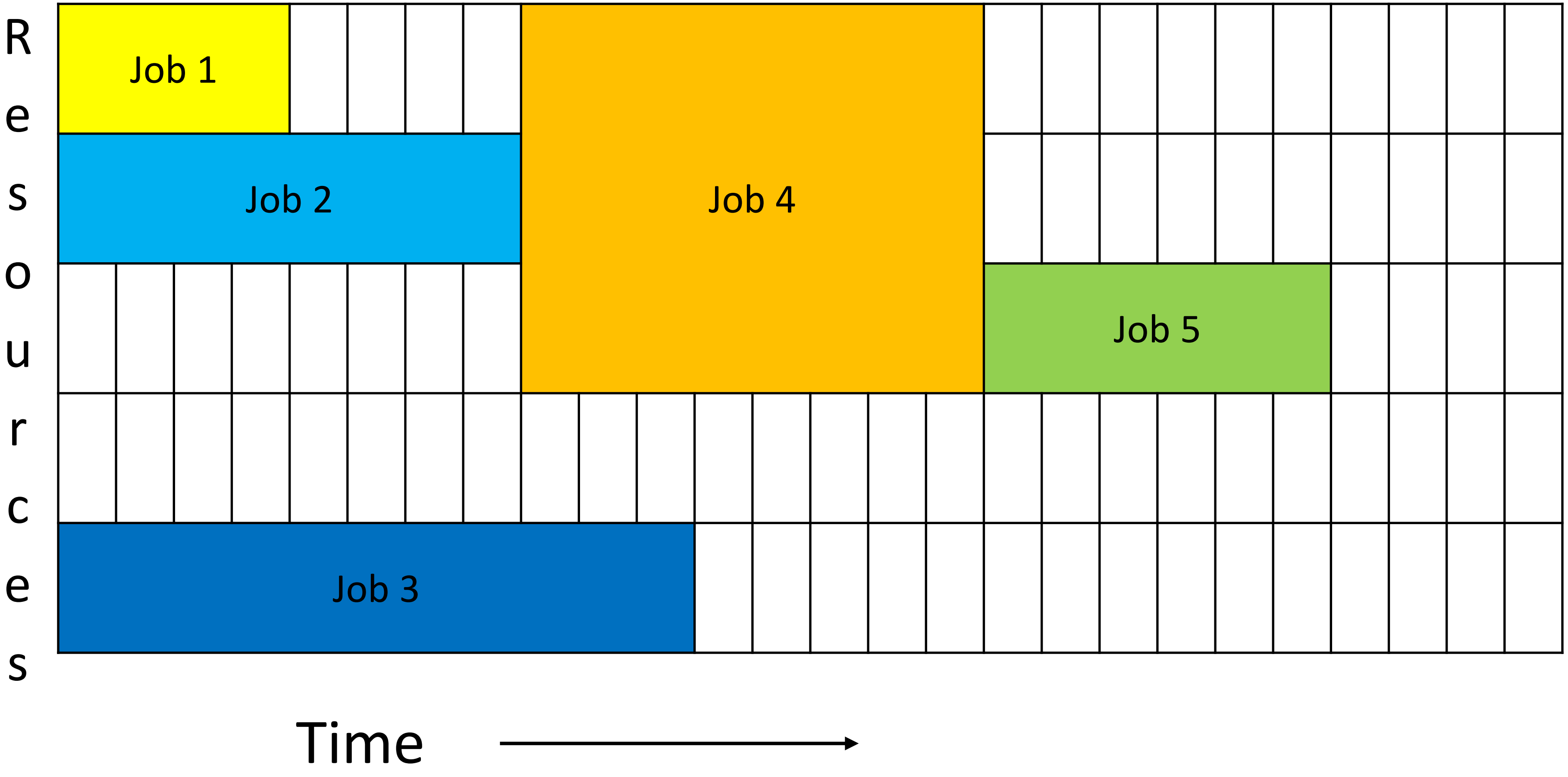
FIFO Scheduling Example



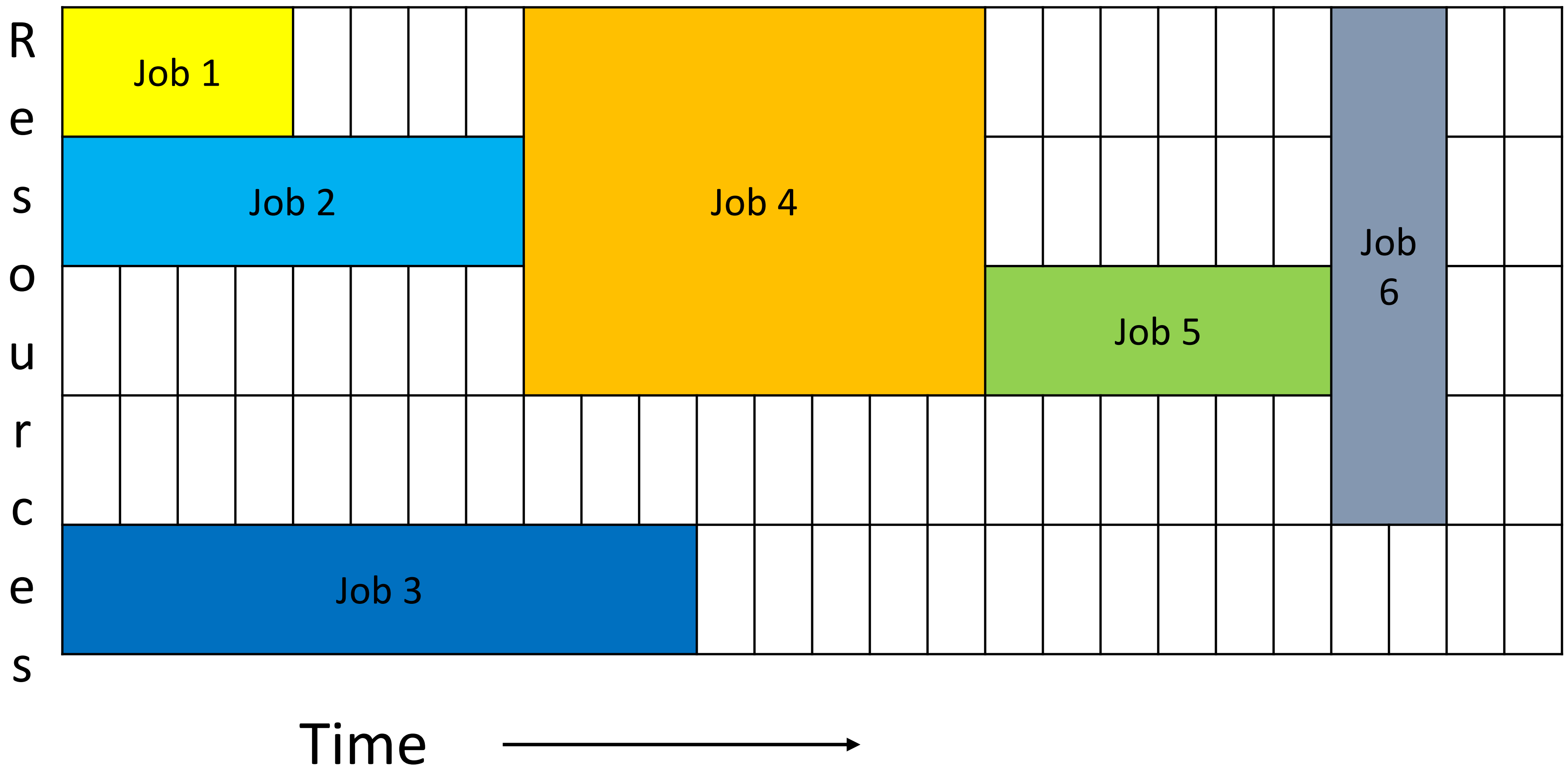
FIFO Scheduling Example



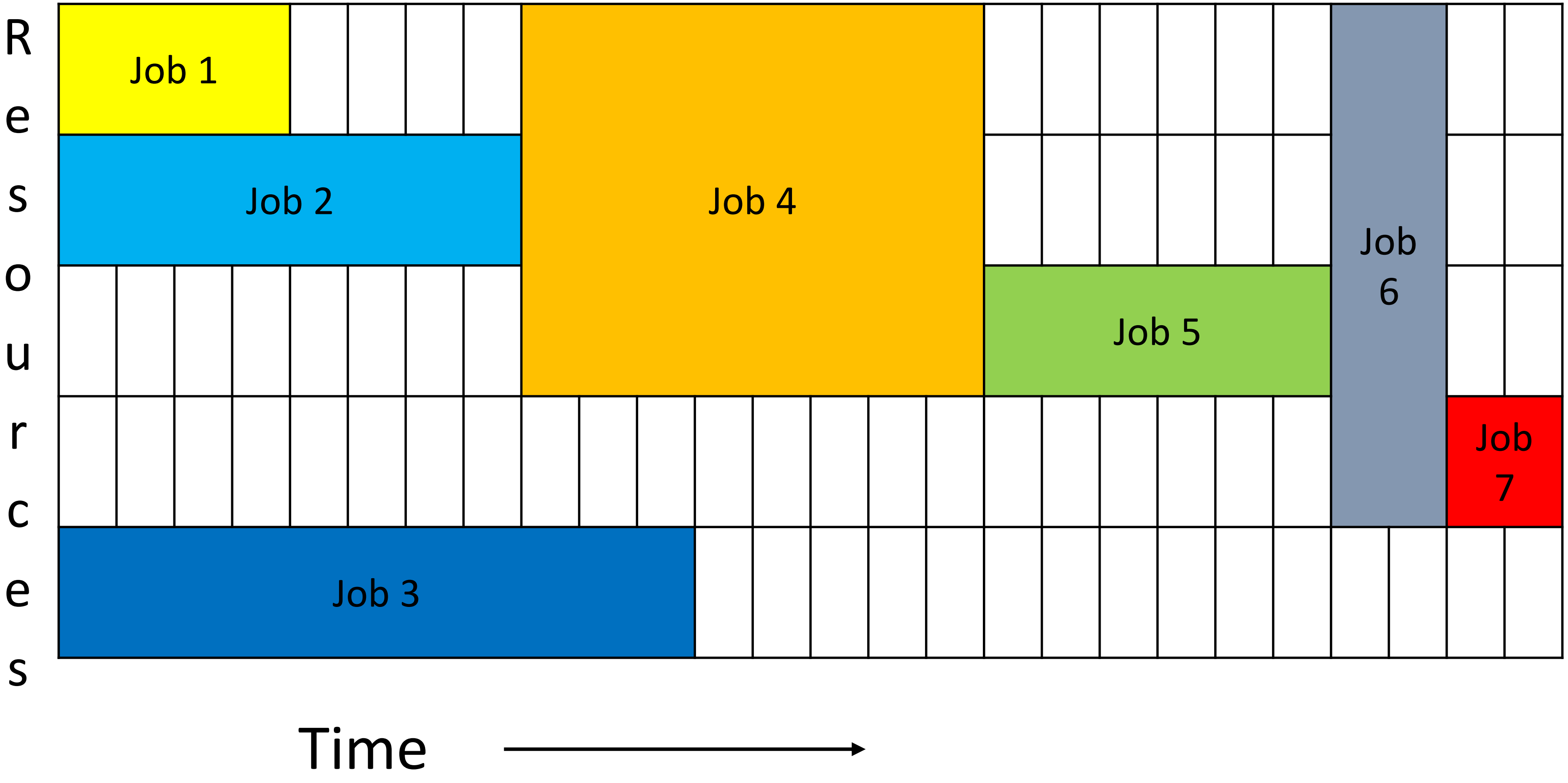
FIFO Scheduling Example



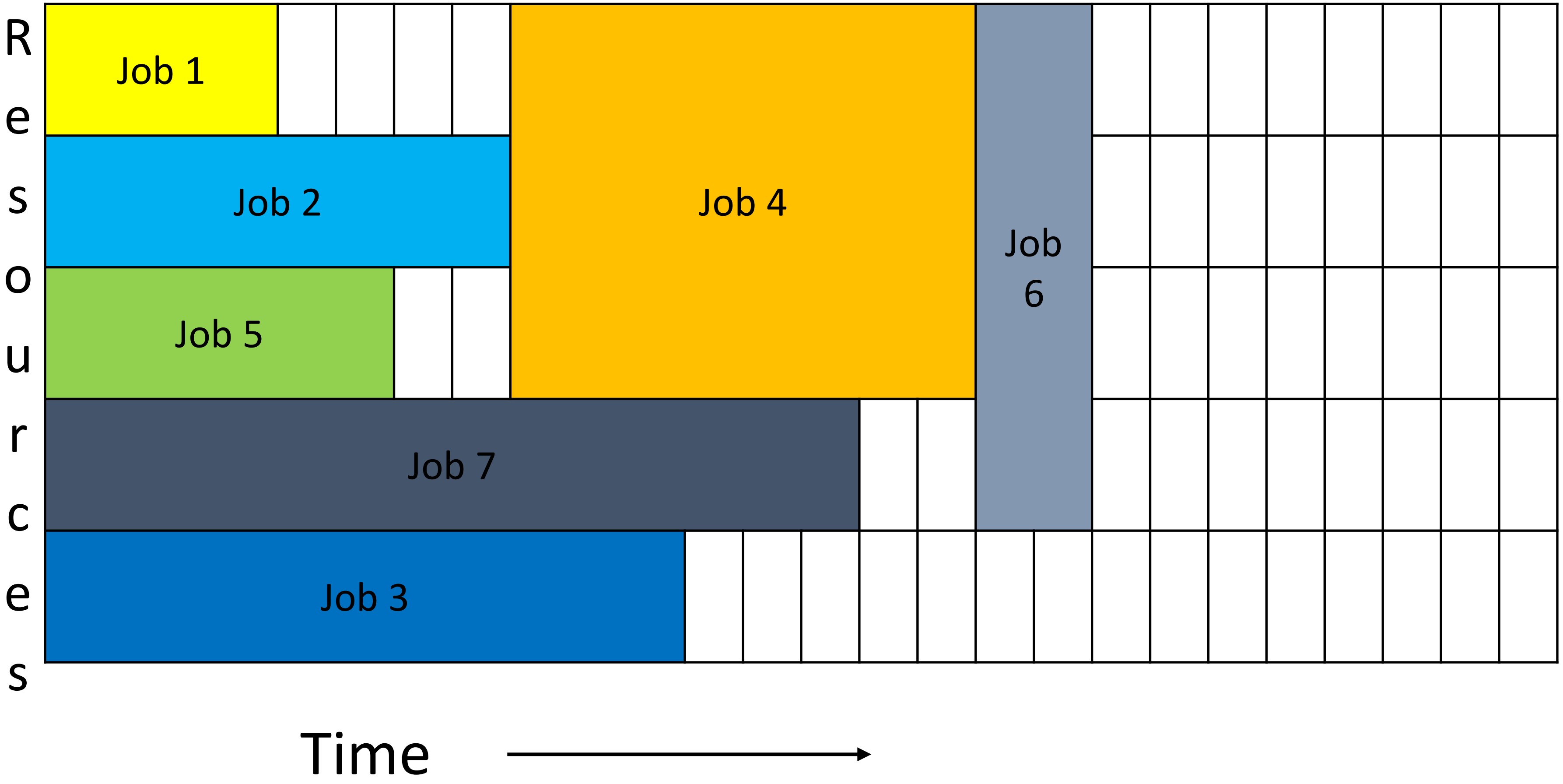
FIFO Scheduling Example



FIFO Scheduling Example



Backfill Scheduling Example



Slurm Scheduling

- Slurm tries to find a better schedule (using quick and simple algorithms) when:
  - A job is submitted;
  - A job completes;
  - A configuration change takes place.
- Slurm also performs slower and more expensive scheduling attempts less frequently
- This design allows nearly instant response; even when thousand of job are submitted at the same time.

Slurm – Quick Scheduling

- Slurm only checks the first X (by default 100) entries of the queue for new scheduling opportunities;

- Once a job in a partition is left pending (i.e. no scheduling is possible), Slurm ignores the other jobs in that partition;

Slurm – Thorough Scheduling

- Slurm checks all jobs in the queue (or until a configurable time limit is reached);

- Jobs are ordered by priority so this operation has low overhead;

- However, jobs in lower priority partitions (queues) have now more opportunities to start.