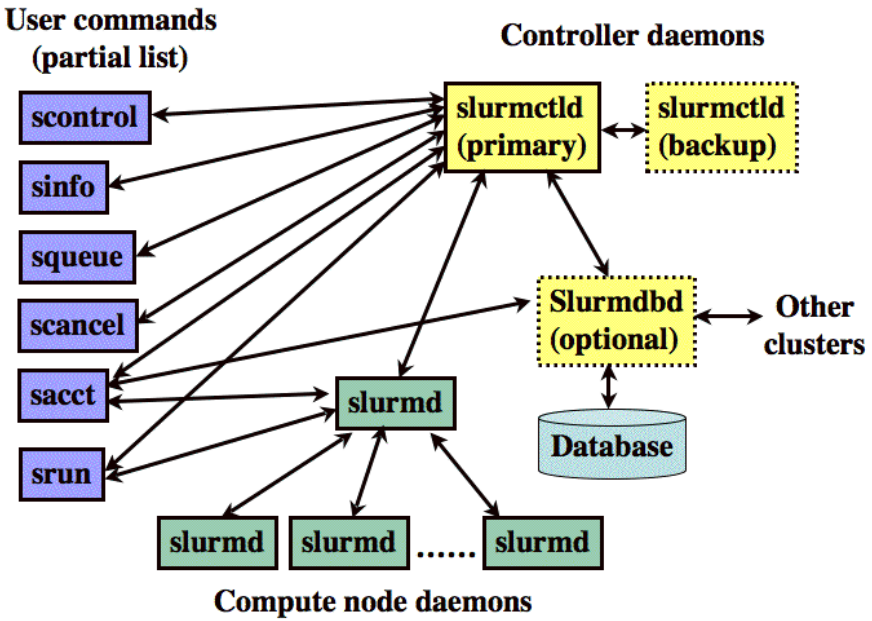
Erőforráskezelő (resource manager)

- "köztesréteg" a klaszter és a párhuzamos programok között
- munkaállomásszerű futtatási környezet biztosítása: `a.out` vs `srun -n8 a.out`
- párhuzamos környezetek támogatása: MPI, OpenMP, array, szerver-kliens

Feladatütemező (job scheduler)

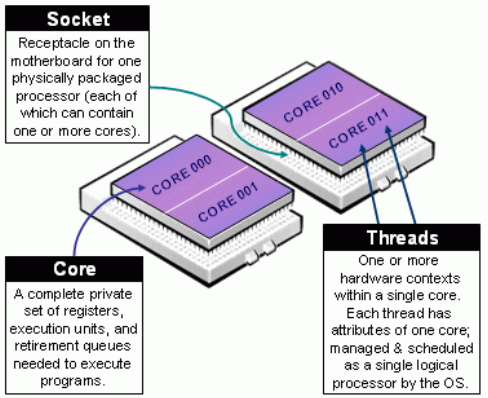
- több a feladat (job), mint az erőforrás
- a feladat sorok kezelése:
  - komplex ütemezései algoritmusok használata:
  - hálózati topológia
  - fair-share
  - advanced reservation (előfoglalás)
  - gang ütemezés (időosztás)
  - erőforráslimitek használata (queue, group, user)

SLURM Architektúra



SLURM entitások

- job: erőforrás kérelem
- job steps: feladatok halmaza
- partitions: ütemezési sorok, QOS
- nodes, CPU, memória, GPU



Állapotinformációk: node

```
BUDAPEST[login] ~ (1)$ scontrol show node cn01
NodeName=cn01 Arch=x86_64 CoresPerSocket=12
   CPUAlloc=24 CPUErr=0 CPUTot=24 CPULoad=25.05 Features=ib,amd
   Gres=(null)
   NodeAddr=cn01 NodeHostName=cn01
   OS=Linux RealMemory=64000 AllocMem=24000 Sockets=2 Boards=1
   State=ALLOCATED ThreadsPerCore=1 TmpDisk=0 Weight=1
   BootTime=2013-03-04T13:08:00 SlurmdStartTime=2013-11-10T01:57:58
   CurrentWatts=0 LowestJoules=0 ConsumedJoules=0
   ExtSensorsJoules=n/s ExtSensorsWatts=0 ExtSensorsTemp=n/s
```

Állapotinformációk: partitions

```
BUDAPEST[login] ~ (0)$ sinfo -la
Mon Feb  3 23:33:35 2014
PARTITION AVAIL  TIMELIMIT   JOB_SIZE ROOT SHARE     GROUPS  NODES       STATE NODELIST
prod*        up 30-00:00:0 1-infinite   no    NO        all     18       mixed cn[01-18]
prod*        up 30-00:00:0 1-infinite   no    NO        all     13        idle cn[19-31]
test         up      30:00 1-infinite   no    NO        all      1        idle cn32

BUDAPEST[login] ~ (1)$ sinfo -lNe
Mon Feb  3 23:54:19 2014
NODELIST   NODES PARTITION       STATE CPUS    S:C:T MEMORY TMP_DISK WEIGHT FEATURES REASON
cn[01-18]     18     prod*       mixed   24   2:12:1  64000        0      1   ib,amd none
cn[19-31]     13     prod*        idle   24   2:12:1  64000        0      1   ib,amd none
cn32           1      test        idle   24   2:12:1  64000        0      1      amd none
```

Állapotinformációk: partitions

```
BUDAPEST[login] ~ (0)$ sinfo -o '%9P %4c %8z %8X %8Y %8Z'
PARTITION CPUS S:C:T    SOCKETS  CORES    THREADS
prod*     24   2:12:1   2        12       1
test      24   2:12:1   2        12       1

BUDAPEST[login] ~ (0)$ sjstat -c

Scheduling pool data:
-------------------------------------------------------------
Pool        Memory  Cpus  Total Usable   Free  Other Traits
-------------------------------------------------------------
prod*      64000Mb    24     31     31     13  ib,amd
test       64000Mb    24      1      1      1  amd
```

Állapotinformációk: partitions

```
BUDAPEST[login] ~ (0)$ squeue -la
Mon Feb  3 23:39:59 2014
       JOBID PARTITION     NAME     USER    STATE       TIME TIMELIMIT  NODES NODELIST(REASON
15645_[1-96]      prod       ar     geza  PENDING       0:00 7-00:00:00      1 (AssociationJobLimit)
    15165_52      prod       ar     geza  RUNNING    5:59:58 7-00:00:00      1 cn17
    15069_56      prod       ar     geza  RUNNING    1:24:28 7-00:00:00      1 cn17
       15790      prod mozyme10     stef  RUNNING    5:24:58 14-00:00:00      1 cn18
```

Állapotinformációk: partitions

```
BUDAPEST[login] ~ (0)$ scontrol show partition
PartitionName=prod
   AllocNodes=ALL AllowGroups=ALL Default=YES
   DefaultTime=NONE DisableRootJobs=NO GraceTime=0 Hidden=NO
   MaxNodes=UNLIMITED MaxTime=30-00:00:00 MinNodes=1 MaxCPUsPerNode=UNLIMITED
   Nodes=cn[01-31]
   Priority=1 RootOnly=NO ReqResv=NO Shared=NO PreemptMode=REQUEUE
   State=UP TotalCPUs=744 TotalNodes=31 SelectTypeParameters=N/A
   DefMemPerNode=UNLIMITED MaxMemPerNode=UNLIMITED

PartitionName=test
   AllocNodes=ALL AllowGroups=ALL Default=NO
   DefaultTime=NONE DisableRootJobs=NO GraceTime=0 Hidden=NO
   MaxNodes=UNLIMITED MaxTime=00:30:00 MinNodes=1 MaxCPUsPerNode=UNLIMITED
   Nodes=cn32
   Priority=1 RootOnly=NO ReqResv=NO Shared=NO PreemptMode=REQUEUE
   State=UP TotalCPUs=24 TotalNodes=1 SelectTypeParameters=N/A
   DefMemPerNode=UNLIMITED MaxMemPerNode=UNLIMITED
```

Állapotinformációk: qos és prioritás

```
sacctmgr show qos
```

```
sprio -l
```

Állapotinformációk: job

- információ a sorban lévő jobokról

```
JobId=15164 ArrayJobId=15069 ArrayTaskId=96 Name=ar
   UserId=geza(11006) GroupId=geza(11006)
   Priority=10443 Account=fazisata QOS=normal
   JobState=PENDING Reason=AssociationJobLimit Dependency=(null)
   Requeue=1 Restarts=0 BatchFlag=1 ExitCode=0:0
   RunTime=00:00:00 TimeLimit=7-00:00:00 TimeMin=N/A
   SubmitTime=2014-02-03T13:25:29 EligibleTime=2014-02-03T13:25:30
   StartTime=Unknown EndTime=Unknown
   PreemptTime=None SuspendTime=None SecsPreSuspend=0
   Partition=prod AllocNode:Sid=login:7455
   ReqNodeList=(null) ExcNodeList=(null)
   NodeList=(null)
   NumNodes=1 NumCPUs=1 CPUs/Task=1 ReqS:C:T=*:*:*
   MinCPUsNode=1 MinMemoryCPU=1000M MinTmpDiskNode=0
   Features=(null) Gres=(null) Reservation=(null)
   Shared=OK Contiguous=0 Licenses=(null) Network=(null)
   Command=/fs01/home/geza/ar/ar.sh
   WorkDir=/fs01/home/geza/ar
```

Állapotinformációk: futó job

```
BUDAPEST[login] ~ (0)$ smemory 14781

 MaxVMSize  MaxVMSizeNode  AveVMSize     MaxRSS MaxRSSNode     AveRSS
---------- -------------- ---------- ---------- ---------- ----------

   117736K           cn01    117736K      8896K       cn01      8896K
    95800K           cn01     95800K     83288K       cn01     83288K

BUDAPEST[login] ~ (0)$ sdisk 14781

 MaxDiskRead MaxDiskReadNode    AveDiskRead MaxDiskWrite MaxDiskWriteNode   AveDiskWrite
------------ --------------- -------------- ------------ ---------------- --------------

       0.25M            cn01          0.25M        0.02M             cn01          0.02M
       0.23M            cn01          0.23M        0.02M             cn01          0.02M
```

Állapotinformáció: futó job

```
BUDAPEST[login] ~ (0)$ sjobcheck 15419
Hostname                      LOAD                          CPU             Gexec
 CPUs (Procs/Total) [     1,     5, 15min] [  User,  Nice, System, Idle, Wio]
cn04.budapest.hpc.niif.hu     24 (   26/  857) [ 25.18, 25.27, 25.18] [  99.9,   0.0,   0.1,   0.0,   0.0] OFF
```

Projekt információk

```
BUDAPEST[login] ~ (0)$ sbank balance statement -a niif
User              Usage |          Account       Usage | Account Limit   Available (CPU hrs)
---------- ----------- + ---------------- ----------- + ------------- -----------

ganzler               0 |             niif           0 |        41,000      41,000
htom *                0 |             niif           0 |        41,000      41,000
kzoli                 0 |             niif           0 |        41,000      41,000
martoni               0 |             niif           0 |        41,000      41,000
portal                0 |             niif           0 |        41,000      41,000
roczei                0 |             niif           0 |        41,000      41,000
szigeti               0 |             niif           0 |        41,000      41,000
```

Állapotinformációk: felhasználás

```
BUDAPEST[login] ~ (0)$ susage niif -4 month
--------------------------------------------------------------------------------
Cluster/Account/User Utilization 2013-10-04T00:00:00 - 2014-02-03T23:59:59 (10630800 secs)
Time reported in CPU Minutes
--------------------------------------------------------------------------------
  Cluster         Account     Login     Proper Name       Used
--------- --------------- --------- --------------- ----------
 budapest            niif                              255566
 budapest            niif      htom   Hornos Tamás     223670
 budapest            niif    roczei    Roczei Gabor      31896
```

srun példa

- –label: TASK ID

```
BUDAPEST[login] ~ (0)$ srun -ptest -n2 --label hostname
srun: job 15791 queued and waiting for resources
srun: job 15791 has been allocated resources
0: cn32
1: cn32

BUDAPEST[login] ~ (1)$ srun -N 2 --exclusive --label hostname
srun: job 15792 queued and waiting for resources
srun: job 15792 has been allocated resources
1: cn20
0: cn19
```

salloc példa

```
BUDAPEST[login] ~ (1)$ salloc --ntasks=48 --time=10 bash
salloc: Pending job allocation 15793
salloc: job 15793 queued and waiting for resources
salloc: job 15793 has been allocated resources
salloc: Granted job allocation 15793
[login.budapest:~]$ env|grep SLURM
SLURM_NODELIST=cn[19-20]
SLURM_BANK_HOME=/opt/nce/packages/global/sbank/1.2
SLURM_NODE_ALIASES=(null)
SLURM_MEM_PER_CPU=1000
SLURM_NNODES=2
SLURM_JOBID=15793
SLURM_BANK=sbank/1.2
SLURM_NTASKS=48
SLURM_TASKS_PER_NODE=24(x2)
SLURM_JOB_ID=15793
SLURM_SUBMIT_DIR=/fs01/home/htom
SLURM_NPROCS=48
SLURM_JOB_NODELIST=cn[19-20]
SLURM_JOB_CPUS_PER_NODE=24(x2)
SLURM_SUBMIT_HOST=login.budapest.hpc.niif.hu
SLURM_JOB_NUM_NODES=2
[login.budapest:~]$ srun --label hostname
37: cn20
29: cn20

[login.budapest:~]$ squeue -u htom; exit
JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
15793      prod     bash     htom  R       1:25      2 cn[19-20]
```

Task ID szerinti szeparáció

master.conf

```
#id #app    #args
0   master
1-4 slave   --rank=%o

srun --ntasks=5 --multi-prog master.conf
```

Statisztika

```
BUDAPEST[login] ~ (0)$ sstat --format=AveCPU,AvePages,AveRSS,AveVMSize,JobID -j 14973
    AveCPU   AvePages     AveRSS  AveVMSize        JobID
---------- ---------- ---------- ---------- ------------
  11:13:50          0     83288K     95800K 14973.0

BUDAPEST[login] ~ (0)$ sacct -S 2013-12-01 -E 2013-12-30
       JobID    JobName  Partition    Account  AllocCPUS      State ExitCode
------------ ---------- ---------- ---------- ---------- ---------- --------
416                 mpi     normal       niif         48  CANCELLED      0:0
8490         array_test       test       niif          2 CANCELLED+      0:0
8490.batch        batch                  niif          1  CANCELLED     0:15
8490.0       MOPAC2012+                  niif          1     FAILED    174:0
8490.1       MOPAC2012+                  niif          1     FAILED    174:0

sacct -l
```

Statisztika

```
BUDAPEST[login] ~ (1)$ sreport cluster Utilization
--------------------------------------------------------------------------------
Cluster Utilization 2014-02-03T00:00:00 - 2014-02-03T23:59:59 (86400*cpus secs)
Time reported in CPU Minutes
--------------------------------------------------------------------------------
  Cluster    Allocated       Down  PLND Down          Idle  Reserved     Reported
--------- ------------ ---------- ---------- ------------ --------- ------------
 budapest       558024         51          0       526213     21631      1105920

sreport: user TopUsage
--------------------------------------------------------------------------------
Top 10 Users 2014-02-03T00:00:00 - 2014-02-03T23:59:59 (86400 secs)
Time reported in CPU Minutes
--------------------------------------------------------------------------------
  Cluster     Login     Proper Name         Account       Used
--------- --------- --------------- --------------- ----------
 budapest      geza    Dr Odor Geza        fazisata     508554
 budapest  jeszenoi Dr. Jeszenoi N+          molkot      41191
 budapest      stef  Horvath Istvan          molkot       8280
```

CPU óra becslése

- array jobok: minden szál számít
- idő megadása kötelező (minimális és maximális)
- rövidebb job hamarabb indul

```
BUDAPEST[login] ~ (0)$ sestimate -N 8 -t 2-10:00:00
Estimated CPU hours: 11136
```

Job szkriptek

- sbatch parancs
- parancssori kapcsolók vagy az #SBATCH direktíva
- kötelező paraméterek:

```
#!/bin/bash
#SBATCH -A ACCOUNT
#SBATCH --job-name=NAME
#SBATCH --time=TIME
APP
```

Job szkript beküldése / törlése

```
BUDAPEST[login] omp (0)$ sbatch slurm.sh
Submitted batch job 15795

BUDAPEST[login] omp (0)$ squeue -u htom
JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
15795      prod      omp     htom  R       0:58      1 cn15

scancel JOBID
scontrol hold / release JOBID
```

Ütemezési sorok és QOS

```
#SBATCH --partition=test
#SBATCH --qos=lowpri

BUDAPEST[login] omp (0)$ sacctmgr list qos Format="Name,Priority,Preempt,PreemptMode,UsageFactor"
      Name   Priority    Preempt PreemptMode UsageFactor
---------- ---------- ---------- ----------- -----------
    normal       1000     lowpri     cluster    1.000000
    lowpri        100                requeue    0.500000
```

Memória és értesítés

```
#SBATCH --mem-per-cpu=MEMORY
#SBATCH --mail-type=ALL
#SBATCH --mail-user=EMAIL
```

Array jobok

```
BUDAPEST[login] arrayjob (0)$ cat slurm.sh
#!/bin/bash
#SBATCH -A niif
#SBATCH --time=00:30:00
#SBATCH --qos=lowpri
#SBATCH --job-name=array
#SBATCH --array=1-120
srun envtest.sh 6000

BUDAPEST[login] arrayjob (0)$ cat envtest.sh
#!/bin/bash
echo -n "$(hostname): "
echo "PROCID=$SLURM_PROCID  NODEID=$SLURM_NODEID  NNODES=$SLURM_NNODES  LOCALID=$SLURM_LOCALID  NPROCS=$SLURM_NPROCS  NTASKS=$SLURM_NTASKS"
set | grep SLURM
echo ""
sleep ${1:-600}

BUDAPEST[login] arrayjob (0)$ cat slurm-1228_48.out
cn06: PROCID=0  NODEID=0  NNODES=1  LOCALID=0  NPROCS=1  NTASKS=1
SLURMD_NODENAME=cn06
SLURM_ARRAY_JOB_ID=1228
SLURM_ARRAY_TASK_ID=48
```

SLURM környezeti változók

```
SLURMD_NODENAME=cn06
SLURM_ARRAY_JOB_ID=1228
SLURM_ARRAY_TASK_ID=48
SLURM_BANK=sbank/1.2
SLURM_BANK_HOME=/opt/nce/packages/global/sbank/1.2
SLURM_CHECKPOINT_IMAGE_DIR=/fs01/home/htom/slurm/test/arrayjob
SLURM_CPUS_ON_NODE=1
SLURM_DISTRIBUTION=cyclic
SLURM_GTIDS=0
SLURM_JOBID=1275
SLURM_JOB_CPUS_PER_NODE=1
SLURM_JOB_ID=1275
SLURM_JOB_NAME=array
SLURM_JOB_NODELIST=cn06
SLURM_JOB_NUM_NODES=1
SLURM_LAUNCH_NODE_IPADDR=192.168.64.6
SLURM_LOCALID=0
SLURM_MEM_PER_CPU=1000
SLURM_NNODES=1
SLURM_NODEID=0
SLURM_NODELIST=cn06
SLURM_NPROCS=1
SLURM_NTASKS=1
SLURM_PRIO_PROCESS=0
SLURM_PROCID=0
SLURM_SRUN_COMM_HOST=192.168.64.6
SLURM_SRUN_COMM_PORT=36477
SLURM_STEPID=0
SLURM_STEP_ID=0
SLURM_STEP_LAUNCHER_PORT=36477
SLURM_STEP_NODELIST=cn06
SLURM_STEP_NUM_NODES=1
SLURM_STEP_NUM_TASKS=1
SLURM_STEP_TASKS_PER_NODE=1
SLURM_SUBMIT_DIR=/fs01/home/htom/slurm/test/arrayjob
SLURM_SUBMIT_HOST=cn06
SLURM_TASKS_PER_NODE=1
SLURM_TASK_PID=17921
SLURM_TOPOLOGY_ADDR=cn06
SLURM_TOPOLOGY_ADDR_PATTERN=node
```

Serial jobok

```
#!/bin/bash
#SBATCH --job-name=serial
#SBATCH --time=24:30:00
#SBATCH -n 3
#SBATCH --partition=test
srun -n 1 program input1 &
srun -n 1 program input2 &
srun -n 1 program input3
wait
```

master.conf

```
#id #app    #args
0   master
1-4 slave   --rank=%o
```

```
srun --ntasks=5 --multi-prog master.conf
```

MPI jobok

- –ntasks-per-node=core-ok száma
- ha nem, akkor –exclusive

```
#!/bin/bash
#SBATCH -A niif
#SBATCH --job-name=mpi
#SBATCH -N 2
#SBATCH --time=00:30:00
#SBATCH --ntasks-per-node=24
#SBATCH -o slurm.out
mpirun -v -report-bindings ./a.out
```

CPU binding: openmpi

```
mpirun --bind-to-core --bycore
[cn05:05493] MCW rank 0 bound to socket 0[core 0]: [B . . . . . . . . . .][. . . . . . . . . . .]
[cn05:05493] MCW rank 1 bound to socket 0[core 1]: [. B . . . . . . . . .][. . . . . . . . . . .]
[cn05:05493] MCW rank 2 bound to socket 0[core 2]: [. . B . . . . . . . .][. . . . . . . . . . .]
[cn05:05493] MCW rank 3 bound to socket 0[core 3]: [. . . B . . . . . . .][. . . . . . . . . . .]

mpirun --bind-to-core --bysocket
[cn05:05659] MCW rank 0 bound to socket 0[core 0]: [B . . . . . . . . . .][. . . . . . . . . . .]
[cn05:05659] MCW rank 1 bound to socket 1[core 0]: [. . . . . . . . . . .][B . . . . . . . . . .]
[cn05:05659] MCW rank 2 bound to socket 0[core 1]: [. B . . . . . . . . .][. . . . . . . . . . .]
[cn05:05659] MCW rank 3 bound to socket 1[core 1]: [. . . . . . . . . . .][. B . . . . . . . . .]

mpirun --bind-to-core --bynode
[cn05:05904] MCW rank 0 bound to socket 0[core 0]: [B . . . . . . . . . .][. . . . . . . . . . .]
[cn05:05904] MCW rank 2 bound to socket 0[core 1]: [. B . . . . . . . . .][. . . . . . . . . . .]
[cn06:05969] MCW rank 1 bound to socket 0[core 0]: [B . . . . . . . . . .][. . . . . . . . . . .]
[cn06:05969] MCW rank 3 bound to socket 0[core 1]: [. B . . . . . . . . .][. . . . . . . . . . .]
```

OpenMP jobok

- OMP_NUM_THREADS=core-ok száma (a SLURM nem állítja be)
- ha nem, akkor –exclusive

```
#!/bin/bash
#SBATCH -A niif
#SBATCH --job-name=omp
#SBATCH --time=01:00:00
#SBATCH -N 1
#SBATCH -o slurm.out
OMP_NUM_THREADS=24 ./a.out#!/bin/bash
```

MPI + OpenMP

- –exclusive kötelező
- –ntasks-per-node -> MPI szálak
- OMP_NUM_THREADS -> OpenMP szálak / MPI szál

```
#!/bin/bash
#SBATCH -A niif
#SBATCH --job-name=mpi
#SBATCH -N 2
#SBATCH --time=00:30:00
#SBATCH --exclusive
#SBATCH --ntasks-per-node=1
#SBATCH -o slurm.out
export OMP_NUM_THREADS=24
mpirun ./a.out
```