Networkshop 2016
Debreceni Egyetem

HPC Tutoriál

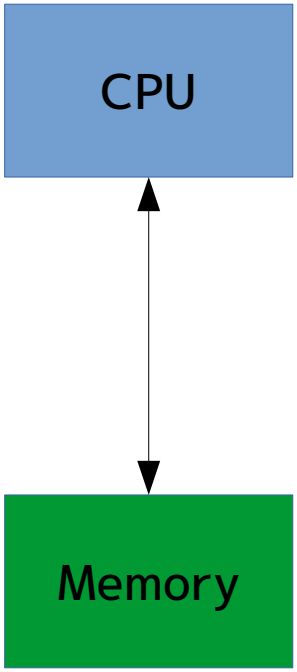
Székelyi Szabolcs
<szekelyi@niif.hu>

Tartalom

- Szupergép típusok
- Párhuzamosítás
- Ütemező: SLURM
- Vizualizáció

A teljesítmény architekturális evolúciója

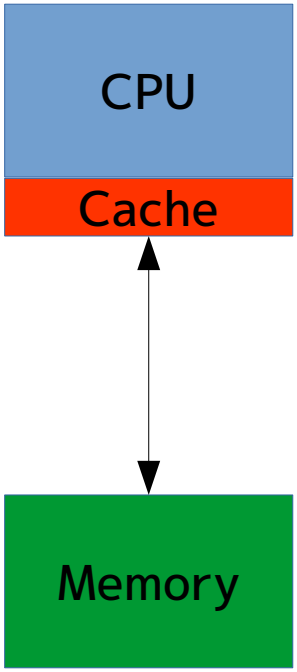
Single-processing



Lassú memória

A teljesítmény architekturális evolúciója

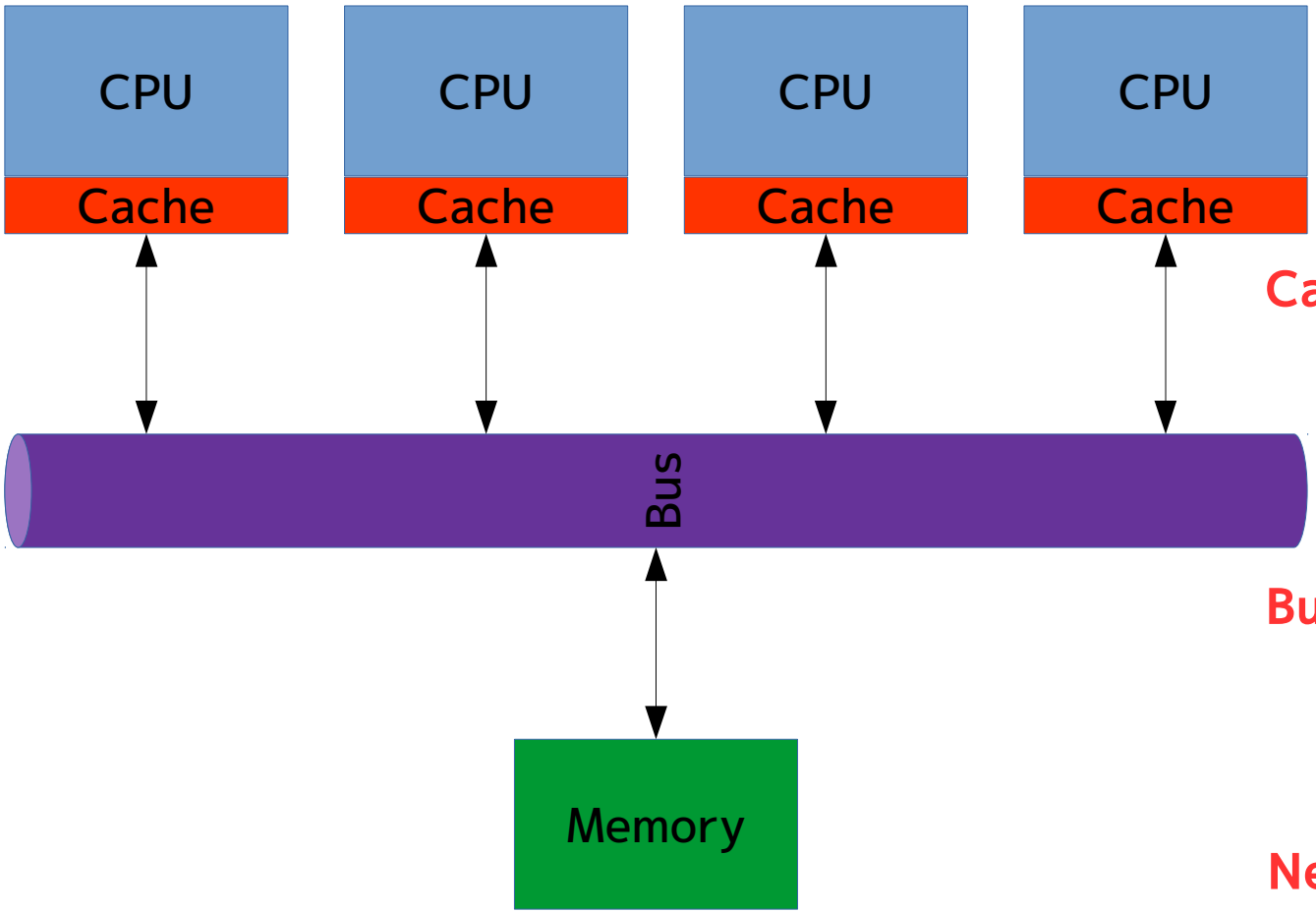
Single-processing + cache



Nem skálázódik

A teljesítmény architekturális evolúciója

Symmetric Multi-Processing (SMP)

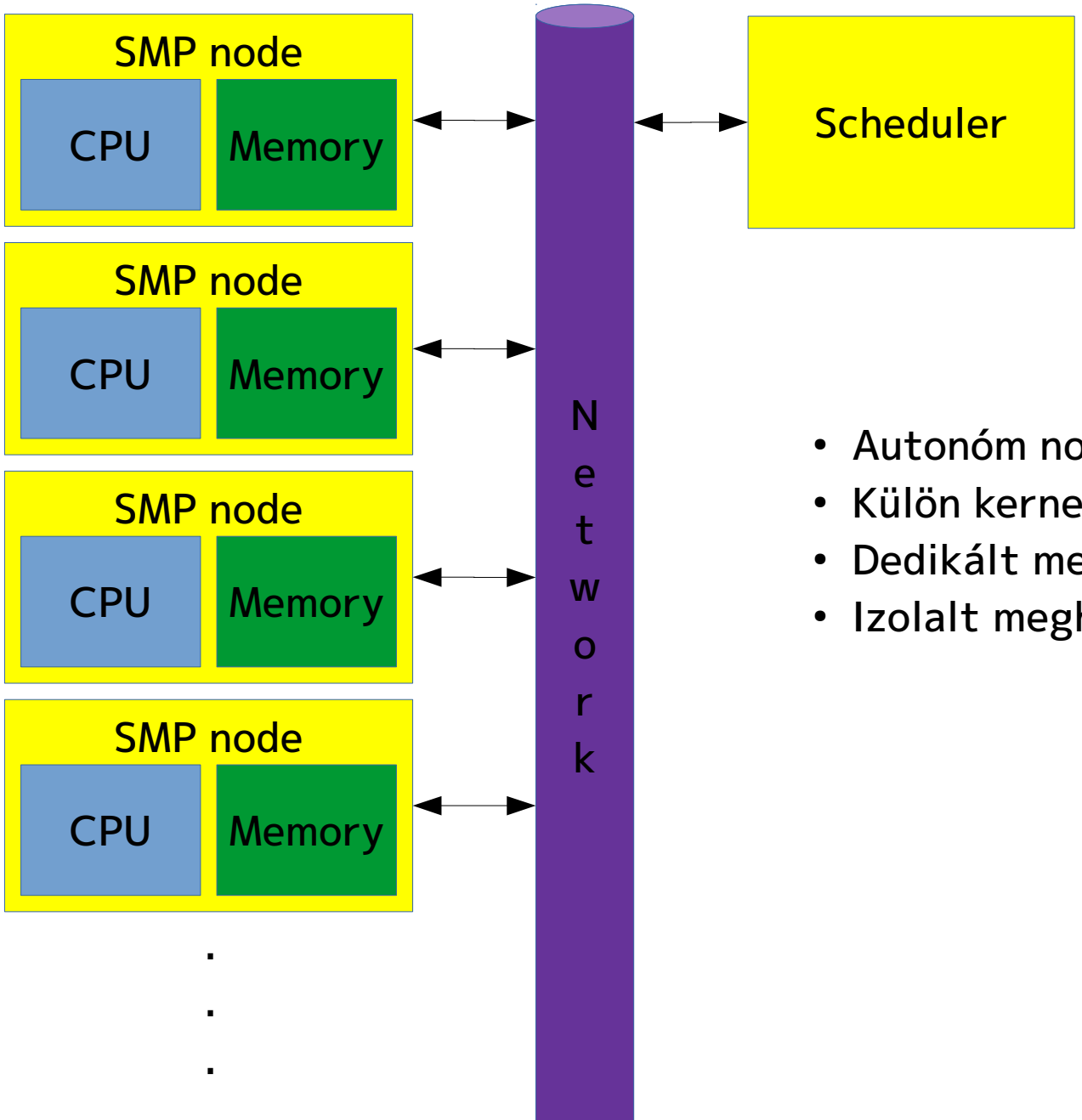


Cache koherencia

Busz arbitráció

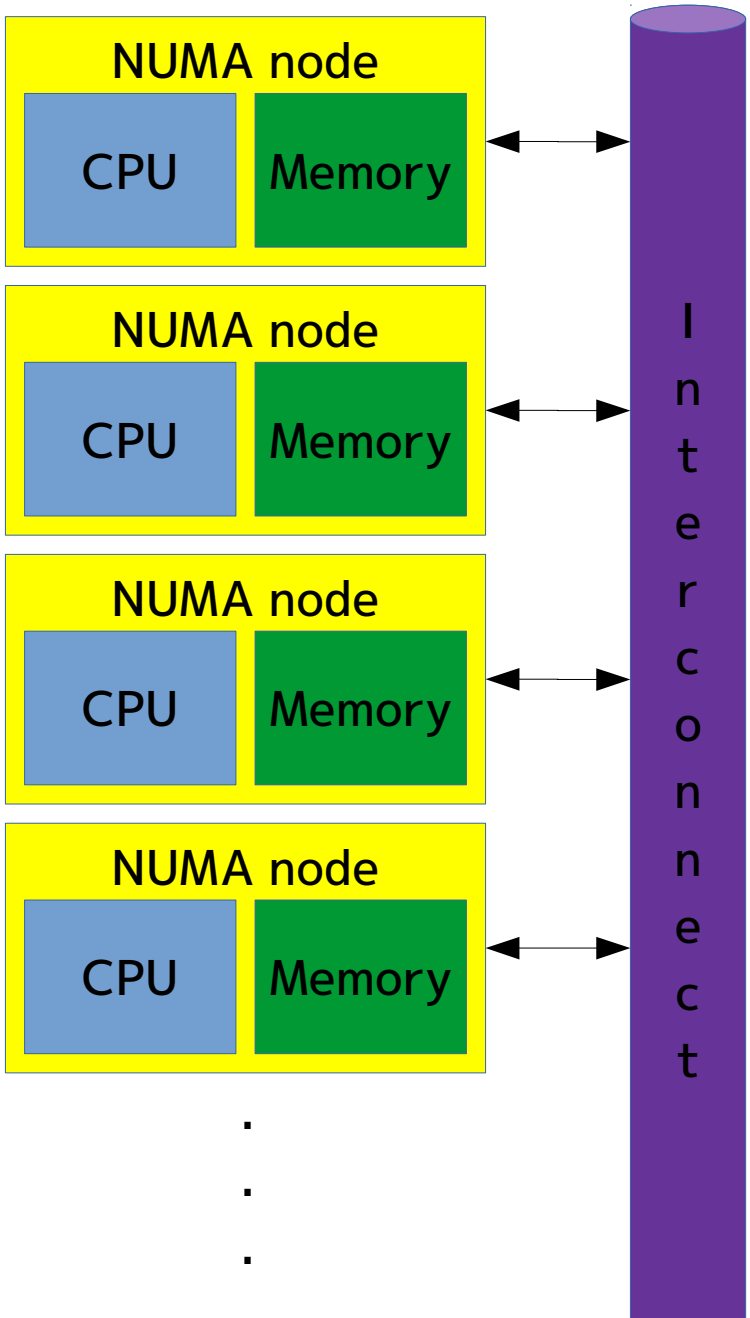
Nem skálázódik

SMP Cluster



- Autonóm node-ok
- Külön kernel
- Dedikált memóriatartomány
- Izolalt meghibásodás

Non-Uniform Memory Access (NUMA)



- Függő “node”-ok
- Egy kernel
- Egyesített memóriatartomány
  - Egymáséba belenyúlhatnak
- Globális meghibásodás
- “Távolság”

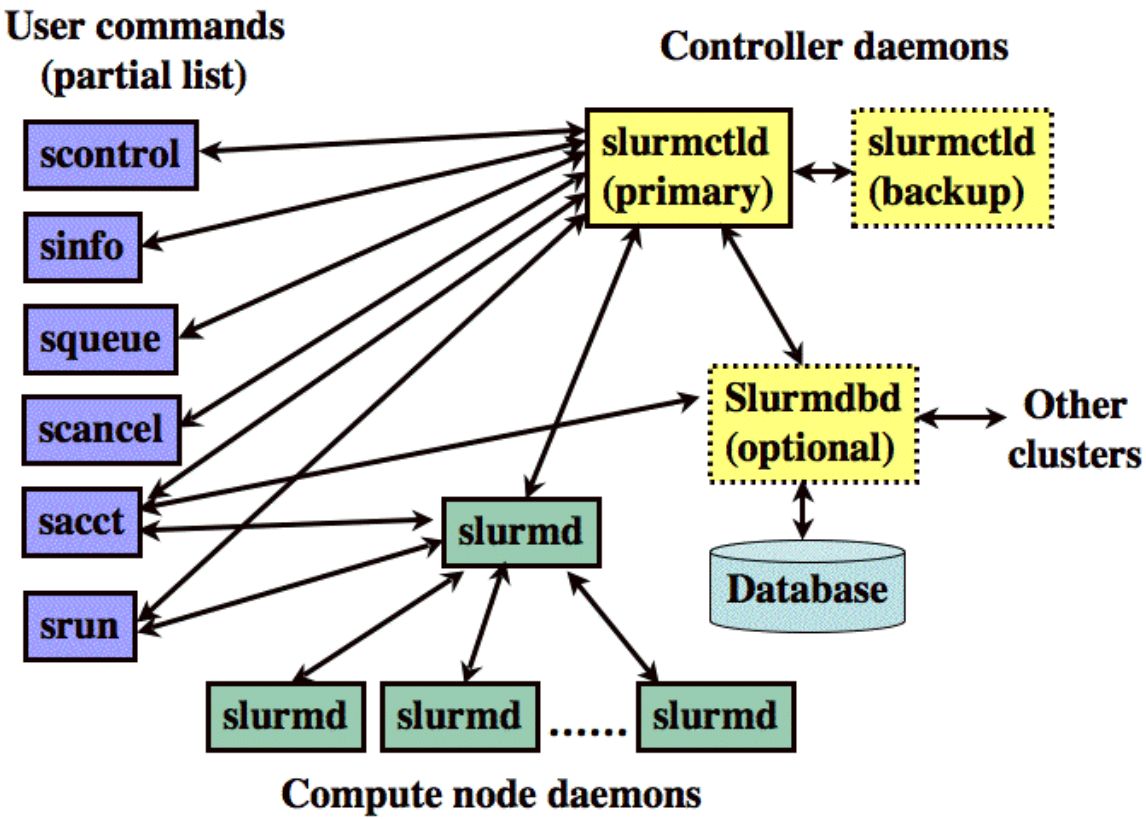
Párhuzamosítás

- Párhuzamos loopok, fork, join, kritikus szakaszok, atomi műveletek, barrier, privát ciklusváltozók, stb.
- OpenMP
  - node-on belül
- MPI
  - node-ok között
  - hardver-/gyátróspecifikus implementációk (Intel MPI, SGI MPT)
- OpenMP + MPI

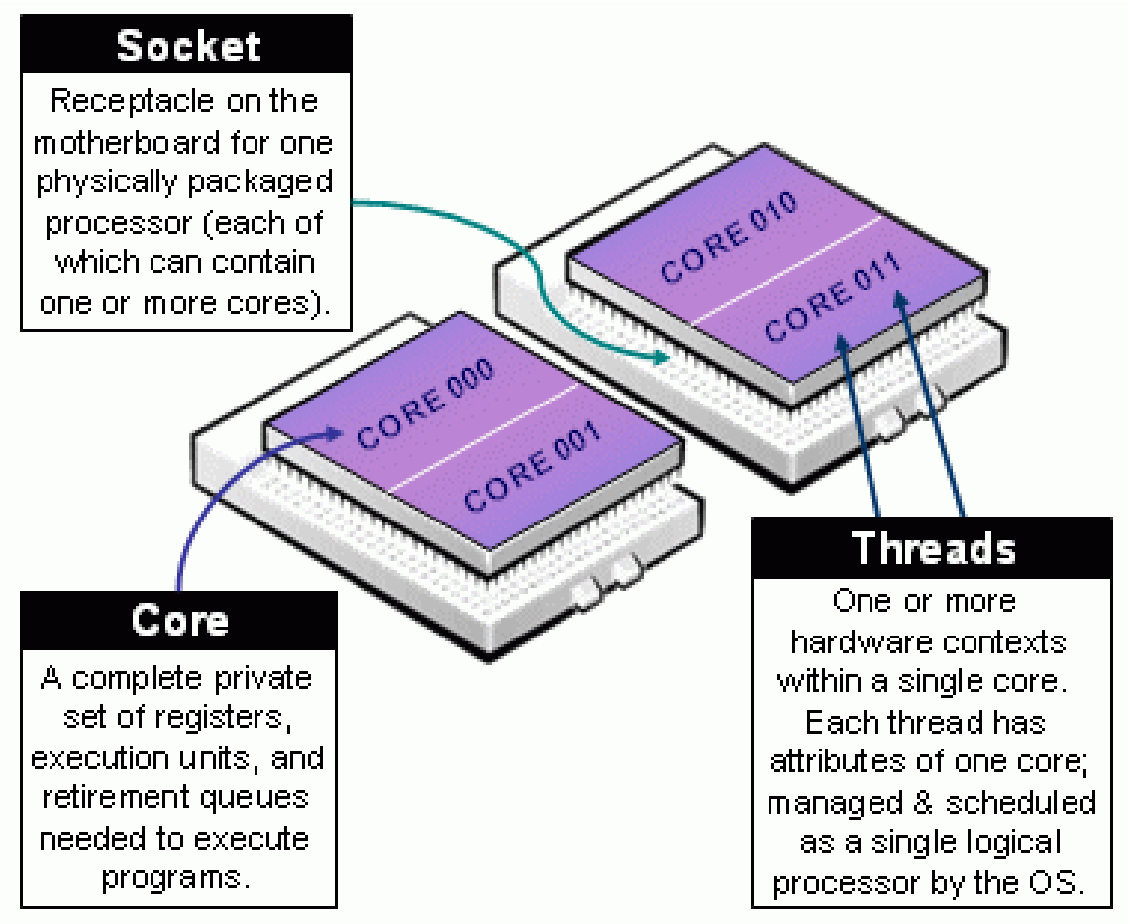
Ütemező: SLURM

- Simple Linux Utility for Resource Management
- Számítási erőforrások leírása
- Párhuzamos futtatási környezet biztosítása
- Feladat-ütemezés: `sbatch`

Simple Linux Utility for Resource Management (SLURM)



Számítási erőforrások leírása



Számítási erőforrások leírása

- Socket/Core/Thread
- Memória (automatikus)
- Features (bináris: van vagy nincs)
- General Resources (GRES): megszámolható erőforrások
  - GPU [Nvidia K20/K40]
  - MIC [Intel Xeon Phi]
- Partició: node-ok csoportja, külön ütemezési sorral
- QoS
  - Prioritás
  - Maximális futási idő, jobok száma, node-ok száma, …
  - Eltérő árazás
  - normal, fast, lowpri, test?

Párhuzamos futtatási környezet biztosítása

sbatch paraméterezés

- `--job-name=<name>`: job neve
- `--account=<account>`: projekt neve
- `--partition=<P*>`: adott partíció(k) használata
- `--qos=<qos>`: adott QoS használata
- `--time=<time>`: maximum mennyi ideig futhat a job
- `--array=<spec>`: array jobok indexei
- `--mpi=<mpi_type>`: használni kívánt MPI típusa

Párhuzamos futtatási környezet biztosítása

sbatch paraméterezés

- `--nodes=<min[-max]>`: allokálandó node-ok száma
- `--exclusive`: node-ok nem megoszthatók
- `--extra-node-info=<S[:C[:T]]>`: minimális socket / core / thread igény
  - `--sockets-per-node=<N>`
  - `--cores-per-socket=<N>`
  - `--threads-per-core=<N>`
- `--mincpus=<N>`: CPU-k minimális száma node-onként
- `--gres=<list>`: Genreal Resource-okbol milyet és mennyit kérünk
- `--mem-per-cpu=<MB>`: igényelt memória processzoronként

Párhuzamos futtatási környezet biztosítása: srun

- `--ntasks=<N>`: párhuzamos feladatok száma
  - `--[n]tasks-per-node`
  - `--ntasks-per-socket`
  - `--ntasks-per-core`
- `--cpus-per-task=<N>`: egy taszk hány szálon fog futni
- `--hint`:
  - `compute_bound`: egy socketben minden core használata
  - `memory_bound`: egy socketben csak egy core használata
  - `[no]multithread`: threadek [nem-]használata
- `--distribution=<node:core>`: taszkok elosztása a rendelkezésre álló nodeok/processzorok között
  - `cyclic`: round-robin, minél jobban szétkenve
  - `block`: addig foglalva, amig be nem telik, aztán továbblépve

srun példa

```
[scheduler] ~ (0)$ srun --label --ntasks=8 hostname | sort
srun: job 59486 queued and waiting for resources
srun: job 59486 has been allocated resources
0: cn001
1: cn001
2: cn001
3: cn001
4: cn001
5: cn001
6: cn001
7: cn001
[scheduler] ~ (0)$
```

srun példa

```
[scheduler] ~ (0)$ srun -l -n 8 --tasks-per-node=2 hostname
| sort
srun: job 59487 queued and waiting for resources
srun: job 59487 has been allocated resources
0: cn005
1: cn005
2: cn006
3: cn006
4: cn007
5: cn007
6: cn008
7: cn008
[scheduler] ~ (0)$
```

srun példa

```
[scheduler] ~ (0)$ srun -l -n 8 --tasks-per-node=2
--distribution=cyclic hostname | sort
srun: job 59488 queued and waiting for resources
srun: job 59488 has been allocated resources
0: cn005
1: cn006
2: cn007
3: cn008
4: cn005
5: cn006
6: cn007
7: cn008
[scheduler] ~ (0)$
```

Komplex allokáció

- 3 node, 2 socket/node, 4 core/socket
- `srun --ntasks=18 --ntasks-per-node=6 --distribution=cyclic:block hostname`

Node	node0								node1								node2							
Socket	0				1				0				1				0				1			
Allokált core	4				2				4				2				4				2			
Core	0	1	2	3	4	5	6	7	0	1	2	3	4	5	6	7	0	1	2	3	4	5	6	7
Taszk	0	3	6	9	12	15			1	4	7	10	13	16			2	5	8	11	14	17		

Komplex allokáció

- 3 node, 2 socket/node, 4 core/socket
- `srun --ntasks=18 --ntasks-per-node=6 --distribution=cyclic:cyclic hostname`

Node	node0								node1								node2							
Socket	0				1				0				1				0				1			
Allokált core	3				3				3				3				3				3			
Core	0	1	2	3	4	5	6	7	0	1	2	3	4	5	6	7	0	1	2	3	4	5	6	7
Taszk	0	6	12		3	9	15		1	7	13		4	10	16		2	8	14		5	11	17	

Feladat-ütemezés

- Job beküldés: `sbatch <opciók> script`
  - Praktikusan csak ezt használjuk
- Várakozás az erőforrásokra
- Erőforrás-allokáció (`salloc`)
  - Egy példányban elindítja a job scriptet
- Párhuzamos feladatfuttatás (`srun`)

Job scriptek

- `sbatch <opciók> script`
- `#SBATCH` markup parancssori opciók helyett/mellett
- Kötelező opciók:
  - `--account`
  - `--job-name`
  - `--time`
- **Egy példányban** indul el az első kijelölt node-on
- Tipikusan `srun`-t hív

```
#!/bin/bash
#SBATCH --account hpcteszt
#SBATCH --job-name teszt
#SBATCH --time 0:30

srun hostname
```

Információ a clusterről

- Node-ok, partíciók: `sinfo`
- Várakozási sor: `squeue`
- QoS: `sacctmgr show qos`
- Prioritások: `sprio`
- `scontrol show`
  - `job [job_id]`
  - `node [node_name]`
  - `partition [partition_name]`
  -

Partíciók

```
[visual] ~ (0)$ sinfo
PARTITION     AVAIL  TIMELIMIT  NODES  STATE NODELIST
prod-gpu-k40     up 7-00:00:00      2    mix cn[003-004]
prod-gpu-k40     up 7-00:00:00     14   idle cn[011-012,021-022,025-026,033-034,037-038,041-04
prod-gpu-k20*    up 7-00:00:00      1  down* cn014
prod-gpu-k20*    up 7-00:00:00      1    mix cn010
prod-gpu-k20*    up 7-00:00:00      1  alloc cn001
prod-gpu-k20*    up 7-00:00:00     65   idle cn[002,005-009,013,015-020,023-024,027-032,035-03
prod-phi         up 7-00:00:00      1   drng apollo001
prod-phi         up 7-00:00:00     20  alloc apollo[002-003,005-014,017-018,023-027,041]
prod-phi         up 7-00:00:00     23   idle apollo[004,015-016,019-022,028-040,042-044]
[visual] ~ (0)$
```

Partíciók – részletek

```
[visual] ~ (0)$ scontrol show partition
PartitionName=prod-gpu-k40
   AllowGroups=ALL AllowAccounts=ALL AllowQos=ALL
   AllocNodes=ALL Default=NO
   DefaultTime=01:00:00 DisableRootJobs=NO GraceTime=0 Hidden=NO
   MaxNodes=UNLIMITED MaxTime=7-00:00:00 MinNodes=1 LLN=NO MaxCPUsPerNode=UNLIMITED
   Nodes=cn[003,004,011,012,021,022,025,026,033,034,037,038,041,042,051,052]
   Priority=1 RootOnly=NO ReqResv=NO Shared=NO PreemptMode=REQUEUE
   State=UP TotalCPUs=256 TotalNodes=16 SelectTypeParameters=N/A
   DefMemPerNode=UNLIMITED MaxMemPerNode=UNLIMITED

PartitionName=prod-gpu-k20
   AllowGroups=ALL AllowAccounts=ALL AllowQos=ALL
   AllocNodes=ALL Default=YES
   DefaultTime=01:00:00 DisableRootJobs=NO GraceTime=0 Hidden=NO
   MaxNodes=UNLIMITED MaxTime=7-00:00:00 MinNodes=1 LLN=NO MaxCPUsPerNode=UNLIMITED
   Nodes=cn[001-2,005-10,013-20,023-24,027-32,035-36,039-40,043-50,053-84]
   Priority=1 RootOnly=NO ReqResv=NO Shared=NO PreemptMode=REQUEUE
   State=UP TotalCPUs=1088 TotalNodes=68 SelectTypeParameters=N/A
   DefMemPerNode=UNLIMITED MaxMemPerNode=UNLIMITED
```

Várakozási sor

```
[visual] ~ (0)$ squeue
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
             59478 prod-gpu-   wt55_4  hegedus  R    8:47:10      1 cn001
             59479 prod-gpu-   wt55_5  hegedus  R    8:47:10      1 cn001
             59480 prod-gpu-   wt55_7  hegedus  R    8:47:10      1 cn010
             59514 prod-gpu-      MDS     fkun  R    8:47:10      1 cn003
             59515 prod-gpu-      MDS     fkun  R    8:47:10      1 cn003
             59516 prod-gpu-      MDS     fkun  R    8:47:10      1 cn003
             59517 prod-gpu-      MDS     fkun  R    8:47:10      1 cn004
             59518 prod-gpu-      MDS     fkun  R    8:47:10      1 cn004
             59451  prod-phi   qemd5K     sule PD       0:00      5 (AssociationJobLimit)
             59156  prod-phi GMX51_GP  pepfold  R 1-02:04:05      1 apollo018
             59154  prod-phi GMX51_GP  pepfold  R 3-18:24:06      1 apollo014
             59113  prod-phi GMX51_GP  pepfold  R 4-20:11:10      1 apollo017
             59111  prod-phi GMX51_GP  pepfold  R 4-22:19:07      1 apollo002
             59450  prod-phi qetemdvv     sule  R 3-20:24:10      5 apollo[009-013]
             59159  prod-phi   qetemd     sule  R 4-19:48:18      5 apollo[023-027]
             59513  prod-phi   qevv5K     sule  R 1-00:56:59      4 apollo[005-008]
             59110  prod-phi GMX51_GP  pepfold  R 4-23:36:55      1 apollo001
             59096  prod-phi GMX51_GP  pepfold  R 4-23:45:38      1 apollo041
```

Job információk

```
[visual] ~ (0)$ scontrol show job 59110
JobId=59110 Name=GMX51_GPU
   UserId=pepfold(11062) GroupId=pepfold(11062)
   Priority=10008 Nice=0 Account=peptid QOS=normal
   JobState=RUNNING Reason=None Dependency=(null)
   Requeue=1 Restarts=0 BatchFlag=1 ExitCode=0:0
   RunTime=4-23:50:09 TimeLimit=7-00:00:00 TimeMin=N/A
   SubmitTime=2016-03-24T10:30:46 EligibleTime=2016-03-24T10:30:46
   StartTime=2016-03-24T10:30:46 EndTime=2016-03-31T11:30:46
   PreemptTime=None SuspendTime=None SecsPreSuspend=0
   Partition=prod-phi AllocNode:Sid=visual:15474
   ReqNodeList=(null) ExcNodeList=(null)
   NodeList=apollo001
   BatchHost=apollo001
   NumNodes=1 NumCPUs=24 CPUs/Task=6 ReqB:S:C:T=0:0:*:*
   Socks/Node=* NtasksPerN:B:S:C=4:0:*:* CoreSpec=0
   MinCPUsNode=24 MinMemoryCPU=2500M MinTmpDiskNode=0
   Features=(null) Gres=(null) Reservation=(null)
   Shared=0 Contiguous=0 Licenses=(null) Network=(null)
   Command=/Lustre01/home/pepfold/entropy_calc/1micro/solo/6/1mu_6.qsub
```

Vizualizáció

- Dedikált grafikus node-ok erős GPU-val
- VNC over SSH alapú hozzáférés
- VirtualGL + TurboVNC

Vizualizáció

```
[visual] ~ (0)$ vncserver

You will require a password to access your desktops.

Password:
Verify:
Would you like to enter a view-only password (y/n)? n

Desktop 'TurboVNC: visual:4 (szekelyi)' started on display visual:4

Creating default startup script /home/szekelyi/.vnc/xstartup.turbovnc
Starting applications specified in /home/szekelyi/.vnc/xstartup.turbovnc
Log file is /home/szekelyi/.vnc/visual:4.log

[visual] ~ (0)$
ssh> -L 5900:localhost:5904
Forwarding port.

[visual] ~ (0)$
```

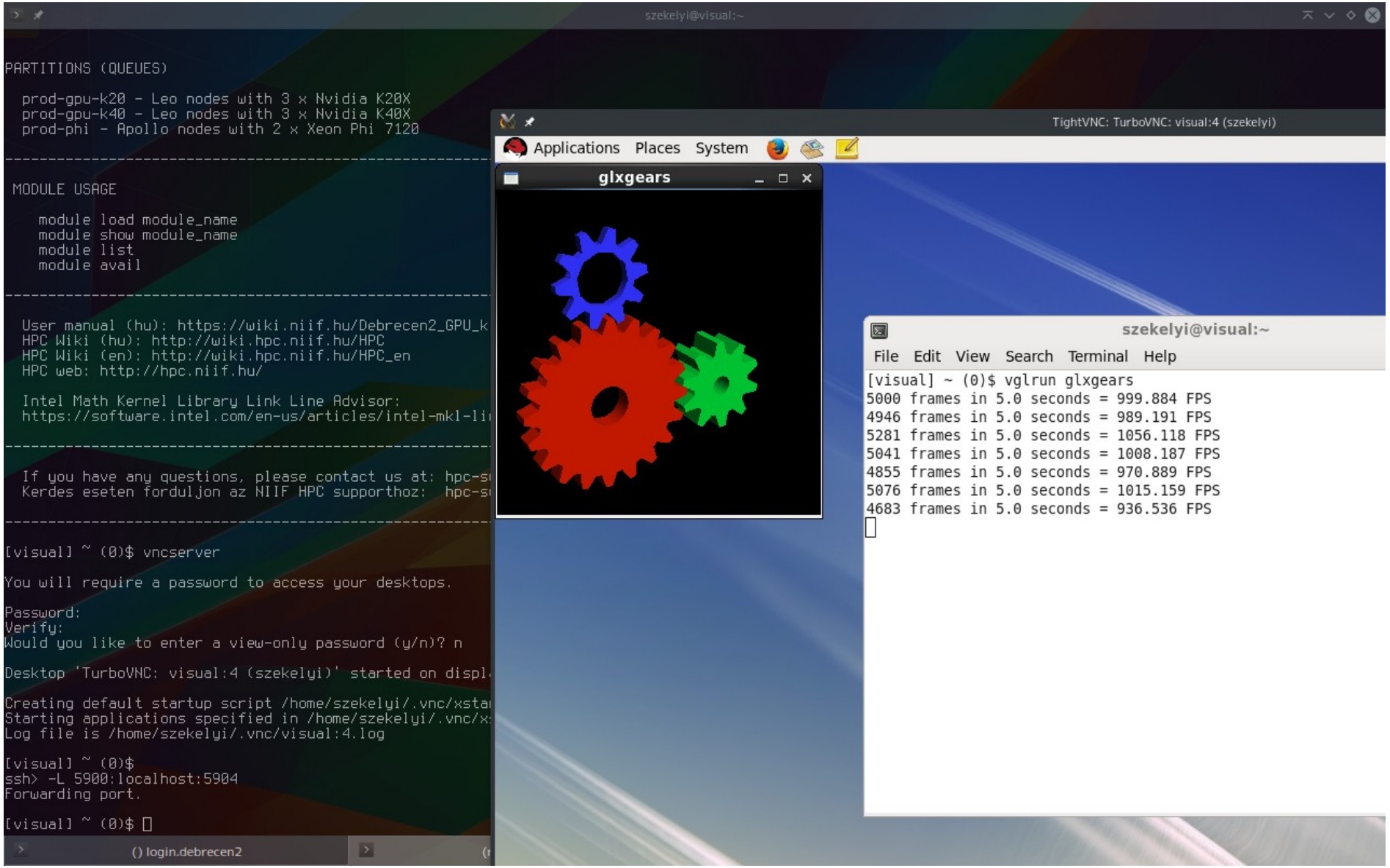
~C

5900 + display

Vizualizáció

```
cc@mranderson:~$ vncviewer localhost
Connected to RFB server, using protocol version 3.8
Enabling TightVNC protocol extensions
Performing standard VNC authentication
Password:
Authentication successful
Desktop name "TurboVNC: visual:4 (szekelyi)"
VNC server default format:
  32 bits per pixel.
  Least significant byte first in each pixel.
  True colour: max red 255 green 255 blue 255, shift red 16 green 8 blue 0
Using default colormap which is TrueColor.  Pixel format:
  32 bits per pixel.
  Least significant byte first in each pixel.
  True colour: max red 255 green 255 blue 255, shift red 16 green 8 blue 0
Same machine: preferring raw encoding
```

Vizualizáció



Vizualizáció

```
[visual] ~ (2)$ vncserver -kill :4
Killing Xvnc process ID 9640
[visual] ~ (0)$
```

Networkshop 2016
Debreceni Egyetem

HPC Tutoriál

Székelyi Szabolcs
<szekelyi@niif.hu>