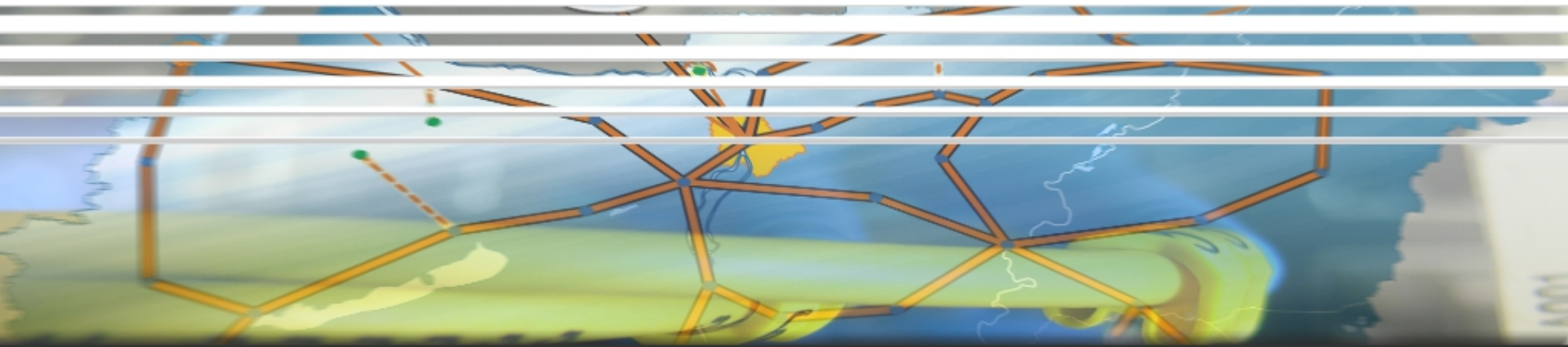


Párhuzamos programok futásának kiértékelése Scalasca profiler segítségével



2014. Április 24.

Pécs, Networkshop 2014

Róczei Gábor

roczei@niif.hu

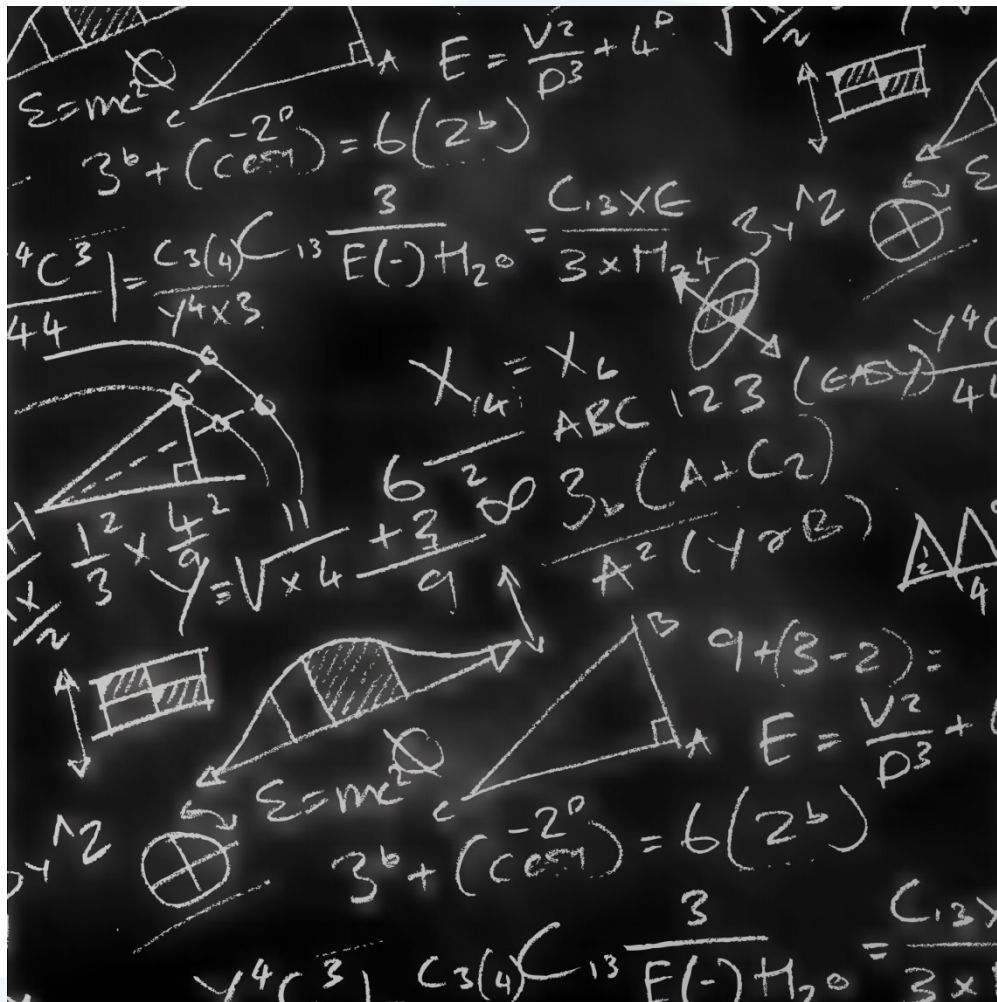


Főbb témák

- Miért használjunk superszámítógépet?!
- Alapfogalmak
- Miért van szükség profiler programra?!
- Scalasca profiler

Miért használjunk szuperszámítógépet?

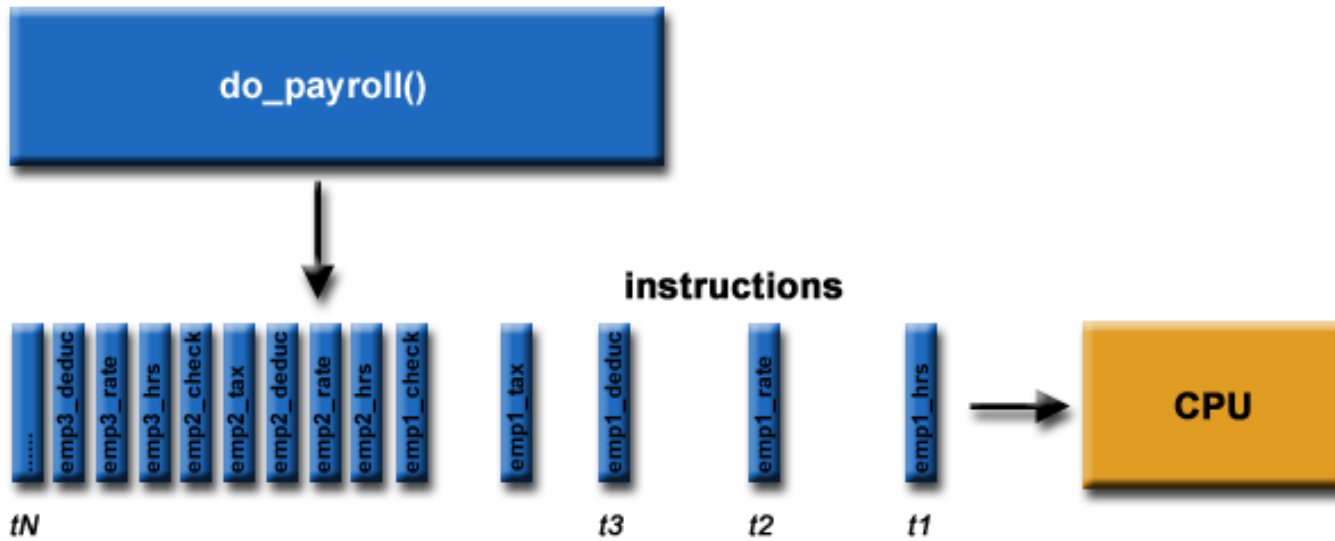
- **Költség csökkentés**
- **Gyors eredmények**



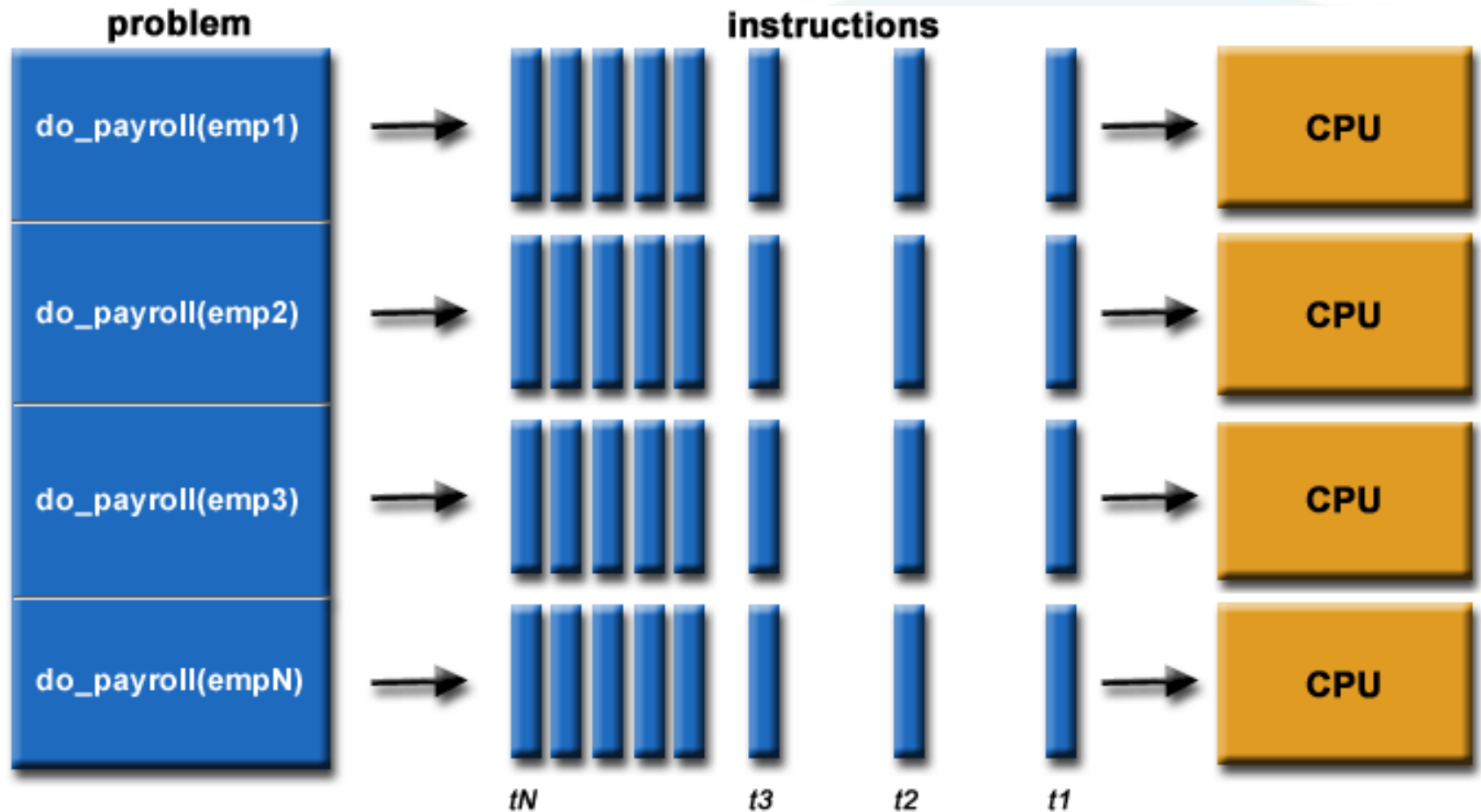
Miért fontos, hogy gyorsak legyünk?

- Elsők szeretnénk lenni
- Új gyógyszerek felfedezése (életek múlhatnak rajta)

Soros (serial) feldolgozás



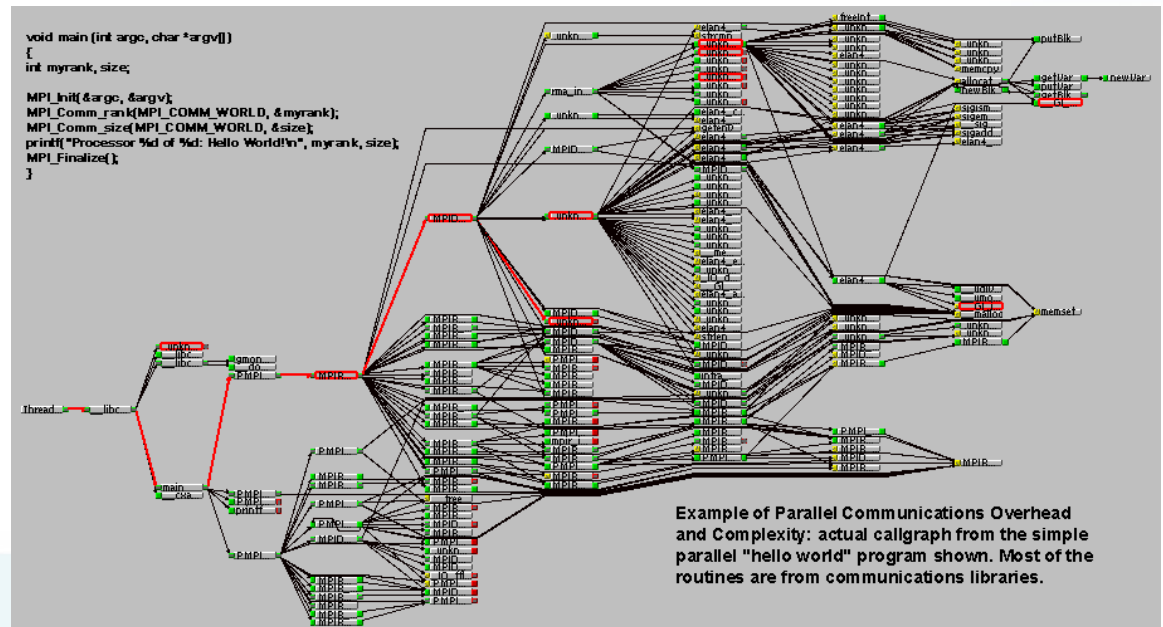
Párhuzamos (parallel) feldolgozás



Alapfogalmak (5/1)

Párhuzamosítási „költség” (Parallel Overhead)

- Feladat elindítása
- Szinkronizáció
- Kommunikáció
- Szoftver „költség” (párhuzamosítási nyelv, matematikai könyvtár, stb.)
- Feladat befejezése



Alapfogalmak (5/2)

Komplexitás



Skálázódás

Több erőforrás, nagyobb sebesség

Elsősorban ezek befolyásolják :

- Hálózat (sávszélesség, késleltetés)
- Algoritmus
- Memória (sávszélesség, összesen mennyi memória / node)
- CPU órajel
- Matematikai könyvtár
- Fordító

Alapfogalmak (5/4)

Gyorsulás (speedup)

- Maximális gyorsulás (Amdahl törvény)

$$S(N) = \frac{1}{(1-P)+P/N}$$

$$S(8) = \frac{1}{(1-0)+0/8} = 1$$

$$S(8) = \frac{1}{(1-1)+1/8} = 8$$

$$S(8) = \frac{1}{(1-0,9)+0,9/8} = 4,706$$

P: párhuzamosítható kód (0-1), valós szám

S: nem párhuzamosítható kód (0-1), valós szám

N: processzor magok száma (1-), egész szám

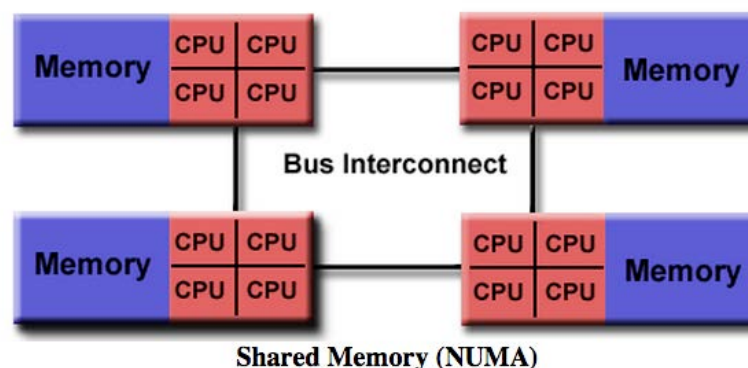
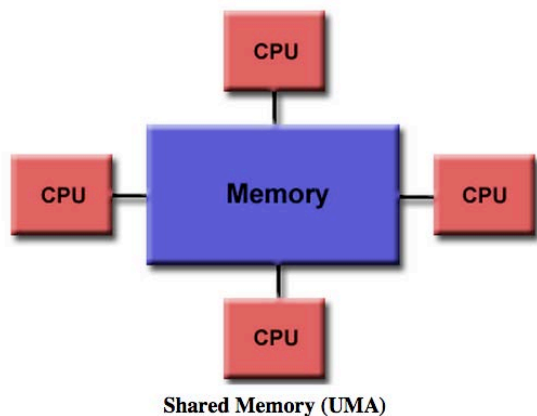
Portolhatóság

- Párhuzamos API-k (szabványok): MPI, POSIX threads, OpenMP, stb.
- Architektúra
- Operációs rendszer

Párhuzamosítási megoldások

Közös memóriás (Shared memory)

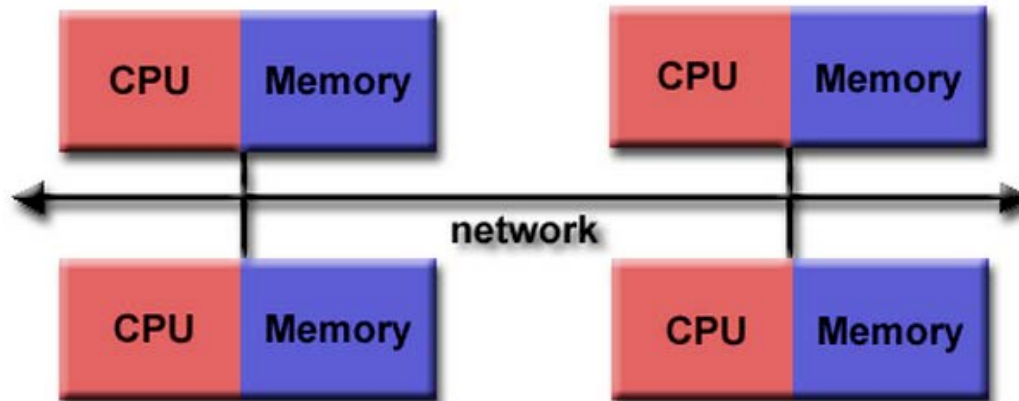
- POSIX threads
 - <https://computing.llnl.gov/tutorials/pthreads>
- OpenMP
 - <https://computing.llnl.gov/tutorials/openMP>



Párhuzamosítási megoldások

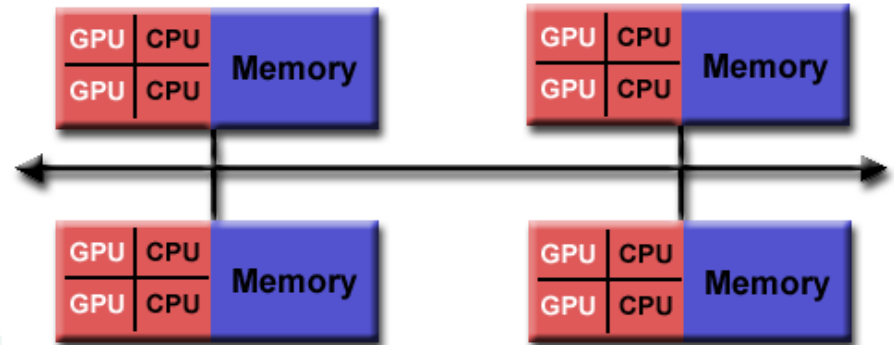
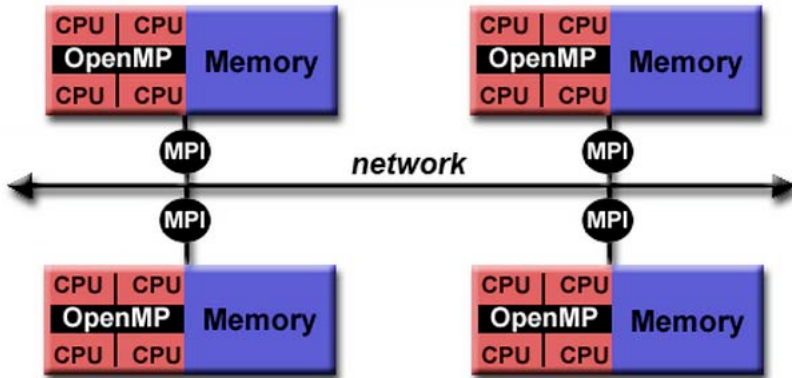
Elosztott memóriás (Distributed memory)

- MPI 1994
- MPI-2 1996
- MPI-3 2012
- <https://computing.llnl.gov/tutorials/mpi>



Párhuzamosítási megoldások

Hybrid model



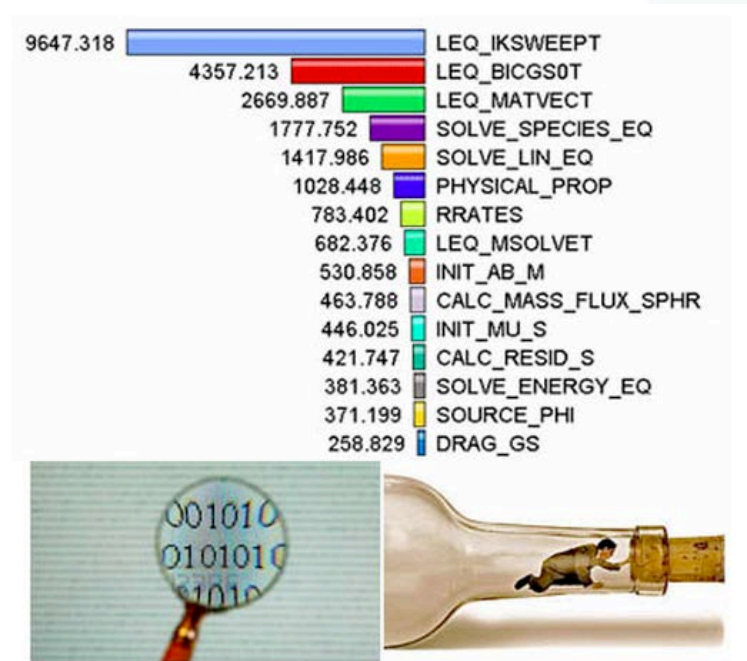
Automatikus

- A fordító analizálja a forráskódot
 - Nem biztos hogy jó lesz
 - Elsősorban ciklusok

Párhuzamosítás

Manuális

- Meg kell érteni azt a problémát amit meg szeretnénk oldani
- Meg kell találni a fontosabb részeket
 - Mi lassú?!
 - Miért lassú?!
 - I/O probléma?!
- Profiler programok segítenek
 - Képesek vagyunk szűrni

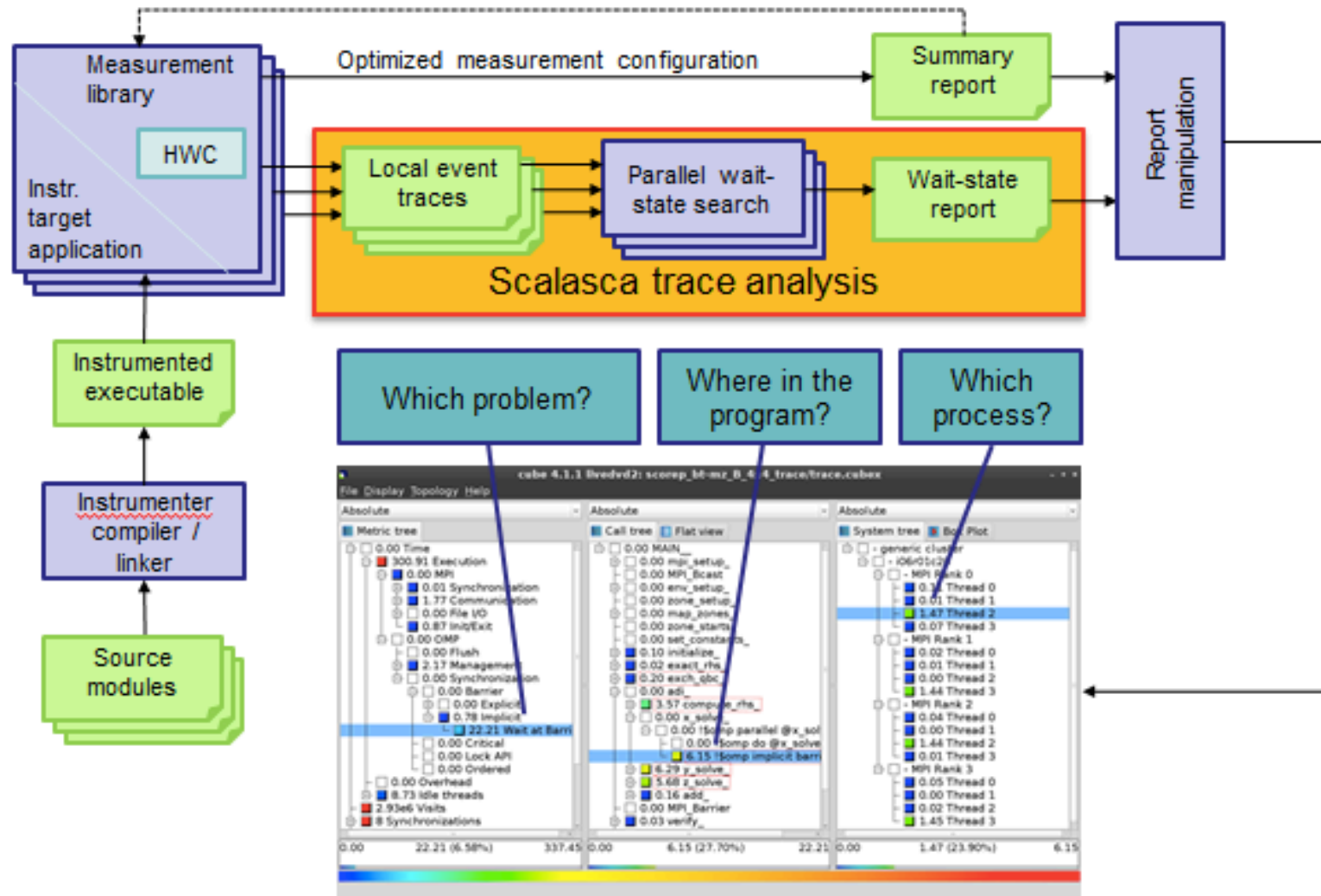


Scalasca profiler

- 1998 KOJAK projekt
- Jülich Supercomputing Centre
- German Research School for Simulation Sciences
- Futási adatok mérése, analizálása
- Több párhuzamosítás megoldást támogat
 - OpenMP, MPI, Hibrid
 - C, C++, Fortran
- Nyílt forráskódú, új BSD licenz
- Nagy rendszerekhez tervezték (több mint 1000 CPU core)
- Integrált mérés, instrumentation, analizálás, trace



Scalasca munkafolyamat



Scalasca használata (9/1)

Gyakorlati példa (Szegedi szuperszámítógépen)

```
SZEGED[loginnode] ~ (0)$ module load scalasca
```

```
/opt/nce/packages/global/scalasca/tutorial/NPB-MZ.tar.gz
```

http://www.training.prace-ri.eu/uploads/tx_pracetmo/BT-Tutorial-Scalasca.pdf

Scalasca használata (9/2)

```
SZEGED[loginnode] NPB3.3-MZ-MPI (1)$ diff -u config/make.def.orig
config/make.def
--- config/make.def.orig      2014-04-24 13:36:35.000000000 +0200
+++ config/make.def          2014-04-24 13:37:04.000000000 +0200
@@ -29,7 +29,7 @@
#-----
# This is the fortran compiler used for fortran programs
#-----
-F77 = mpif77
+F77 = scalasca -instrument mpif77
#F77 = mpiifort
# This links fortran programs; usually the same as ${F77}
FLINK  = $(F77)
SZEGED[loginnode] NPB3.3-MZ-MPI (1)$
```

Scalasca használata (9/3)

```
SZEGED[loginnode] NPB3.3-MZ-MPI (0)$ mkdir bin
SZEGED[loginnode] NPB3.3-MZ-MPI (0)$ make bt-mz CLASS=B NPROCS=4
=====
=   NAS PARALLEL BENCHMARKS 3.3   =
=   MPI+OpenMP Multi-Zone Versions =
=   F77                             =
=====

cd BT-MZ; make CLASS=B NPROCS=4 VERSION=
make[1]: Entering directory `/fs01/home/roczei/scalasca/NPB3.3-MZ-MPI/BT-MZ'
make[2]: Entering directory `/fs01/home/roczei/scalasca/NPB3.3-MZ-MPI/sys'
cc -o setparams setparams.c -lm
make[2]: Leaving directory `/fs01/home/roczei/scalasca/NPB3.3-MZ-MPI/sys'
../sys/setparams bt-mz 4 B
make[2]: Entering directory `/fs01/home/roczei/scalasca/NPB3.3-MZ-MPI/BT-MZ'
scalasca -instrument mpif77 -c -O3 -fPIC -fopenmp bt.f

...

scalasca -instrument mpif77 -c -O3 -fPIC -fopenmp verify.f
scalasca -instrument mpif77 -c -O3 -fPIC -fopenmp mpi_setup.f
cd ../common; scalasca -instrument mpif77 -c -O3 -fPIC -fopenmp print_results.f
cd ../common; scalasca -instrument mpif77 -c -O3 -fPIC -fopenmp timers.f
scalasca -instrument mpif77 -O3 -fopenmp -o ../bin/bt-mz.B.4 bt.o initialize.o exact_solution.o
exact_rhs.o set_constants.o adi.o rhs.o zone_setup.o x_solve.o y_solve.o exch_qbc.o solve_subs.o
z_solve.o add.o error.o verify.o mpi_setup.o ../common/print_results.o ../common/timers.o
INFO: Instrumented executable for OMP+MPI measurement
make[2]: Leaving directory `/fs01/home/roczei/scalasca/NPB3.3-MZ-MPI/BT-MZ'
make[1]: Leaving directory `/fs01/home/roczei/scalasca/NPB3.3-MZ-MPI/BT-MZ'
SZEGED[loginnode] NPB3.3-MZ-MPI (0)$
```

Scalasca használata (9/4)

```
SZEGED[loginnode] bin (0)$ cat job.sh
#!/bin/bash
#$ -pe mpi 16
#$ -q test.q

module load scalasca

OMP_NUM_THREADS=4 scalasca -analyze mpirun -np 4 ./bt-mz.B.4
SZEGED[loginnode] bin (0)$ qsub job.sh
Your job 722071 ("job.sh") has been submitted
SZEGED[loginnode] bin (0)$
```

Scalasca használata (9/5)

```
SZEGED@loginnode] bin (0)$ head -n 15 job.sh.o722071
S=C=A=N: Scalasca 1.4.3 runtime summarization
S=C=A=N: ./epik_bt-mz_4x4_sum experiment archive
S=C=A=N: Thu Apr 24 14:43:58 2014: Collect start
/opt/nce/packages/global/openmpi/1.6.3-gcc-4.7.2/bin/mpirun -np 4
/home/roczei/scalasca/NPB3.3-MZ-MPI/bin/bt-mz.B.4
[00000.0]EPIK: Created new measurement archive ./epik_bt-
mz_4x4_sum
[00000.0]EPIK: Activated ./epik_bt-mz_4x4_sum [NO TRACE] (0.225s)
[00000.0]EPIK: MPI-2.1 initialized 4 ranks
```

NAS Parallel Benchmarks (NPB3.3-MZ-MPI) - BT-MZ MPI+OpenMP
Benchmark

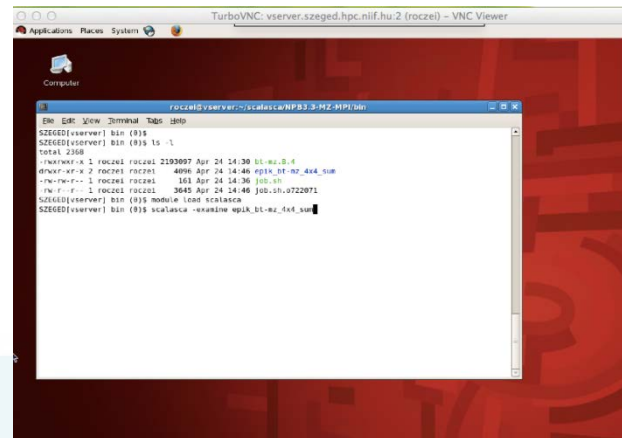
```
Number of zones:  8 x  8
Iterations: 200  dt:  0.000300
Number of active processes:  4
```

```
SZEGED@loginnode] bin (0)$
```

Scalasca használata (9/6)

TurboVNC használata: <http://www.niif.hu/node/674>

```
SZEGED[vserver] bin (0)$ ls -l
total 2368
-rwxrwxr-x 1 roczei roczei 2193097 Apr 24 14:30 bt-mz.B.4
drwxr-xr-x 2 roczei roczei 4096 Apr 24 14:46 epik_bt-mz_4x4_sum
-rw-rw-r-- 1 roczei roczei 161 Apr 24 14:36 job.sh
-rw-r--r-- 1 roczei roczei 3645 Apr 24 14:46 job.sh.o722071
SZEGED[vserver] bin (0)$ module load scalasca
SZEGED[vserver] bin (0)$ scalasca -examine epik_bt-mz_4x4_sum
```



The screenshot shows a TurboVNC window titled "TurboVNC: vserver.szeged.hpc.niif.hu:2 (roczei) - VNC Viewer". Inside the window, a terminal window titled "roczei@vserver:scalasca@P33-3-MZ-MPI/bin" displays the same terminal output as the text block above. The terminal output shows the execution of 'ls -l', 'module load scalasca', and 'scalasca -examine epik_bt-mz_4x4_sum' commands.

Scalasca használata (9/8)

The image shows a terminal window and the Cube 3.4 QT GUI. The terminal window displays the following commands and output:

```
roczei@vserver:~/scalasca/NPB3.3-MZ-MPI/bin
File Edit View Terminal Tabs Help
SZEGED[vserver] bin (0)$
SZEGED[vserver] bin (0)$ scalasca -examine epik_bt-mz_4x4_sum/
INFO: Displaying ./epik_bt-mz_4x4_sum/summary.cube.gz...
```

The Cube 3.4 QT GUI displays three main views:

- Metric tree:** Shows a hierarchical view of performance metrics. The 'Point-to-point' metric is highlighted with a value of 102.21. A context menu is open over this item, showing options like 'Info', 'Online description', 'Expand/collapse', etc.
- Call tree:** Shows the call stack of the program. The 'main' function is highlighted with a value of 0.00.
- System tree:** Shows the system topology, including the Linux Cluster and the processes (Process 0, Process 1, Process 2, Process 3) and threads (Thread 0, Thread 1, Thread 2, Thread 3).

At the bottom of the GUI, there is a table showing the values for the selected items:

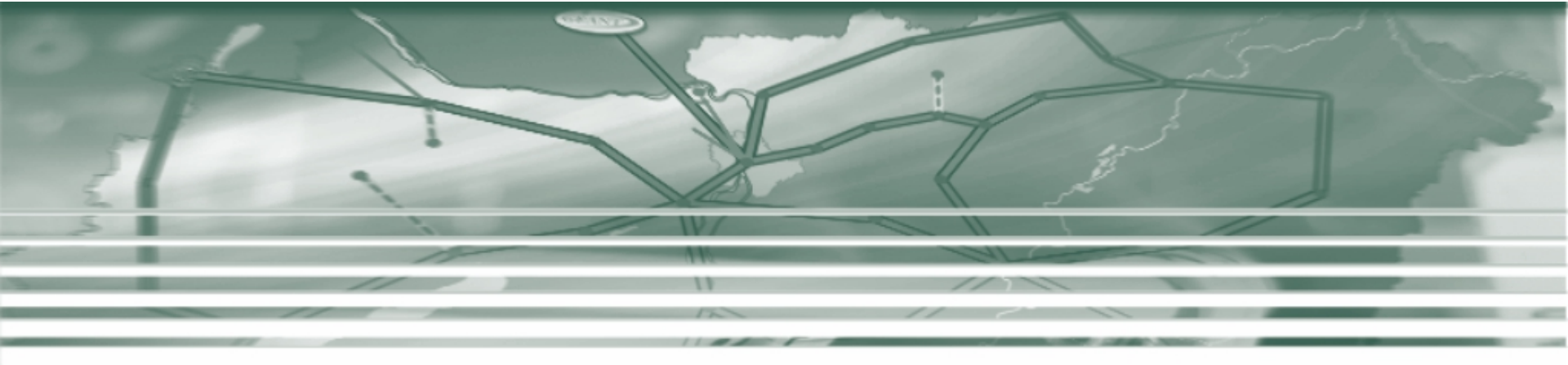
Metric tree	Call tree	System tree
0.00	0.00	0.00
102.21 (4.23%)	0.00 (0.00%)	0.00
2415.71	0.00	0.00

Shows the online description of the clicked item

Scalasca használata (9/9)

The screenshot displays the Scalasca performance analysis tool interface. In the background, a terminal window shows the execution of Scalasca on a program: `SZEGED[vserver] bin (8)$ scalasca -examine`. The main window, titled "Performance properties", is focused on the "Late Sender, Wrong Order Time / Different Sources" metric. It includes a description: "This specialization of the Late Sender, Wrong Order pattern refers to wrong order situations due to messages received from different source locations." Below the text is a timeline diagram with three processes (0, 1, 2) on the y-axis. Process 0 has a "Send" event. Process 1 has a "Send" event that occurs before Process 2's first "Recv" event. Process 2 has two "Recv" events. A red double-headed arrow indicates the time interval between the end of Process 1's "Send" and the start of Process 2's first "Recv". The "Unit" is "Seconds" and the "Diagnosis" suggests checking the proportion of "Point-to-point Receive Communications" that are "Late Sender, Wrong Order Instances (Communications)".

Köszönöm a figyelmet!



Róczei Gábor
roczei@niif.hu