Distribuované výpočty a GRID: prostředky a zkušenosti

Martin Petřek, Petr Kulhánek

Jan Kmuníček

Obsah

1. Distribuované výpočty
   - výpočetní chemie
2. Gridové systémy a software pro řazení úloh
   - METACentrum
     - PBS (fronty, vlastnosti, zadávání úloh, paralelní úlohy)
   - EGEE
     - LCG-2 (jazyk JDL, zadávání úloh)
3. Systém CHARON
   - požadavky a koncepce systému
   - Module - správa aplikačního software
   - ukázka
   - konfigurace

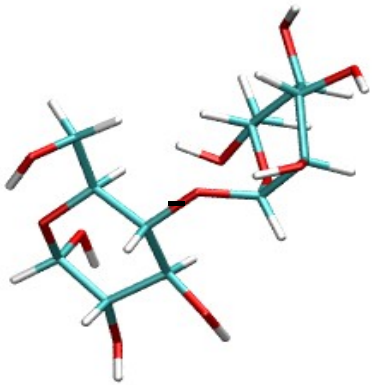
Distribuované výpočty

Chemické aplikace

- studie chování a simulace biologických systémů
- návrh léčiv – (studium interakcí protein x lék)
- molekulové dokování, konformační analýza molekuly
- zkoumání reakčních mechanizmů, určení tranzitních stavů, odhady energetických rozdílů pro reakční cestu, výpočty Gibsovy 'volné energie'

Matematické modely

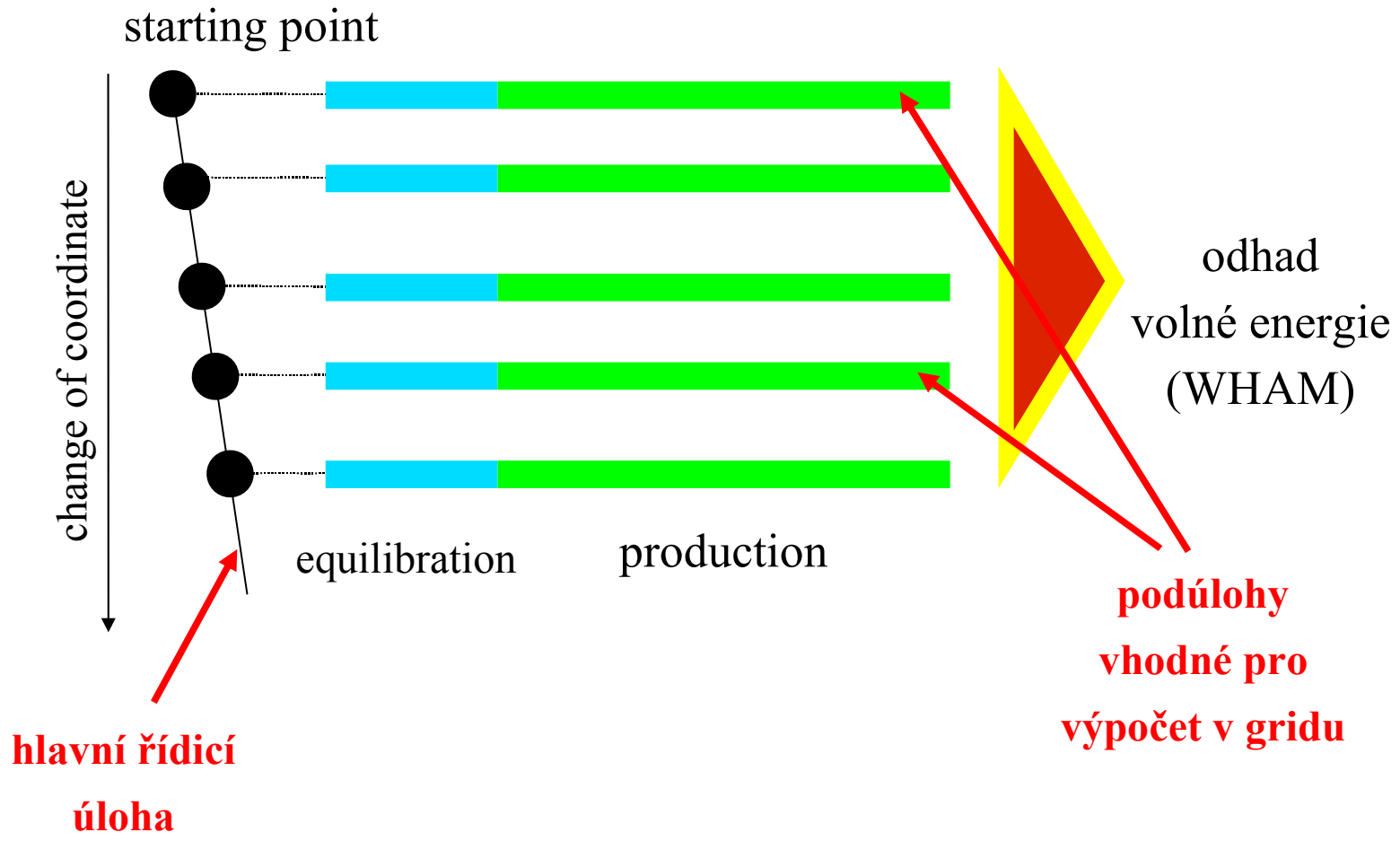
- QM, MM, CPMD



Distribuované výpočty

Výpočet volné energie metodou UMBRELLA

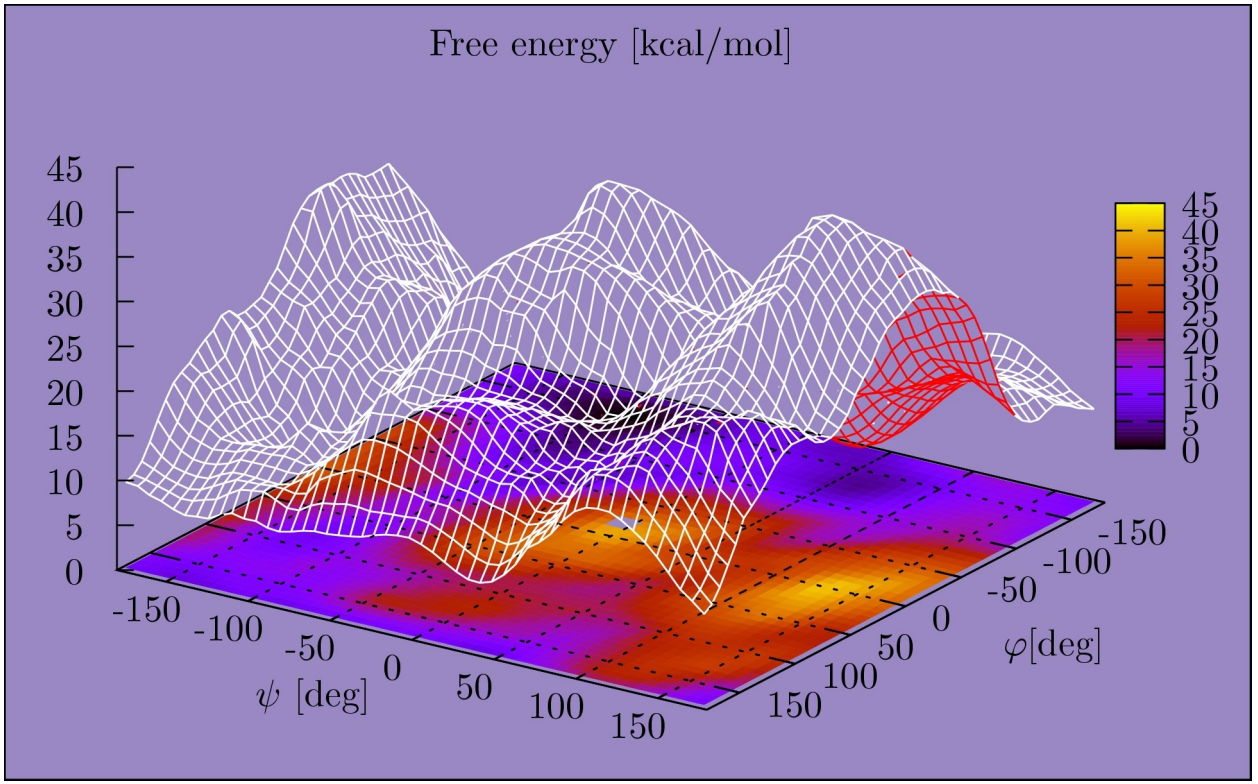
- isomaltose



Distribuované výpočty

Výpočet volné energie metodou UMBRELLA

- isomaltose



Distribuované výpočty

Výpočet numerického hessianu energie (QM)

- 3N*2 nezávislých výpočtů gradientu energie podle souřadnic

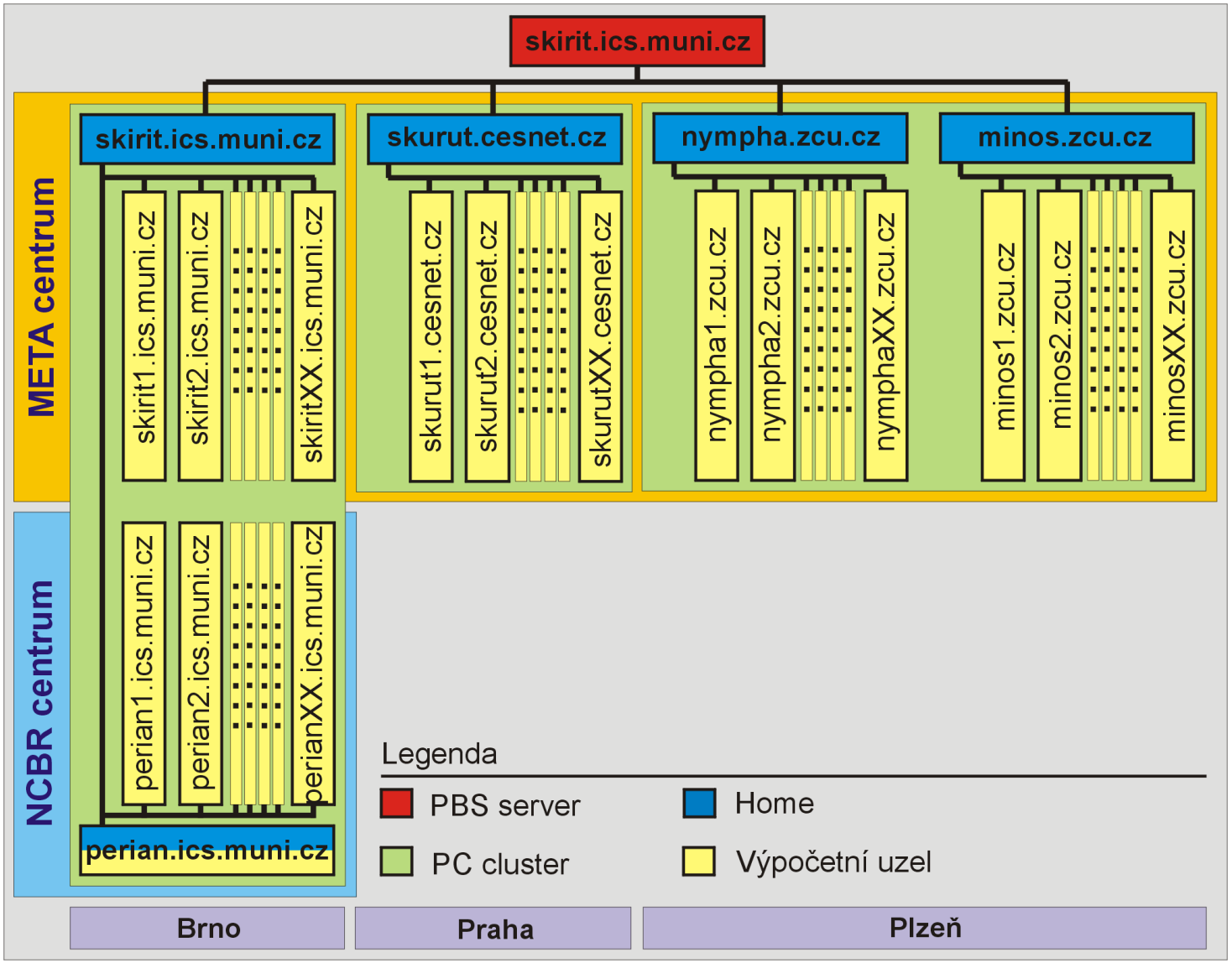
$$\begin{pmatrix}
\frac{\partial^2 E}{\partial x_1 \partial x_1} & \frac{\partial^2 E}{\partial x_1 \partial y_1} & \frac{\partial^2 E}{\partial x_1 \partial z_1} & \cdots & \frac{\partial^2 E}{\partial x_1 \partial x_n} & \frac{\partial^2 E}{\partial x_1 \partial y_n} & \frac{\partial^2 E}{\partial x_1 \partial z_n} \\
\frac{\partial^2 E}{\partial y_1 \partial x_1} & \frac{\partial^2 E}{\partial y_1 \partial y_1} & & \cdots & & \frac{\partial^2 E}{\partial y_1 \partial y_n} & \frac{\partial^2 E}{\partial y_1 \partial z_n} \\
\frac{\partial^2 E}{\partial z_1 \partial x_1} & & & & & & \frac{\partial^2 E}{\partial z_1 \partial z_n} \\
\vdots & & & \ddots & & & \vdots \\
\frac{\partial^2 E}{\partial x_n \partial x_1} & & & & & & \frac{\partial^2 E}{\partial x_n \partial z_n} \\
\frac{\partial^2 E}{\partial y_n \partial x_1} & \frac{\partial^2 E}{\partial y_n \partial y_1} & & \cdots & & \frac{\partial^2 E}{\partial y_n \partial y_n} & \frac{\partial^2 E}{\partial y_n \partial z_n} \\
\frac{\partial^2 E}{\partial z_n \partial x_1} & \frac{\partial^2 E}{\partial z_n \partial y_1} & \frac{\partial^2 E}{\partial z_n \partial z_1} & \cdots & \frac{\partial^2 E}{\partial z_n \partial x_n} & \frac{\partial^2 E}{\partial z_n \partial y_n} & \frac{\partial^2 E}{\partial z_n \partial z_n}
\end{pmatrix}$$

- -> k výpočtu vibračních modů molekuly

Gridové systémy - METACENTRUM

- http://meta.cesnet.cz (informace o projektu)
- sdružení CESNET
- distribuovaný výpočetní systém
  - Superpočítačové centrum Brno MU (http://scb.ics.muni.cz/static)
  - Superpočítačové centrum UK (http://supercomp.cuni.cz)
  - Superpočítačové centrum ZČU (http://zsc.zcu.cz)
- techinfo
  - 213 uzlů, 413 CPU
  - SMP stroje (shared memory), klastry (1-2 procesorové PC)
  - 1Gb/s (GE, Gigabit Ethernet) nebo 2.5Gb/s (Myrinet)

Gridové systémy - METACENTRUM



Gridové systémy - METACENTRUM

- programové prostředky
  - distribuovaný souborový systém AFS
  - autentizační systém Kerberos (kinit, kauth, SSH protokol)
  - systém správy aplikačního software (meta)moduly
  - přístup na centrální uzel pomocí SSH
  - přístup pomocí hardwarových klíčů (Token s certifikátem)
- software pro řazení úloh (dávkové systémy)
  - PBSPro – Portable Batch System, dávkový systém pro PC klastr
  - NQE – Network Queuing Environment, dávkový systém pro eru.ics.muni.cz
  - LSF – Load Share Facilities, dávkový systém pro grond.ics.muni.cz a gandalf.ics.muni.cz

PBS – Portable Batch System

(dávkový systém pro PC klastr)

METACENTRUM

- **Fronty**

Jméno fronty	Max. doba běhu	Maximum úloh	Maximum/Uživatel
• short	2 hodiny	12	8
• normal	24 hodin	24	12
• long	720 hodin	96	32
• ncbr	720 hodin	120	32
• cpmd	720 hodin	120	16

- **Vlastnosti výpočetních uzlů**

Vlastnosti (meta):
- linux
- praha
- brno
- plzen
- iti

Vlastnosti (ncbr):
- lcc
- ibp
- cpmd

Vlastnosti (obecné):
- p3
- xeon
- athlon

PBS – Portable Batch System

(dávkový systém pro PC klastr)

METACENTRUM

- Příkazy
  - **Zadávání úloh:** qsub
  - **Vymazání úlohy z fronty:** qdel
  - **Informace o ulohách:** qstat
  - **Informace o uzlech:** pbsnodes, xpbs

- Zařazení úlohy

```
[petrek@skirit petrek]$ qsub -r n -m abe -j oe -o test.out \
                        -e test.err -N "Test cislo 1" \
                        -q normal -l "node=1:brno:xeon" \
                        -v "BACKUPDIR" test
```

standardní a chybový výstup

fronta a vlastnosti

proměnné prostředí

skript

Gridové systémy -

- http://egee.cesnet.cz (informace o projektu)
- mezinárodní projekt Evropské Unie (CESNET za ČR)
- celoevropská gridová infrastruktura pro vědeckou komunitu i průmysl
  - 27 zemí, 70 organizací
- pilotní aplikace
  - HEP (High Energy Physics) – zpracování a analýza dat z experimentů částicové fyziky (Atlas, CMS, Alice, LHCb, ...)
  - biomedicínské gridy
  - výpočetně-chemické simulace biologických systémů
  - zpracování bioinformatikých a lékařských dat

Gridové systémy -

- přístup a autentifikace
  - virtuální organizace
  - hesla, certifikáty, SSH
- softwarové prostředky
  - LCG-2, EDG, Genius
  - gLITE
- EDG, LCG-2
  - OOP - (ComputingElements, UserInterface, StorageElements, ResourceBroker, InformationService, Monitoring and Discovering Service)
  - lcg_utils – API pro operace s úlohami

LCG-2

(low-level API pro EGEE)

- **Příkazy**
  - **Zadávání úloh:** edg-job-submit
  - **Informace o úloze:** edg-job-status
  - **Operace s daty na StorageElement:** lcg-cp, lcg-cr, lcg-del

- **Postup pro zařazení úlohy do fronty – na úrovni API**
  1) **Sestavení popisovacího skriptu pro úlohu** (*.JDL)
  2) **Nahrátí vstupních dat na storage element** (lcg-cp)
  3) **Vlastní zařazení úlohy** (edg-job-submit)
  4) **Sledování stavu úlohy** (edg-job-status)
  5) **Stáhnutí výsledku ze storage elementu** (lcg-cr)

- **Samotná úloha (běžící na ComputingElementu) musí zajistit načtení a uložení dat z/do Storage Elementu**

LCG-2

(low-level API pro EGEE)

- JobDescriptionLanguage

```
# JDL Test.jdl
Type = "Job";
JobType = "Normal";
Executable = "Test";
StdOutput = "Test.stdout";
StdError = "Test.stderr";
InputSandbox = {"in1.xml","in2.xml"};
OutputSandbox = {"out1.xml",”out2.xml”};
Environment = {
 "AMBERPATH=/var/amber",
 "BIGFILE1=guid:645c2af0-498e-4657-8154-8295380b349e"
 };
Arguments = "";
RetryCount = 1;
```

předává se s spolu s úlohou

identifikátor souboru na SE

```
$ export  VOCONFIG=edg_wl_ui.conf
$ edg-job-submit --config-vo $VOCONFIG -o JID test.jdl
```


DON'T PANIC

Přímé použití low-level API

- PBS
  - nutná znalost front, vlastností
  - uživatel musí znát poměrně dost informací o systému
  - kopírování vstupních dat na výpočetní uzel a stažení výsledku provádí až samostatný skript
  - paralelní úlohy - speciální volby ve spouštěcím skriptu ohledně architektury (shmem, p4, mpich-gm)
  - nastavení cest k software – uživatel musí opět znát, co je kde nainstalováno, jakou architekturu použít
  - => různé skripty pro různé architektury
  - informace o úloze svázané s identifikačním číslem jobu

Přímé použití low-level API

- LCG
  - nutná znal
  - ukládání a … soubory
  - chybí centr… – uživatel musí opět znát, co je … ogramy na SE spolu s úlohou
  - nelze se lo… …vat průběh výpočtu
  - paralelní spouštění úloh *(ve vývoji)*

Tohle obyčejného ! uživatele nezajímá !

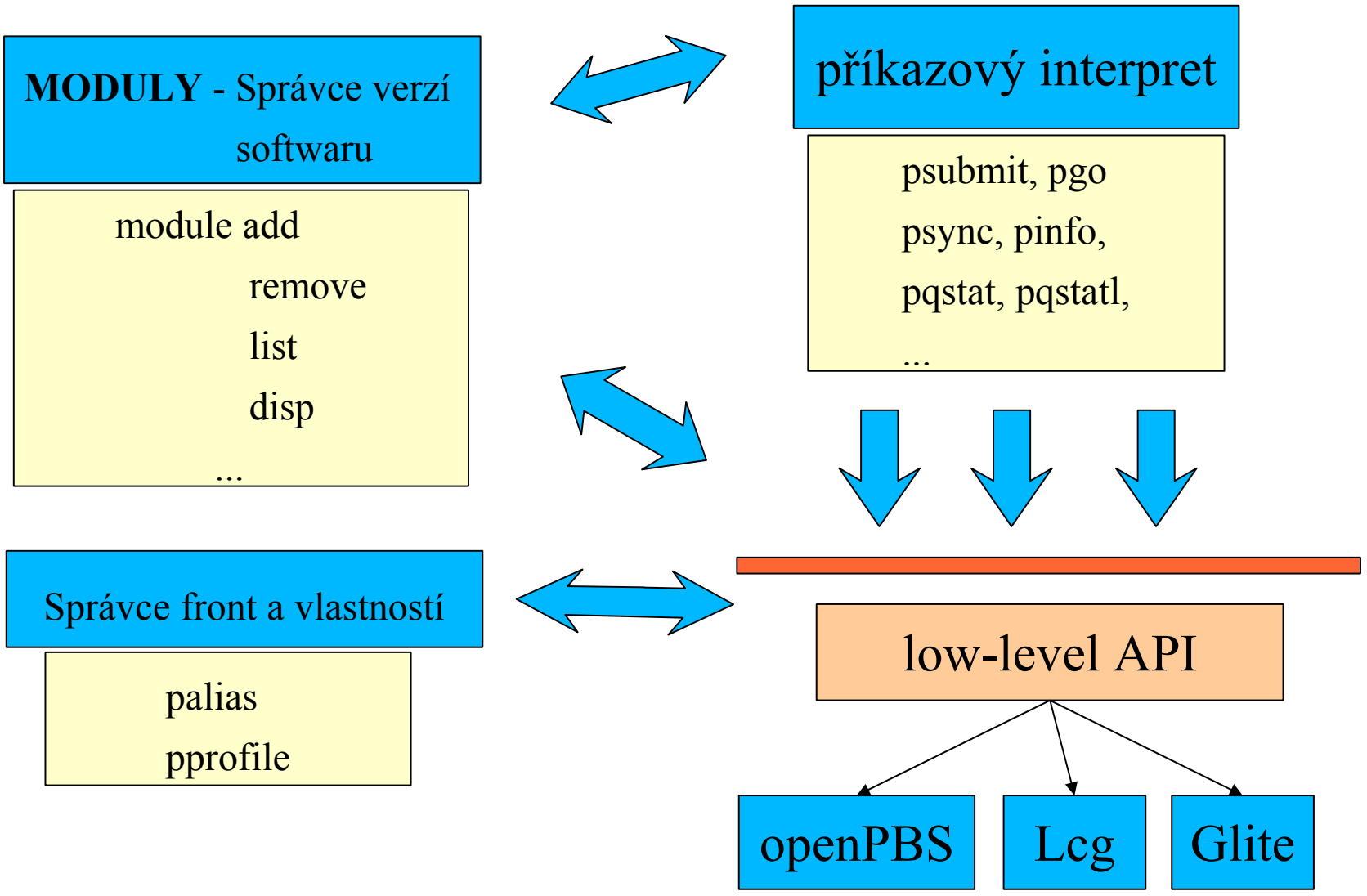
- Uživatelské požadavky
  - specifikace vstupních souborů programu
  - odeslaní programu do fronty
  - (nastavení vlastností, kde má úloha běžet a na kolika procesorech)

CHARON – koncepce systému

(Petr Kulhánek, Martin Petřek)

- nadstavba nad (Open)PBS(Pro), LCG, Glite, atp.
- jednotné prostředí v dávkových systémech
- Cíl: maximální zjednodušení práce
  - spouštění úloh
  - sledování průběhu výpočtu
  - správa softwaru a výběr verze pro dané architektury
  - spouštění paralelních úloh
  - spouštění na uzlech s požadovanými vlastnostmi
  - snadná instalace

CHARON – pohled do systému



CHARON ukázka

1) Příprava úlohy

```
[jobdir]$ ls
equi.rst isomaltose.top myjob prep.in
```

```
module add amber
sander -O -i prep.in \
       -o prep.out -p isomaltose.top \
       -c equi.rst -x prep.traj -r prep.rst
```

2a) Odeslání úlohy

```
[jobdir]$ psubmit normal myjob
```

```
[jobdir]$ psubmit voce myjob
```

METACENTRUM

Enabling Grids for E-sciencE

CHARON ukázka

2b) Odeslání úlohy

```
Job name      : myjob
Grid job name : myjob (Job type: generic)
Job directory : skurut4.cesnet.cz:/home/petrek/jobdir
Job project   : -none-
================================================================
Alias         : -none-
Organization  : voce
Profile       : default
----------------------------------------
NCPU          : 1
Resources     : -job match-
Properties    : -none-
Sync mode     : gridcopy
----------------------------------------
Start after   : -not defined-
================================================================
Do you want to submit job to GRID environment (YES/NO) ? YES

Please wait packing data ...
Submiting job ...
```

CHARON ukázka

3) Informace o úloze

```
[jobdir]$ pinfo
Job name    : myjob
JOb ID      : https://skurut3.cesnet.cz:9000/bx06C-RuqZawpCPQ
Grid job name : myjob (Job type: generic)
Job directory : skurut4.cesnet.cz:/home/petrek/jobdir
Job project   : -none-
==================================================
Alias         : -none-
Organization  : voce
Profile       : default
----------------------------------
NCPU          : 1
Resources     : -job match-
Properties    : -none-
Sync mode     : gridcopy
----------------------------------
Start after   : -not defined-
==================================================
Job was submited at     : 2005-10-12 14:16:28
  and was queued for  :  0d 00:04:28
Job was started at      : 2005-10-12 14:20:56
```

CHARON ukázka

3) synchronizace (v případě EGEE)

```
[jobdir]$ psync
Starting synchronization procedure.
  downloading sandbox ...
  completing data ...
  dowloading data from SE ...
  unpacking result archive ...
  cleaning ...
```

4) Výsledek

```
[jobdir]$ ls
equi.rst        myjob.ces       myjob.stdout
prep.in         myjob.cesout    mdinfo
isomaltose.top  myjob.jdl       prep.traj
myjob           myjob.info      prep.rst  prep.out
```

vstupní soubory

kontrolní soubory

výstupní soubory

CHARON správa softwaru

Hierarchie verzí a realizací softwaru

balík:verze:platforma:architektura

realizace

např.

`octave:2.1.71:i686:single`

`amber:8.0:xeon:single`

- **systém modulů volí nejvhodnější realizaci, pokud neuspěje, postupuje podle hierarchie**

xeon -> pn3 -> i686 -> i386 -> noarch

athlon#debian -> athlon -> pn3#debian -> pn3 -> i386#debian -> i386 -> noarch#debian -> noarch

- **paralelní realizace: p4, shmem, single, para**

CHARON správa softwaru

Konfigurace modulů

- **XML, jednoduché přidání nového softwaru**

Dokumentace – verze, build, patche

- **XML -> xslt() -> formát Wikipedia -> HTML**

Další vývoj

- ***implementace CHARONa pro gLITE (EGEE)***
- ***PHP rozhraní (web aplikace na odesílání jobů)***

Martin Petřek
Petr Kulhánek
Jan Kmuníček