MetaCentrum

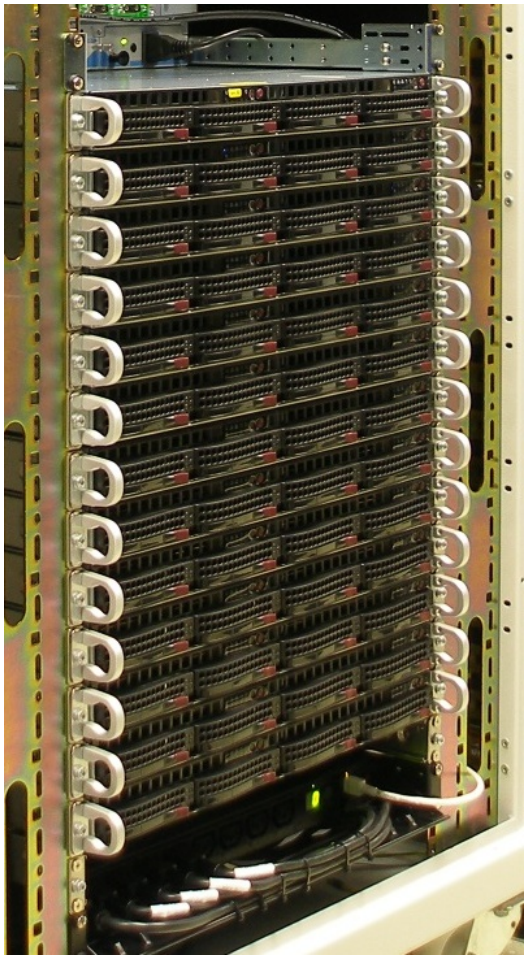
Výběr zdrojů, zadávání a správa úloh v MetaCentru

Martin Kuba

CESNET

O superpočítání

- MetaCentrum jsou spojená výpočetní a superpočítačová centra MU, UK, ZČU, PřF JU, FEKT VUT
- místo superpočítačů se sdílenou pamětí se používají clustery
- clustery PC-kompatibilních rack-mounted strojů (1U,2U,4U) s OS typu UNIX (Linux,IRIX,SunOS)
- klimatizované sály, zálohování napájení UPS/motorgenerátory, redundantní zdroje, hotswap, disková pole, zálohování dat na pásky, 1Gb/s Ethernet, 20Gb/s Infiniband, 10Gb/s spojení na páteř CESNETu



O rychlosti a výkonu

- Když jeden dělník vymaluje místnost za 1 hodinu, za jak dlouho to udělá 360 dělníků?
- sekání lánu obilí versus oprava hodinek
- clustery a superpočítače nabízejí více dělníků, ale je umění jim rozdělit práci
- Amdahlův zákon říká, o kolik lze maximálně zrychlit práci paralelním zpracováním
- lineární nárůst je nedostižný ideál

Plánovací systém

- zdroje
  - procesory (počet, typ)
  - místo na lokálním disku
  - stroje s určitými vlastnostmi (síť,OS, město,...)
  - paměť
  - licence software
- uživatelé
  - účty na clusterech a strojích
  - účast ve skupinách (příslušnost k organizaci)
- efektivní přidělování zdrojů lze zajistit jen s plánovacím systémem
- přidělování zdrojů ve formě úloh (jobs)
  - interaktivní
  - dávkové

Úloha (Job)

- životní cyklus:
  - vytvořena (požadavek na určité zdroje)
  - čeká ve frontě
  - spuštěna (zdroje jsou přiděleny)
  - ukončena (zdroje jsou uvolněny)
- interaktivní úloha
  - uživatel aktivně pracuje, obvykle s nějakým grafickým rozhraním, rozhoduje o ukončení úlohy
- dávková úloha
  - uživatel se věnuje něčemu jinému, úloha běží sama bez dozoru, vhodné při velkém množství úloh

Fronty úloh

- priorita je řízena pomocí **front úloh** (queue)
- tři **běžné** fronty
  - short – úlohy do 2 hodin
  - normal – úlohy do 24 hodin
  - long – úlohy do 30 dnů
- **prioritní** fronty
  - ncbr, jcu, iti, zsc, lsd, quark – pro vlastníky strojů
  - privileged – pro uživatele s alespoň 3 publikacemi s poděkováním
- **označující**
  - maintenance – stroje v údržbě nebo opravě
  - reserved – vyhrazeny
- **speciální** pro administrátorské účely a pokusné
- pro účely tohoto kurzu je speciální fronta qckurs@arien

Fronty

10.11.2009 12:36:53

O METACentru
Vybavení
Aktuality
Dokumentace
Přihláška
Můj účet
Stav zdrojů
Osobní pohled
Fyzické stroje
Virtuální stroje
Fronty úloh
Úlohy
Čekající úlohy
Uživatelé
Vlastnosti strojů
Mapa portálu
Diskuzní fórum
Interní část

Server skirit-f.ics.muni.cz - Produkční prostředí

fronta		priorita	časové limity	nutná vlastnost	úloh				
					ve frontě	běžících/max	hotových	celkem	max. na uživatele
maintenance	🔒	99	0 - 0		0	0 / 1000	0	0	1000
preempt_ncbr	🔒	95	0 - 720:00:00		0	0 / 3	0	0	2
reserved	🔒	90	0 - 0		0	0 / 1000	0	0	1000
globus	🔒	80	0 - 720:00:00	globus	0	0 / 4	0	0	4
pa177	🔒	80	0 - 24:00:00	pa177	0	0 / 4	0	0	15
priority	🔒	80	0 - 720:00:00		0	0 / 16	2	2	16
xentest	🔒	80	0 - 720:00:00		0	0 / 4	0	0	4
cpmd	🔒	71	0 - 0	per2	0	0 / 80	0	0	50
orca16g	🔒	71	0 - 720:00:00	orca16g	0	0 / 120	0	0	80
ncbr	🔒	70	0 - 720:00:00	per	19	27 / 1200	9	55	32
interactive	🔒	70	0 - 04:00:00	q_normal	0	0 / 6	0	0	6
jcu	🔒	70	0 - 720:00:00	jcu	0	0 /	0	0	
zsc	🔒	70	0 - 720:00:00	zsc	0	0 /	0	0	
orca	🔒	70	0 - 720:00:00	orca	0	7 / 120	0	7	80
iti	🔒	70	0 - 720:00:00	iti	0	0 / 500	10	10	48
privileged	🔒	65	0 - 720:00:00	forprivileged	2	3 / 24	3	8	16
short		63	0 - 02:00:00	q_short	4738	6 / 12	218	4962	8
long		62	24:00:01 - 720:00:00	long	22	24 /	6	52	400
preemptible		62	00:00:00 - 720:00:00	q_preemptible	2	41 /	17	60	400
normal		50	02:00:01 - 24:00:00	q_normal	33	207 / 2000	123	364	1000

fronta	popis	omezení
maintenance	Speciální fronta označující stroje určené k údržbě	pro uživatele: mulac ruda salvet xhavlik1
preempt_ncbr	Fronta umožňující na strojích NCBR okamžitě pozastavit úlohy v doménách -1 a spustit akutní úlohu v doménách -2. Používat vyjímečně.	pro uživatele: kulhanek makub salvet
reserved	Speciální fronta označující stroje dočasně vyhrazené ke speciálním účelům	pro uživatele: mulac salvet wiesner xhavlik1
globus	Speciální fronta pro úlohy zadávané přes systém Globus	pro uživatele: ruda
pa177	Fronta pro studenty předmětu PA177 na Masarykově Univerzitě	pro skupiny: scb pa177
priority	Speciální fronta pro administrátorské zásahy	pro uživatele: izaak knourek mulac ruda salvet
xentest	Pokusná fronta pro druhé virtuální stroje	pro uživatele: mulac ruda salvet
cpmd	Fronta pro uživatele z NCBR používající program CPMD	pro skupiny: scb ncbr
orca16g	Fronta pro úlohy vyžadující 16GB paměti na strojích clusteru Orca	pro uživatele: armladek judit kulhanek otyepka salvet
ncbr	Fronta pro uživatele z NCBR - National Center for Biomolecular Research (PřF MU)	pro skupiny: scb ncbr
interactive	Pokusná fronta pro interaktivní paralelní úlohy s možností preempce	pro uživatele: ruda
jcu	Fronta pro uživatele z JČU - Jihočeské Univerzity	pro skupiny: scb jcu
zsc	Fronta pro uživatele ZSC - Západočeského Superpočítačového Centra	pro skupiny: scb zsc
orca	Fronta pro úlohy na strojích clusteru Orca. Spouští svoje úlohy okamžitě na úkor úloh z fronty preemptible, které pozastavuje.	pro uživatele: armladek judit kulhanek otyepka salvet
iti	Fronta pro uživatele z ITI - Institutu Teoretické Informatiky (ZČU)	pro skupiny: scb iti kky kma
privileged	Fronta pro uživatele, kteří mají alespoň 3 publikace s poděkováním MetaCentru	pro uživatele: jhouska jkoca jsebera kosovan kraus manik msob pelikan stepan uhlik vrbka zeleny
short	Běžná fronta pro úlohy trvající maximálně 2 hodiny	
long	Běžná fronta pro úlohy trvající až 30 dnů	
preemptible	Fronta umožňující běžným uživatelům využít i stroje, na kterých jejich vlastníci mají možnost pozastavovat cizí běžící úlohy a spouštět svoje úlohy okamžitě. Úlohy mohou být zdrženy až o 30 dnů.	
normal	Běžná fronta pro úlohy trvající maximálně 24 hodin	
quark	Fronta pro uživatele clusteru Quark	pro skupiny: scb quark
mikroskop	Fronta pro zpracování obrazů z mikroskopu na clusteru Quark	pro uživatele: jfeit xhejtman
default	Směrovací fronta rozdělující úlohy podle délky běhu do front short a normal	
maintenance@arien	Speciální fronta označující stroje určené k údržbě	pro uživatele: knourek mulac ruda salvet xxhavlik1

Produkční a testovací prostředí

- dilema stabilita versus inovace
- řešením jsou dvě prostředí
  - **produkční** – stabilita – PBS server skirit-f
  - **testovací** – inovace – PBS server arien
- testovací prostředí
  - nový hardware
  - novější verze OS Linux
  - vylepšení prostředí

Požadavek na zdroje

- úlohy se zadávají příkazem qsub
  - třeba `qsub -q qckurs -l mem=3gb -I`
- parametry
  - fronta `-q normal`
  - počet strojů a CPU `-l nodes=1:ppn=4`
  - velikost paměti `-l mem=2gb`
  - vlastnosti strojů (viz dále)
  - interaktivní úloha `-I`
- dokumentace na http://meta.cesnet.cz/
- sestavovač příkazu qsub na portálu

Osobní pohled

10.11.2009 13:24:26

O METACentru
Vybavení
Aktuality
Dokumentace
Přihláška
Můj účet
Stav zdrojů
Osobní pohled
Fyzické stroje
Virtuální stroje
Fronty úloh
Úlohy
Čekající úlohy
Uživatelé
Vlastnosti strojů
Mapa portálu
Diskuzní fórum
Interní část
Hledat
RSS
CESNET

Tato stránka zobrazuje osobní pohled na PBS pro uživatele **makub**, tj. fronty a stroje uživateli přístupné.

Úlohy uživatele makub

uživatel	Počet úloh					Počet CPU úloh				
	celkem	ve frontě	běžících	dokončených	ostatních	celkem	ve frontě	běžících	dokončených	ostatních
makub	0	0	0	0	0	0	0	0	0	0

Sestavovač příkazu qsub

qsub -q normal -l mem= 400 mb -l nodes= 40 :ppn= 2 :x86 :linux
uloha.sh

Význam vlastností viz **Vlastnosti strojů**.

Nalézt stroje odpovídající požadavku

Výsledek

Výběr: qsub -q normal -l mem=400mb -l nodes=40:ppn=2:x86:linux

OK

Požadováno bylo 40 strojů, a právě teď je volných 43 strojů z 162 strojů splňujících požadavky, úloha bude spuštěna okamžitě.

Stroje volné právě teď

hermes11-1 (4 CPU, 15200mb (14gb))	**loslab1-1** (4 CPU, 3498mb)	**loslab2-1** (4 CPU, 3417mb)	**loslab3-1** (4 CPU, 3417mb)	**loslab4-1** (4 CPU, 3417mb)
loslab5-1 (4 CPU, 3417mb)	**loslab6-1** (4 CPU, 3417mb)	**manwe5** (3 CPU, 61534mb (60gb))	**manwe6** (4 CPU, 62134mb (60gb))	**nympha1-1** (7 CPU, 14800mb (14gb))

Osobní pohled

10.11.2009 13:25:02

O METACentru
Vybavení
Aktuality
Dokumentace
Přihláška
Můj účet
Stav zdrojů
Osobní pohled
Fyzické stroje
Virtuální stroje
Fronty úloh
Úlohy
Čekající úlohy
Uživatelé
Vlastnosti strojů
Mapa portálu
Diskuzní fórum
Interní část
Hledat
RSS
CESNET

Tato stránka zobrazuje osobní pohled na PBS pro uživatele **makub**, tj. fronty a stroje uživateli přístupné.

Úlohy uživatele makub

uživatel	Počet úloh					Počet CPU úloh				
	celkem	ve frontě	běžících	dokončených	ostatních	celkem	ve frontě	běžících	dokončených	ostatních
makub	0	0	0	0	0	0	0	0	0	0

Sestavovač příkazu qsub

qsub -q normal -l mem= 400 mb -l nodes= 44 :ppn= 2 :x86 :linux
uloha.sh

Význam vlastností viz **Vlastnosti strojů**.

Nalézt stroje odpovídající požadavku

Výsledek

Výběr: qsub -q normal -l mem=400mb -l nodes=44:ppn=2:x86:linux

Upozornění

Požadováno bylo 44 strojů, ale právě teď je volných jen 43 strojů, úloha bude čekat ve frontě.

Všechny stroje odpovídající požadavku

hermes01-1	hermes02-1	hermes03-1	hermes05-1	hermes06-1
hermes07-1	hermes09-1	hermes10-1	hermes11-1 (4 CPU, 15200mb (14gb))	konos1
konos2 (2 CPU, 1953mb)	konos3 (2 CPU, 1953mb)	konos4 (2 CPU, 1953mb)	konos5 (2 CPU, 976mb)	konos6 (2 CPU, 976mb)
konos7 (2 CPU, 976mb)	konos8 (2 CPU, 976mb)	konos9 (2 CPU, 976mb)	konos10 (2 CPU, 976mb)	konos13-1 (1 CPU, 2900mb)

Osobní pohled

10.11.2009 13:34:58

O METACentru
Vybavení
Aktuality
Dokumentace
Přihláška
Můj účet
Stav zdrojů
Osobní pohled
Fyzické stroje
Virtuální stroje
Fronty úloh
Úlohy
Čekající úlohy
Uživatelé
Vlastnosti strojů
Mapa portálu
Diskuzní fórum
Interní část

Hledat

RSS

CESNET

Tato stránka zobrazuje osobní pohled na PBS pro uživatele **kmunicek**, tj. fronty a stroje uživateli přístupné.

Úlohy uživatele kmunicek

uživatel	Počet úloh					Počet CPU úloh				
	celkem	ve frontě	běžících	dokončených	ostatních	celkem	ve frontě	běžících	dokončených	ostatních
kmunicek	0	0	0	0	0	0	0	0	0	0

Sestavovač příkazu qsub

qsub -q normal@arien -l mem= 400 mb -l nodes= 1 :ppn= 32 :x86 :linux
uloha.sh

Význam vlastností viz **Vlastnosti strojů**.

Nalézt stroje odpovídající požadavku

Výsledek

Výběr: qsub -q normal@arien -l mem=400mb -l nodes=1:ppn=32:x86:linux

Problém

Požadován byl 1 stroj, ale požadavku neodpovídá žádný stroj, na kterém máte účet, úloha nebude nikdy spuštěna !

Můžete-li, zřiďte si účty na následujících strojích:
eru1 eru2

Fronty dostupné uživateli kmunicek

Vlastnosti strojů

- **geografické umístění** – vliv na latenci sítě
  - brno, plzen, praha, budejovice, feec
- **typ a počet CPU**
  - i386, amd64, emt64t, x86_64, x86, ia64, xeon, opteron
  - nodecpus1, nodecpus2, ..., nodecpus32
- **operační systém**
  - linux, debian, debian40, debian50, sarge, lenny, suse, redhat
- **síťové karty**
  - myrinet2000, infiniband, infiniband2
- **pro frontu** – není nutné zadávat
  - orca, q_normal, q_short, q_preemptible, forprivileged, ncbr, globus, jcu
- **ostatní**
  - nfs4, per, per2 až per5, ibp, lcc, data-kky, loslab

i386	hydra1	hydra2	hydra3	hydra4	hydra5	hydra6	hydra7	hydra8
	hydra9	hydra10	perian18	perian19	perian20	perian21	perian22	perian23
	perian24	perian25	perian26	perian28	perian29	perian30	perian31	perian32
	perian33	perian34	perian36	perian37	perian38	perian39	perian40	perian41
	perian43	perian46	perian47	perian48	perian49	perian51	perian52	perian53
	perian54	perian55	perian56	perian57	perian58	perian59	perian61	perian62
	perian63	perian64	perian65	perian66	perian67	perian68	quark1	quark2
	quark3	quark4	quark5	skirit18	skirit19	skirit20	skirit21	skirit22
	skirit23	skirit24	skirit25	skirit26	skirit27	skirit28	skirit29	skirit30
	skirit31	skirit32	skirit33	skirit34	skirit35	skirit36	skirit37	skirit38
	skirit39	skirit40	skirit41	skirit42	skirit44	skirit45	skirit46	skirit47
	skirit48	skurut33	skurut34	skurut35	skurut36	skurut37	skurut38	skurut39
	skurut40	skurut41	skurut42	skurut43	skurut44	skurut45	skurut46	skurut47
	skurut48	skurut49	skurut50	skurut51	skurut52	skurut53	skurut54	skurut55
	skurut56	skurut57	skurut58	skurut59	skurut60	skurut61	skurut62	skurut63
	skurut64	skurut65	skurut66	skurut67	skurut68			
ia64	acharon	ajax						
ibp	perian33	perian34	perian36	perian53	perian54	perian55	perian56	perian57
	perian58	perian59	perian61	perian62	perian63	perian64	perian65	perian66
	perian67	perian68	perian69	perian70	perian71	perian72	perian73	perian74
	perian75	perian76						
infiniband	skirit49-1	skirit50-1	skirit51-1	skirit52-1	skirit53-1	skirit54-1	skirit55-1	skirit56-1
	skirit57-1	skirit58-1	skirit59-1	skirit60-1	skirit61-1	skirit62-1	skirit63-1	skirit64-1
	skirit65-1	skirit66-1	skirit67-1	skirit68-1	skirit69-1	skirit70-1	skirit71-1	skirit72-1
	skirit73-1	skirit74-1	skirit75-1	skirit76-1	skirit77-1	skirit79-1	skirit80-1	skirit81-1
	skirit82-1	skirit83-1						
infiniband2	alela1-1	alela2-1	alela3-1	alela4-1	alela5-1	alela6-1	alela7-1	alela8-1
iti	konos1	konos2	konos3	konos4	konos5	konos6	konos7	konos8
	konos9	konos10	konos11-1	konos11-2	konos13-1	konos13-2	konos14-2	konos14-1
	konos15-1	konos15-2	konos16-1	konos16-2	konos17-1	konos17-2	konos18-1	konos18-2
	konos19-1	konos19-2	konos20-1	konos20-2	konos21-1	konos21-2	konos22-2	konos22-1
	konos23-2	konos23-1	konos24-1	konos24-2	konos25-2	konos25-1	konos26-1	konos26-2
	konos27-2	konos27-1	konos28-2	konos28-1	konos29-1	konos29-2	konos30-1	konos30-2
	konos31-1	konos31-2	konos32-2	konos32-1	konos33-1	konos33-2	konos34-1	konos34-2
	konos35-2	konos35-1	konos36-1	konos36-2	konos37-1	konos37-2		

Činnost plánovače

- plánovač setřídí úlohy podle
  - priority fronty
  - fair share
  - CPU čas (pokud není uveden, je nastaven podle fronty)
- fair share – uživatel s méně propočítaným časem má přednost
- má informace o tom, kde má uživatel účet a do kterých patří skupin

Práce s daty

- na každém stroji jsou dostupné 4 druhy diskového prostoru
  - rychlý nesdílený lokální disk ve /scratch
  - adresář sdílený NFSv3 přes jeden cluster v /home
  - adresář sdílený NFSv4 přes všechny stroje v /storage
  - adresář sdílený AFS přes všechny stroje v /afs
- předinstalovaný software je na AFS
- data uživatelů na NFSv3 a NFSv4
- data v době výpočtu na lokálním disku

Přenos dat

- do/z MetaCentra pomocí WinScp, scp, sftp, NFSv4
- uložit do /storage na NFSv4
- na začátku úlohy přenést na lokální disk
  - nejlépe lokálním kopírováním „cp -r /storage/home/$USER/data /scratch/$USER/”
  - nebo pomocí stagein/stageout, to zajistí i smazání dat po skončení úlohy

Časté chyby uživatelů

- žádost o konkrétní stroj
- nesplnitelná kombinace
  - nejsou vůbec takové stroje (např. ia64:brno)
  - jsou takové stroje, ale nemají na nich účet
- přímé přihlášení na stroj bez úlohy (Viktor Čistič)
- používají více CPU než rezervovali
- data pro výpočet jsou v $HOME, ale požadavek splňují i stroje mimo cluster
- nerezervují si licenci pro software
- pro dočasná data nepoužívají scratch
- pracují na frontendu místo použití jednouzlové interaktivní úlohy

Tento kurs

- vyhražená fronta qckurs@arien
- přihlásit se na frontend
  - `ssh QC01@skirit.ics.muni.cz`
- spustit interaktivní úlohu
  - `qsub -q qckurs@arien -l nodes=1:ppn=4 -l mem=3gb -I`
- dostanete shell na jednom ze strojů
- cd /scratch/$USER

Virtualizace a preempce

- všechny fyzické stroje s dostatečnou pamětí jsou virtualizovány
- obvykle dva virtuální stroje -1 a -2
- využito na clusterech orca a skirit č.49-81
- fronty **orca** a **preempt@arien** posílají úlohy do virt. strojů -2 a tím pozastavují -1

Doporučený postup

- zjistit přes *Sestavovač qsub* v *Osobním pohledu* na portálu, zda jsou dostupné vhodné zdroje
- požádat o účty na všech dostupných strojích
- použít aplikaci z modulů nebo mít vlastní na AFS
- zadat úlohu s co nejméně omezujícími požadavky
- po skončení úlohy jsou stdout a stderr nakopírovány do adresáře na stroji, odkud byla úloha zadána. Vhodné je použít adresář ve /storage kvůli místu a dostupnosti.
- velké objemy výstupů zapisovat do /scratch
- vstupní a výstupní data kopírovat přes staging do a z /scratch

Dotazy ?

- Děkuji za pozornost