Příloha č. 1 zadávací dokumentace

Dodávka komponent výpočetního clusteru národní gridové infrastruktury pro projekt Velká infrastruktura CESNET

Technická dokumentace, specifikace požadovaného plnění a popis hodnocení

Předmětem veřejné zakázky je dodávka, instalace a zprovoznění výpočetního clusteru složeného z SMP uzlů s alespoň čtyřmi procesory se sdílenou pamětí, včetně prostoru pro ukládání pracovních dat uživatelů a sdíleného diskového prostoru pro dočasné soubory. Současně zadavatel požaduje poskytnutí rozšířené záruky včetně technické podpory pro jednotlivé komponenty výpočetního clusteru - požadovaný rozsah těchto plnění je uveden v odst. 4.2.3 zadávací dokumentace.

Zadavatel požaduje kompletní řešení, sestávající se z totožných výpočetních uzlů, diskových prostor pro dočasná data (scratch filesystem), diskového prostoru pro semi-permanentní data (home filesystem), síťového propojení, hardwarového managementu clusteru, včetně racku a montáže, tříletou (36 měsíců) rozšířenou záruku včetně technické podpory ve formě next-business day, on site (viz odst. 4.2.2 a 4.2.3 zadávací dokumentace).

Zadavatel požaduje nabídky na výpočetní cluster s následujícími vlastnostmi:

- celkový počet alespoň **600 fyzických CPU jader** (bez hyperthreadovaných jader)
- celkový výkon (součet výkonů jednotlivých výpočetních uzlů) alespoň 9000 bodů SPECfp2006 rate baseline
- fileserver exportující NFSv4 pro sdílená pracovní data uživatelů (/home) s užitnou kapacitou 100 TB včetně odpovídající síťové infrastruktury.
- fileserver exportující vhodný souborový systém pro dočasná data uživatelů (/scratch) včetně odpovídající síťové infrastruktury; u tohoto systému nepředpokládáme zálohovaní či vysoké zabezpečení proti ztrátě dat, dále nepředpokládáme re-export jiným souborovým systémem mimo dodaný cluster.

V nabídce musí být explicitně uvedena cena jednoho výpočetního uzlu.

Požadavky zadavatele na jednotlivé části výpočetního clusteru

1. Každý výpočetní uzel musí splňovat tyto podmínky:

1.1. Provedení do standardního 19" racku nebo dle vlastního řešení. Rack musí být součástí dodávky. Na rozměry racku neklade zadavatel omezení.

1.2. V případě sdílení některých komponent více počítači (například při provedení blade) redundance komponent společných pro všechny počítače (zdroje apod.). Redundance komponent v jednotlivých počítačích není nutná, v případě HW chyby může dojít k výpadku jednoho počítače, ale nesmí dojít k výpadku více než 2 počítačů vlivem selhání jedné komponenty.

1.3. V případě provedení blade možnost vyměnit za chodu jednotlivé komponenty (servery, zdroje, switche apod.) blade chassis.

1.4. Každý počítač (výpočetní jednotka se samostatnou pamětí, chipsetem, procesory, diskem, atd.) musí být vybaven minimálně čtyřmi procesory se sdílenou pamětí.

Procesory musí být v architektuře x86_64. Minimální výkon celého uzlu měřený nástrojem SPECfp2006 ve variantě rate base musí být 579 bodů[1].

1.5. Operační paměť alespoň 64 GB ECC na jeden fyzický procesor, paměťové moduly musí být v kanálech rozmístěny rovnoměrně.

1.6. Každý počítač musí mít přístup k lokálnímu disku, na kterém bude nainstalován operační systém a odkládací prostor (swap). Velikost tohoto lokálního úložiště musí být minimálně 30 GB na systém + velikost RAM * 2. Tento požadavek platí i pro blade provedení.

1.7. Každý počítač musí mít přístup k síťovému paralelnímu souborovému systému pro dočasné soubory (scratch). Je-li tento souborový systém zpoplatněn, musí být licence součástí nabídky, váže-li se na nějakou zpoplatněnou variantu OS, musí být tato také součástí nabídky.

1.8. Rozhraní 1Gb Ethernet a InfiniBand QDR na každou započatou čtveřici fyzických procesorů, obě rozhraní musí být připojena do systému tak, aby jejich průchodnost nebyla omezena přípojným místem (např. dvouportová IB karta s rozhraním PCI-E x8 nestačí).

1.9. Každý počítač umožňuje centralizovaný přístup ke konzoli (klávesnice + monitor) a zároveň podporuje bootování z externího zařízení, a to jak lokálně (KVM switch, boot z USB – CD-ROM, flash disk, harddisk), tak po síti (síťový KVM nebo BMC, boot z virtuálního média).

1.10. Základní deska musí umožňovat změnu pořadí bootovacích zařízení.

1.11. Základní deska musí obsahovat management controller (BMC) kompatibilní se specifikací IPMI 2.0 nebo vyšší. BMC musí umět monitorovat minimálně funkčnost ventilátorů, teplotu CPU a základní desky; dále musí BMC poskytovat základní vzdálený power management (vypnout, zapnout, reset). Požadujeme možnost změny bootovacího zařízení vzdáleně pomocí BMC nebo KVM.

1.12. Funkcionalita IPMI musí být přístupná z příkazové řádky běžící na vzdáleném linuxovém systému připojeném k BMC přes LAN.

1.13. Uzly clusteru by mělo být možno koupit bez jakéhokoliv software. Pokud je programové vybavení nutnou součástí nabídky (například SW pro vzdálenou správu), musí být jasně specifikovány důvody a cena za takový SW musí být zahrnuta do ceny dodávky (na dobu neurčitou; pokud autor / výrobce / dodavatel SW neposkytuje licenci na dobu neurčitou, je uchazeč povinen tuto skutečnost zadavateli prokázat a zajistit licenci nejméně do konce roku 2015 – viz odst. 9.20 zadávací dokumentace).

2. Síťová infrastruktura

2.1. Odpovídající gigaethernetové a InfiniBandové switche, kabeláž apod. včetně připojení diskových prostor (viz dále). U InfiniBandu neblokující propustnost pro skupiny alespoň 24 portů, blocking faktor 1/3 nebo lepší pro celou instalaci.

2.2. Blocking faktor InfiniBandové infrastruktury definujeme jako minimum poměru celkové propustnosti mezi skupinami uzlů k součtu nominálních kapacit rozhraní uzlů v menší skupině přes všechny možné způsoby rozdělení clusteru na dvě skupiny uzlů. V obvyklé topologii skládající se z vnitřně neblokujících koncových přepínačů propojených neblokujícím jádrem sítě se blocking faktor rovná podílu počtu uplinků a počtu připojených koncových uzlů na koncovém přepínači (při stejných typech portů na obou stranách), např. při připojení 24 koncových uzlů na přepínač je tedy nezbytné mít v nabízené konfiguraci alespoň 8 uplinků do zbytku sítě.

[1] 80% současného nejlepšího čtyřprocesorového skóre, vyhovět mohou např. čtyřprocesorové systémy založené na CPU AMD Opteron 6176 (4x12core) nebo Intel Xeon E7-4850 (4x10core)

2.3. Uplink gigaethernetových switchů musí být minimálně 1x10 Gbps. Gigaethernetové switche musí umožnit připojení front-end diskových polí (viz níže) rychlostí 10 Gbps. Nabídnutou infrastrukturu musí být možně připojit do vnější sítě jedním uplinkem rychlostí 10 Gbps.

3. Diskové pole /home

3.1. Rackmount systém.

3.2. Součástí dodávky diskového pole jsou 2 front-endy, které diskové pole zpřístupní.

3.3. Jedno nebo více diskových polí připojených ke dvěma front-endům exportující NFSv4[2]. Export NFSv4 musí podporovat autentizaci systémem Kerberos. Oba front-endy musí exportovat stejná data a adresářové struktury. Pokud je potřeba k realizaci této vlastnosti další SW, musí být součástí nabídky. Váže-li se zmíněný SW na konkrétní placený operační systém (RHEL, SLES, Solaris, AIX, ...), musí být tento rovněž součástí nabídky. Zadání vyhovuje active-passive přístup (jeden z front-endu exportuje data a druhý je v té době v režimu stand-by). Active-active přístup je výhodou (souborový systém na diskovém poli umožňuje paralelní přístup 2 front-endů k jednomu svazku, tento souborový systém musí být v tomto případě součástí nabídky). Dva uzly jsou požadovány z důvodu odolnosti proti výpadku a rozložení zátěže. NFSv4 může reexportovat jiný souborový systém jako např. CXFS, GPFS, Lustre a jiné.

3.4. Celková kapacita musí být minimálně 100 TB. Do kapacity 100 TB nejsou počítány paritní disky. Zabezpečení disků musí být pomocí RAID 5 nebo RAID 6. Dále musí být dodány nejméně 2 hot spare disky, přidělitelné k libovolnému RAIDu. RAID musí být nakonfigurován tak, aby rebuild neběžel více jak 24 hodin (během plného provozu, je přípustná degradace výkonu). Uchazečem dodané výsledky výkonnostního měření musí být provedeny na uchazečem navržené konfiguraci vyhovující tomuto zadání (není tedy možné dodat výkonnostní charakteristiky pouze pro RAID 0 nebo pro jinou RAID konfiguraci nesplňující uvedené požadavky).

3.5. Pole a servery mohou být samostatné jednotky. Součástí nabídky musí být veškeré propojovací prvky jako např. FC kabely a switche.

3.6. Plná redundance diskových polí, včetně řadičů, zdrojů napájení, ventilátorů a případných FC switchů a FC řadičů (v diskových serverech i polích).

3.7. Front-end servery musí mít připojení rychlostí 10 Gbps (ethernet) a musí být touto rychlostí připojitelné do síťové infrastruktury nabízené spolu s worker nody. Dále musí být servery integrované do nabízené InfiniBand infrastruktury. Každý front-end musí mít alespoň 32 GB RAM. Každý front-end musí mít alespoň 8 fyzických CPU jader (nepočítáme hyperthreadovaná jádra). Dále musí být každý front-end vybaven dvěma systémovými disky s kapacitou alespoň 100 GB každý.

3.8. Zabezpečení cache hardwarových RAID řadičů při výpadku proudu nebo poruše jednoho z řadičů.

3.9. Disky a zdroje v serverech i polích typu hot-plug.

3.10. Vzdálený management a monitoring serverů i diskových polí, varování o poruchách disků a řadičů pomocí SNMP zpráv. Vzdálený management musí být plně použitelný z Linuxu.

3.11. Sestava musí poskytovat průchodnost alespoň 150 MB/s při sekvenčním čtení velkého souboru z jednoho uzlu a 150 MB/s při sekvenčním zápisu velkého

[2] Není požadována plná implementace protokolu NFSv4, za dostatečnou považujeme implementaci v linuxovém jádře verze 2.6.18 (RHEL), nepožadujeme podporu NFSv4.1.

souboru z jednoho uzlu (čtení a zápis nebude měřen paralelně, viz příkaz iozone níže).

3.12. Sestava musí poskytovat celkovou průchodnost alespoň 800 MB/s při sekvenčním čtení velkých souborů z 8 uzlů zároveň a 500 MB/s při sekvenčním zápisu velkých souborů z 8 uzlů zároveň (čtení a zápis nebude měřen paralelně, viz příkaz iozone níže). Průchodnost pro 8 uzlu a pro jeden uzel nebude měřena paralelně.

3.13. Oba požadavky na průchodnost musí být dosažitelné na identické dodané konfiguraci.

3.14. Ověření výkonu bude prováděněno pomocí iozone -t 1 -Mce -s200g -r256k -i0 -i1 -F „cesta k souboru v /home" pro bod 3.11, iozone -t 8 -Mce -s200g -r256k -i0 -i1 -+m pro bod 3.12. Podstatné pro průchodnost jsou údaje „Children see throughput for 1(8) initial writers" (pro zápis) a "Children see throughput for 1(8) readers" (pro čtení).

4. Síťový scratch

4.1. Rackmount systém.

4.2. Součástí dodávky je příslušný počet front-endů, které síťový scratch zpřístupní vhodným protokolem na klienty.

4.3. Velikost scratche musí být alespoň 450 GB * počet fyzických procesorů (socketů, nikoli jader) ve výpočetních uzlech clusteru, disky pro scratch musí mít rychlost otáčení alespoň 10000 RPM.

4.4. Konfigurace zabezpečení síťového scratche musí být taková, aby průměrná dostupnost scratche pro výpočetní uzly odpovídala nebo byla lepší než v konfiguraci s RAID 0 přes dva lokální disky v každém uzlu (jeden velký RAID 0 nebo obdobná konfigurace je nepřijatelné). Nepožadujeme konkrétní implementaci typu mirror nebo RAID 5.

4.5. Body 5. až 10. a 13. předchozí sekce platí i pro síťový scratch.

4.6. Komponenty síťového scratche mohou být sdílené s diskovým polem pro /home, je ale nutné aby byly dodrženy požadované výkonové parametry i při zátěži obou částí pole (specifikovaný výkon každé části musí být dosažitelný i současném testování obou souborových systémů, tj. bod 7 a 8 níže bude testován současně s body 3.11 a 3.12).

4.7. Sestava musí poskytovat průchodnost alespoň 300 MB/s při sekvenčním čtení velkého souboru z jednoho uzlu a 300 MB/s při sekvenčním zápisu velkého souboru z jednoho uzlu.

4.8. Sestava musí poskytovat celkovou průchodnost alespoň 1400 MB/s při sekvenčním čtení velkých souborů z 8 uzlů zároveň a 800 MB/s při sekvenčním zápisu velkých souborů z 8 uzlů zároveň.

4.9. Oba požadavky na průchodnost musí být dosažitelné na identické dodané konfiguraci.

4.10. Ověření výkonu bude prováděno pomocí iozone -t 1 -Mce -s200g -r256k *-i0 -i1* -F „cesta k souboru v /home" pro bod 4.7, iozone -t 8 -Mce -s200g -r256k *-i0 -i1* -+m pro bod 4.8. Podstatné pro průchodnost jsou údaje „Children see throughput for 1(8) initial writers" (pro zápis) a "Children see throughput for 1(8) readers" (pro čtení).

5. Ostatní

5.1. Všechny výpočetní uzly, které jsou touto technickou specifikací požadovány, musí být použitelné v prostředí operačního systému Linux (zejména, ale nikoliv výhradně Debian a openSuse), tj. musí být podporovány distribučním nebo originálním jádrem nebo s využitím externích ovladačů dostupných ve zdrojovém kódu. Front-endy diskového pole a síťového scratche musí být provozovány na free nebo komerční verzi Linuxu nebo UNIXu; komerční verze musí být součástí nabídky. Na front-endy musíme mít možnost plného administrátorského přístupu (root účet v Unixu, většina NAS appliance neposkytuje administrátorský přístup). Klienti souborového systému pro scratch musí být použitelní v prostředí operačního systému Linux ve formě zdrojových kódů (jejichž součástí může být binární objekt).

5.2. Všechny hardwarové komponenty musí být umístěny do dodaných racků chlazených vzduchem. Tepelný výkon všech komponent umístěných ve vzduchem chlazeném racku nesmí nikdy přesáhnout 15 kW. Počet racků, v nichž je celé řešení umístěno, není součástí hodnocení.

5.3. Každý počítač (vč. servisních strojů, front-endů) umožňuje centralizovaný přístup ke konzoli (klávesnice + monitor) a zároveň podporuje bootování z externího zařízení, a to jak lokálně (KVM switch, boot z USB – CD-ROM, flash disku, harddisku), tak po síti (síťový KVM nebo BMC, boot z virtuálního média). Ke KVM switchi požadujeme LCD monitor a klávesnici, je tedy nutno dodat fyzický KVM switch. Tyto komponenty nemusí mít nezbytně provedení optimalizované pro montáž do racku a úsporu prostoru, musí být ale při instalaci do racku vhodným způsobem zabudovány (nesmí vyžadovat dodatečný stůl apod.)

5.4. Součástí nabídky musí být celková maximální spotřeba sestavy (maximální spotřeba odpovídá spotřebě při plném zatížení všech komponent, tedy všech výpočetních uzlů, front-endů, diskových polí, síťových switchů).

6. Měření výkonu

Součástí nabídky budou výkonnostní testy dle následujícího popisu.

6.1. Zadavatel v akceptačních testech ověří deklarované výsledky měření (dle bodů 1.4, 3.11, 3.12, 4.7 a 4.8) na dodané sestavě nakonfigurované dle výše uvedené technické specifikace, tj. v konfiguraci headnodů podle bodu 3.7, v konfiguraci RAID dle bodu 3.4 resp. 4.4.

6.2. Testy dodané pro účely hodnocení nemusejí být pořízeny na stejném hardware, který bude dodán, případně v dodávané konfiguraci. Dodavatel nicméně odpovídá za to, že skutečně naměřené hodnoty během akceptačních testů na skutečně dodané konfiguraci nebudou horší, než jaké přikládá k nabídce. Nevadí, budou-li skutečně naměřené hodnoty lepší.

6.3. Pro rychlost úložiště home a scratch je pro zadavatele podstatná rychlost naměřená programem iozone (verze 3.347 z http://www.iozone.org). **Výstup programu iozone je nutné přiložit k nabídce.**

6.4. Rychlost úložiště bude měřena na jednom, resp. 8 fyzických klientech dodaných v konfiguraci dle sekce 1.

6.5. Pro home bude rychlost měřena nad protokolem NFSv4. Pro scratch bude rychlost měřena nad protokolem, který je součástí nabídky a je určen pro připojení scratch úložiště.

7. Hodnocení nabídek

Celkové hodnocení nabídek bude zohledňovat celkovou výši nabídkové ceny v Kč bez DPH s váhou 90 % a celkovou maximální spotřebu elektrické energie v kW s váhou 10 %. Body v jednotlivých dílčích hodnotících kritériích budou přiděleny takto:

- (nejnižší cena ze všech nabídek / hodnocená nabídnutá cena) * 100 * 0,90
- (nejnižší spotřeba v kW ze všech nabídek / hodnocená udaná spotřeba) * 100 * 0,10

Celková hodnotící tabulka

Hodnocené kritérium	Hodnota	Váha kritéria v %	Počet bodů
Celková nabídková cena		90	
Celková spotřeba kW		10	
Celkem	**---**	**100**	