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 uzlů se dvěma procesory se sdílenou pamětí, včetně prostoru pro ukládání pracovních dat uživatelů a diskového prostoru pro dočasné soubory. Instalací a zprovozněním se rozumí instalaci HW do RACK skříní, zapojení do elektrické sítě a spuštění HW a ověření bezchybného chodu všech HW komponent, instalace SW vybavení a operačního systému je požadována jen na fileserveru. Pokud bude dodavatel preferovat prověření výkonových požadavků na vlastní instalaci linuxových strojů, musí být součástí instalace a zprovoznění také instalace příslušného počtu klientských stanic. 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, diskového prostoru pro semi-permanentní data (home filesystem), 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)
- fileserver exportující NFSv4 pro sdílená pracovní data uživatelů (/home) s užitnou kapacitou 100 TB
- výběrové řízení je vypsáno na výpočetní uzly, které budou dodány do dvou oddělených lokalit. Cluster s racky, diskovým prostorem a výpočetními uzly bude umístěn na Přírodovědecké fakultě Jihočeské univerzity v Českých Budějovicích, několik samostatných výpočetních uzlů nahradí staré výpočetní uzly clusteru, který je umístěn na CESNETu v Praze.

V nabídce musí být explicitně uvedena cena a spotřeba 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 vysokého minimálně 42U. 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 dvěma procesory se sdílenou pamětí. Procesory musí být v architektuře x86_64. Minimální výkon celého uzlu měřený nástrojem SPECfp2006 ve variantě rate base musí být 450 bodů. Zároveň výkon v tomto benchmarku přepočtený na jedno jádro CPU, tj. výkon celého uzlu vydělený počtem fyzických jader v uzlu, dosahuje alespoň 28. Počítají se pouze fyzická jádra, nikoli technologie hyperthreading. Zájemce uvede v nabídce deklarované hodnoty, které jeho řešení dosahuje, tyto hodnoty budou ověřeny v akceptačních testech. Zadavatel preferuje CPU se spotřebou nižší než 120W/CPU.

1.5. Operační paměť alespoň 32 GB na jeden fyzický procesor, paměťové moduly musí být v kanálech rozmístěny rovnoměrně, všechny musí být stejné velikosti a typu ECC DDR3-1333 nebo lepší.

1.6. Každý počítač musí mít přístup ke dvěma lokálním diskům s kapacitou alespoň 900 GB každý. Tento požadavek platí i pro blade provedení. Jsou přípustné pouze disky typů SAS, FC, SCSI nebo SATA s NCQ a rychlostí otáčení ploten 10k RPM.

1.7. Každý uzel musí mít rozhraní 1Gb Ethernet a InfiniBand 4xQDR.

1.8. 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).

1.9. Základní deska musí umožňovat změnu pořadí bootovacích zařízení.

1.10. 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.

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

1.12. 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. Diskové pole /home

2.1. Rackmount systém.

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

2.3. Jedno nebo více diskových polí připojených ke dvěma front-endům exportující NFSv4[1]. Export NFSv4 musí podporovat autentizaci systémem Kerberos. Front-endy budou nakonfigurovány v režimu active-passive, NFSv4 může exportovat např. XFS souborový systém.

2.4. Celková kapacita musí být minimálně 100 TB. Do kapacity 100 TB nejsou počítány paritní ani hot-spare disky. Zabezpečení disků musí být pomocí RAID 5 nebo RAID 6. Dále musí být dodány nejméně 4 hot spare disky, přidělitelné k libovolnému RAIDu. RAID musí být nakonfigurován tak, aby rebuild neběžel více jak 48 hodin (během plného provozu, je přípustná degradace výkonu). Uchazečem dodané výsledky výkonnostního měření musí být provedeny na uchazečem navržené

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

konfiguraci vyhovující tomuto zadání (není tedy možné dodat výkonnostní charakteristiky pouze pro RAID 0 nebo pro jinou RAID konfiguraci nesplňující uvedené požadavky).

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

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

2.7. Front-end servery musí mít připojení rychlostí 10 Gbps (ethernet) a 4xQDR InfiniBand. Každý front-end musí mít alespoň 96 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.

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

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

2.10. 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.

2.11. Sestava musí poskytovat průchodnost alespoň 250 MB/s při sekvenčním čtení velkého souboru z jednoho uzlu a 250 MB/s při sekvenčním zápisu velkého souboru z jednoho uzlu (čtení a zápis nebude měřen paralelně, viz příkaz iozone níže).

2.12. Sestava musí poskytovat celkovou průchodnost alespoň 1200 MB/s při sekvenčním čtení velkých souborů z 8 uzlů zároveň a 700 MB/s při sekvenčním zápisu velkých souborů z 8 uzlů zároveň (čtení a zápis nebude měřen paralelně, viz příkaz iozone níže). Průchodnost pro 8 uzlů a pro jeden uzel nebude měřena paralelně.

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

2.14. Ověření výkonu bude prováděněno pomocí iozone -t 1 -Mce –s400g -r256k -i0 -i1 -F „cesta k souboru v /home“ pro bod 2.11, iozone -t 8 -Mce –s400g -r256k -i0 -i1 -+m pro bod 2.12. 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). V případě ověření výkonu z bodu 2.11 bude test puštěn z právě jednoho uzlu. V případě ověření výkonu z bodu 2.12 bude příkaz puštěn z právě jednoho uzlu a díky volbě -+m budou automaticky spuštění klienti iozone na 8 uzlech.

3. Ostatní

3.1. 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 musí být provozovány na free nebo komerční verzi Linuxu nebo UNIXu; licence musí být součástí nabídky. Na front-endy musíme mít možnost plného administrátorského přístupu (root účet v Unixu, většina NAS appliance neposkytuje administrátorský přístup).

3.2. Všechny hardwarové komponenty musí být umístěny do maximálně dvou 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.

3.3. 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).

3.4. Součástí nabídky nejsou prvky síťové infrastruktury, switche, InifniBand switche a dále KVM switche.

3.5. Na celou sestavu je požadována záruka v délce 36 měsíců, on site, s reakcí NBD.

4. Měření výkonu

Součástí nabídky budou výkonnostní testy dle následujícího popisu.

4.1. Zadavatel v akceptačních testech ověří deklarované výsledky měření (dle bodů 1.4, 2.11, 2.12) na dodané sestavě nakonfigurované dle výše uvedené technické specifikace, tj. v konfiguraci headnodů podle bodu 2.7, v konfiguraci RAID dle bodu 2.4.

4.2. 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.

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

4.4. Rychlost úložiště bude měřena na jednom, resp. 8 fyzických klientech dodaných v konfiguraci dle sekce 1. Rychlost bude měřena nad protokolem NFSv4 z jednoho front-endu.

5. 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**	