Příloha č. 1

Základní technická specifikace

1. návrh technického řešení Malého clusteru

1.1. Návrh Mobilního datového centra

Společnost Bull s.r.o. předkládá komplexní řešení Mobilního kontejnerového datového centra, jenž poskytuje vhodné prostředí pro umístění a provoz ICT systémů, které budou do datového centra instalovány, tj. Systému pro náročné výpočty a Specializovaného systému.

Mobilní datové centrum zajišťuje zejména bez-výpadkové napájení, chlazení, rozvody elektrické energie, umístění IT zařízení, odolnost proti vnějším vlivům, zabezpečení, vzdálený dohled a správu technologií.

Řešení je navrženo, dimenzováno a bude implementováno tak, aby zajistilo spolehlivý, bezpečný a efektivní provoz Systému pro náročné výpočty a Specializovaného systému.

Řešení je komplexní a soběstačné bez potřeby dalších zadavatelem poskytovaných nebo zadavatelem zajišťovaných zařízení, systémů či infrastruktury.

Systémy Mobilního datového centra, které jsou klíčové pro zajištění provozního prostředí ICT, mají s výjimkou motorgenerátoru redundantní provedení.

Při návrhu řešení jsme vycházeli zejména z následujících parametrů:

- Pro umístění Mobilního datového centra Zadavatel vyhradil venkovní prostor obdélníkového charakteru o rozměrech 15x13 m (ploše cca 195 m^2). Plocha bude provedena ze štěrkodrti tl. 300 mm, zhutněná na 15MPa. Plocha bude oplocena pletivem do výšky 1800 mm, vybavena přípojkou elektro z vedlejší budovy. Zadavatel zajistí ochranu lokality formou dálkového dohledu v režimu 24/7 a fyzické ostrahy. Pro přístup na plochu bude Zadavatelem připraven přístupový chodník.
- Instalace jednotlivých kompletovaných částí mobilního datového centra na vyhrazenou plochu bude provedena pomocí jeřábu, a to z prostoru přilehlého parkoviště p. č. 1738/87 v k. ú. Poruba.
- Celkový maximální příkon ze vstupní elektrické sítě nesmí překročit 300 kW, přičemž musí být zajištěno rovnoměrné zatížení jednotlivých fází sítě
- V předmětné lokalitě je možno umístit motorgenerátor o maximálním jmenovitém výstupním výkonu 200kW.
- Mobilní datové centrum musí poskytovat bez výpadku napájení ICT systémů i v případě výpadku napájení z elektrické sítě po dobu až 6 hodin a to bez zásahu obsluhy. Po tuto dobu se může snížit výkon Systému pro náročné výpočty dle podmínek zadání.
- Mobilní datové centrum musí umožňovat provoz v rozsahu venkovních teplot -25°C až +43°C, pro teploty nad 35 °C je povolen určitý stupeň řízeného omezení výkonu ICT systémů
- Územní rozhodnutí obsahuje podmínku, že v rámci provozu kontejnerů s výpočetní technikou budou dodrženy limitní hodnoty hygienických limitů hluku u nejbližších chráněných objektů

dle §12 Nařízení vlády č. 272/2011 o ochraně zdraví před nepříznivými účinky hluku a vibrací, přičemž hluková studie zpracovaná pro účely posouzení vlivu hluku z provozu kontejnerů stanovila, že pro modelovou situaci 3 kontejnerů s výpočetní technikou s chlazením a kontejnerem s diesel agregátem maximální instalovaný akustický výkon na jednom kontejneru s výpočetní technikou nesmí přesáhnout $L_{WA,max} \leq 78$ dB. Pro kontejner s diesel agregátem byla uvažována hladina akustického tlaku 69 dB ve vzdálenosti 7m od zdroje s tím, že agregát bude v provozu při výpadku elektrické energie a jeho pravidelné provozní zkoušky se budou konat výhradně v denní době a budou trvat maximálně 1 hodinu za den.

- V lokalitě není stálý zdroj vody, chlazení je řešeno s uzavřeným okruhem chladicí kapaliny, chladící okruh obsahuje i zásobníky o celkovém objemu 3 000l.
- K dispozici bude optická přípojka sestávající se z 24 ks (12 párů) single mode vláken zakončených konektory SC/APC.
- Mobilní datové centrum bude umožňovat umístění a provoz dalšího systému tzv. „Specializovaného systému", který není předmětem této nabídky. Mobilní datové centrum poskytne veškeré technologie, infrastrukturu a zdroje potřebné pro zajištění souběžné funkcionality „Systému pro náročné výpočty" a „Specializovanému systému" (napájení, chlazení, racky, elektroinstalaci, atd.). Pro „Specializovaný systém" bude dostupný vyhrazený příkon ICT 70kW s tím, že v době výpadku napájení elektrické energie z veřejné sítě lze při přechodu na napájení ze záložního generátoru požadovat snížení vyhrazeného příkonu pro „Specializovaný systém" na 50 kW.

1.1.1. TECHNICKÉ A PROSTOROVÉ ŘEŠENÍ

Naše nabídka pro mobilní datové centrum obsahuje tyto základní části:

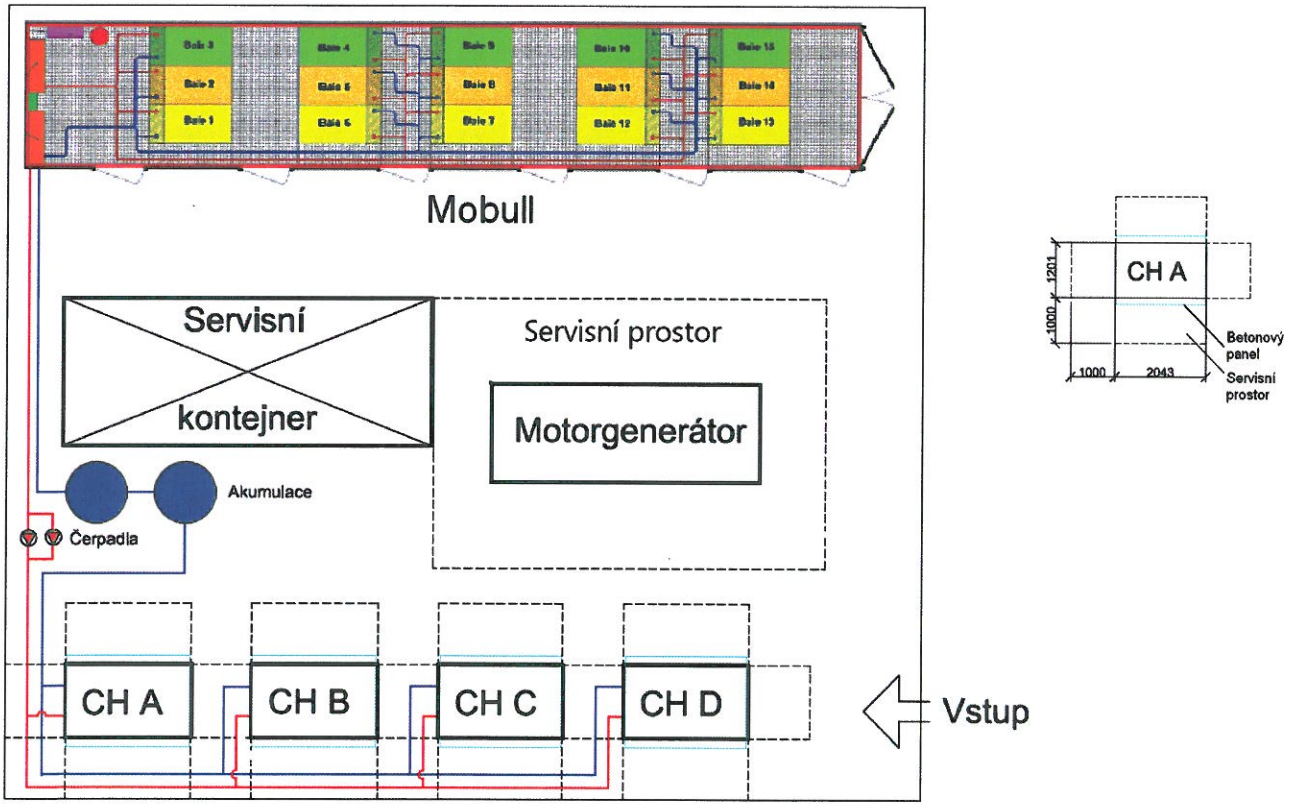
- Kontejner Mobull (45 stop) pro datové centrum (ICT systémy) o rozměrech půdorysu 13700 x 2440 mm a výšce 2900 mm.
- Servisní kontejner (20 stop) pro umístění UPS, rozvaděčů a ústředny pro řízení a monitorování rozvodu elektrické energie, půdorys servisního kontejneru je 6060 x 2440 mm a výška 2600 mm.
- Motor-generátor s dieselovým agregátem a generátorem o výstupním instalovaném výkonu 200 kW včetně nádrže na 500 l nafty. Vše umístěno v zamykatelné odhlučněné kapotě o půdorysu 4430 x 1640 mm a výšce 2230 mm. Uvedený objem nádrže umožňuje dodávat energii po dobu delší než 9 hodin při 100% zátěži.
- Komplet chladícího okruhu, který se skládá ze 4 kusů externích blokových chladících jednotek (každá o chladícím výkonu 64,4 kW tepelného výkonu), vyrovnávací expanzní jednotky a dvou zásobníků s rezervou 3000 l chladícího media (směs vody a propylenglykolu), dvou čerpadel na pohon chladícího media a propojovacího potrubí. Pro zajištění redundance chladicího systému jsou ze čtyř kusů jednotek vždy tři aktivní a čtvrtá je rezervní (vypnutá), role si v pravidelných intervalech mění (předpokládáme po týdnu – délku intervalu lze nastavit). Obdobně ze dvou oběhových čerpadel je vždy jedno v činnosti a druhé je rezervní a role si pravidelně mění. Jednotlivá bloková chladící jednotka zabírá půdorys 2050 x 1210 mm a je vysoká 1880 mm, každá bude umístěna na betonovém panelu (který je součástí dodávky).

Veškerá ICT technologie, která bude umístěna v racích mobilního datového centra, bude provozována v následujících podmínkách:

- napájení napětí 230V ±10 V AC
- napájení frekvence 50Hz
- teplota nasávaného vzduchu - může být nastavena v rozmezí 18-28 °C
- nekondenzující vlhkost
- relativní vlhkost 20-60%
- gradient vlhkosti maximálně 10% za hodinu

Dodávka Mobilního datového centra obsahuje umístění obou kontejnerů, kapotovaného motorgenerátoru, blokových chladících jednotek a zásobníku chladicí kapaliny na předem určené místo, instalaci venkovního rozvodu chladicí kapaliny a oběhových čerpadel, instalaci rozvodu napájecích NN a komunikačních kabelů mezi jednotkami v souladu s technickými normami a příslušnými předpisy, instalaci předepsaných jímacích zemnících tyčí, připojení na centrální napájení a na optickou síťovou, oživení jednotlivých systémů a provozní zkoušky. Nedílnou součástí je i zajištění předepsaných revizí (únik chladicí kapaliny, instalace UPS a rozvodů).

Předmětem dodávky je i pravidelná údržba a servis komponent Mobilního datového centra. Po dobu pronájmu budou dodržovány servisní požadavky a doporučení výrobců jednotlivých systémů, zajišťovány revize systémů Mobilního datového centra požadované legislativou a předpisy, zajišťovány opravy potřebné pro zajištění provozuschopnosti.



Rozmístění komponent mobilního datového centra

Kontejnery a kapotovaný motorgenerátor budou uloženy přímo na připravenou plochu, chladící jednotky budou instalovány na betonové panely, které jsou součástí dodávky.

Mobilní datové centrum je navrženo tak, aby bylo možno umístit, napájet a chladit jak infrastrukturu „Systému pro náročné výpočty", která je předmětem této nabídky, tak i infrastrukturu „Specializovaného systému".

V kontejneru datového centra bude instalováno kromě racků pro „Systém pro náročné výpočty" dalších 9 prázdných standardních 19" 42U racků EIA 310, celkem tedy 378 RU, připravených pro instalaci produktů třetích stran. Součástí návrhu je i veškerá infrastruktura a výkon pro chlazení „Specializovaného systému". Maximální příkon zařízení „Specializovaného systému" instalovaných v rámci jednom racku může být 40kW, což je hodnota vyšší, než bylo požadováno Zadavatelem (25 kW / rack).

Každý z racků je vybaven vodou chlazenými dveřmi a PDU.

Kontejner datového centra Mobull

Mobull jsou kontejnerová řešení pro modulární mobilní datacentrum poslední generace, jehož komponenty umožňují dosahovat vysoké prostorové výpočetní hustoty, která může dosáhnout až 160 TFLOPs na pouhých 30m2 podlahové plochy. To vše za použití standardních 19" EIA-310 racků, umožňujících použití běžného komoditního hardwaru.
Do jediného kontejneru je možné umístit až 15 42U racků do 5 řad po 3 racích. Až 40kW tepelného výkonu na každý z racků je chlazeno kapalinovým okruhem.
Řešení mobull spojuje vlastní vývoj společnosti Bull s průmyslovými standardy, čímž umožňuje použití nejen hardwaru Bull, ale většiny na trhu dostupných komponent.

Základní vybavení kontejneru datového centra:

- kontejner 45 stop – rozměry (d x h x v) 13700x2440x 2900 mm
- tepelně izolovaný kontejner
- boční zamykatelné servisní vstupy, čelní zamykatelný instalační vstup
- napájení 3-fázové 400V AC
- obsahuje 15 x 19" EIA-310 42U racků s dveřmi chlazenými kapalinou
- hydraulický rozvod chladicí kapaliny pro instalaci 15ti kapalinou chlazených dveří pro racky
- 5 montážních rámů, každý pro skupinu tří racků
- technické podlaží – prostor pro instalaci vedení pod falešnou kovovou podlahou
- 1 rozvaděč pro napájení 15 racků včetně rozvodu k jednotlivým rackům
- 1 rozvaděč pro napájení pomocných systémů (světla, protipožární, monitorovací systém atd.)
- 1 systém pro kontrolu vstupu
- systém pro detekci kouře (využívá snímače VESDA s dvojitou detekcí)
- hasicí systém - založený na zásobníku inertního nebo zhášecího plynu, který je uvolňován do prostoru v případě detekce kouře, je snižována koncentrace kyslíku a nebo i odebíráno teplo hořícímu plamenu (pro zhášecí plyn), a tím je dosaženo vysokého hasícího účinku
- systém pro oznámení aktivace hasícího systému (výstražný alarm, světelná signalizace pro evakuaci kontejneru)

- systém pro monitorování prostoru uvnitř kontejneru – dvě kamery
- systém pro detekci úniku chladícího media
- zařízení pro úvodní ohřev prostoru uvnitř kontejneru pro případ "studeného startu
- systém pro monitorování parametrů prostředí uvnitř kontejneru

Charakteristika a vybavení instalovaných racků

- hustota tepelného vyzařování maximálně 40 KW/h na 1 rack
- redundantní Power Distribution Units (PDUs)
- kapalinou chlazené dveře
- maximální průtok vzduchu 8,000 m^3/hod

Konfigurace kapalinou chlazené dveře racku

- dveře se montují na zadní stranu racku
- pro zvýšení chladící účinnosti pouze jeden tepelný výměník (vzduch voda) připojený přímo na primární rozvod chladicí kapaliny
- možnost otevření dveří o 180° - pro snadný přístup a možnost kontroly zadní části serverů
- kit pro fixaci kabelů
- sada 14 větráčků s elektronicky řízenými otáčkami (vzájemná redundance)
- 4 teplotní čidla
- trojcestný ventil pro regulaci průtoku chladicí kapaliny
- čidlo otevření dveří
- čidlo úniku chladicí kapaliny
- čidlo tlaku vzduchu
- řídící jednotka s Ethernetovým a sériovým portem, včetně vestavěného nástroje „Cool Door Console" pro vzdálenou správu a monitorování dveří
- dva napájecí zdroje (redundance N+N)

Servisní kontejner

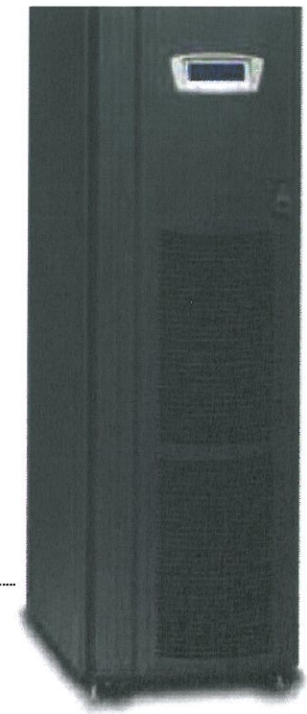
- kontejner 6060 x 2990 x 2800 mm (d x h x v)
- tepelně izolovaný rám kontejneru
- maximální akustický výkon $L_{WA\,max}$ 73 dB
- uzamykatelné dveře opatřené mříží
- obsahuje NN rozvaděče
- obsahuje modulární UPS Eaton 9390 typu 3/3fáze, 360kVA - 9390-3x120-NHS-4X1 (3 paralelně spojené moduly po 120 kVA) včetně externích baterií
- obsahuje nezávislé chlazení – zabudované dvě chladící jednotky (dva split komplety – venkovní jednotka, vnitřní jednotka, ovladač), dohromady s chladícím výkonem 17 kW
- obsahuje Regulátor a monitor spotřeby energie ATS-C120-LM-16-8-ETH s možností komunikace s PC prostřednictvím ethernetového portu, tři modulové elektroměry KWZ a měřící transformátory. Systém je určen pro monitorování spotřeby jednotlivých částí a bude poskytovat údaje pro vyhodnocování parametru PUE; zároveň se bude podílet na řízení

automatického řízeného snížení příkonu tak, aby byl možný dlouhodobý provoz ze záložního zdroje napájení – motorgenerátoru v době výpadku napájení ze vstupní elektrické sítě

- obsahuje ústřednu Elektronického zabezpečovacího systému EZS, vybavenou ethernetovým modulem pro komunikaci a hlášení alarmů, vnitřní prostor bude monitorován PIR čidlem a o čidle pro detekci kouře
- obsahuje kamerový systém CCTV pro sledování vnějšího prostoru Mobilního datového centra vymezeného oplocením – tvořen čtyřmi kamerami ve venkovním provedení a jednotkou pro nahrávání kamerového záznamu s dobou záznamu 24 hod, vybavenou ethernetovým portem pro vzdálený dohled a přehrávání záznamu,

Kapotovaný motor-generátor 200kW

- specializovaná odhlučněná zamykatelná kapota (na základě kontejneru ISO 20) s bočním a čelním vstupe
  - Rozměry (d x h x v) 4430 x 1640 x 2230 mm.
  - tlumič výfuku -30dB, určený do obytné zástavby
  - hladina akustického tlaku 69 dB ve vzdálenosti 7m
  - motorgenerátor MP 250 O, motor Volvo Penta, generátor MECC ALTE
  - jmenovitý základní výstupní výkon 250 kVA / 200kW
  - jmenovitý proud 361 A, napětí 400 V / 230 V, frekvence 50 Hz
  - ekologická záchytná vana integrovaná s kapotou pro zachycení provozních kapalin v případě havárie
  - nádrž na 500l motorové nafty, dodávka včetně plné nádrže
  - předehřev motoru
  - řídící panel automatický AMF 5
  - automatika startu: TTS - ATS box
  - automatická regulace napětí
  - rychlý náběh
  - dobíjení baterie
  - spotřeba 52l / hod při 100% zatížení



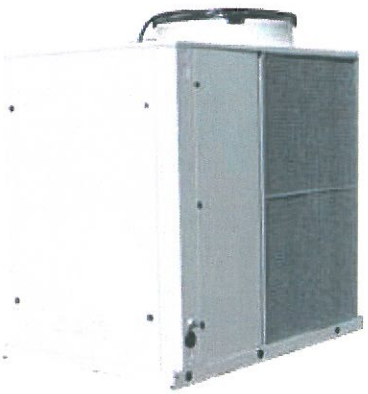
UPS pro zajištění bez-výpadkového provozu ICT systémů a pro překlenutí krátkodobých výpadků napájení z veřejné elektrické sítě

- Eaton 9390 typu 3/3fáze, 240kVA - 9390-3x120-NHS-4X1 (3 moduly 120 kVA)
- paralelní redundantní řazení Powerware HotSync® umožňuje paralelní řazení až osmi jednotek UPS, buď pro zvýšení výkonu, nebo zvýšení dostupnosti (redundance) – umožňuje efektivní sdílení zátěže bez jakéhokoli komunikačního propojení
- topologie dvojité konverze (bez transformátorová IGBT + PWM) izolující výstupní napájení od všech nepravidelností na vstupu

- hodnota vstupního účiníku 0,99 je zajištěna aktivní korekcí účiníku (PFC) a nelineární zkreslení vstupního proudu (ITHD) menší než 4,5%, zabraňuje interferenci s ostatními zařízeními na téže rozvodné síti a rozšiřuje kompatibilitu s motorgenerátory
- účinnost v režimu ESM (režim vysoké účinnosti) až 99%
- účinnost v režimu dvojité konverze 94% při maximálním výkonu
- maximální výstupní výkon 360 kVA / 324 kW (při účiníku 0,9)
- při výpadku jednoho modulu max.í výstupní výkon 240 kVA/216 kW
- možnost přetížení UPS
- bateriové stojany osazené 80ti kusy baterií
- možnost omezení maximálního nabíjecího proudu v konfiguraci UPS
- doba běhu na baterie minimálně 18 minut při zatížení 250 kW a (při poklesu napětí na článku na 1,8 V a s ohledem na účinnost UPS a stárnutí baterií)

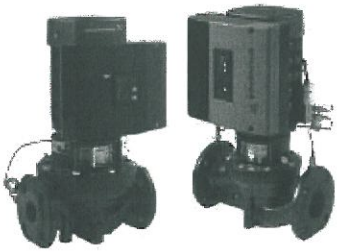
Blokové chladící jednotky

- moderní bloková chladící jednotka výrobce Emerson Network Power typu FG0 006-800
- freecooling aktivní již při minimálním rozdílu mezi teplotou vracející se chladicí kapaliny a vzduchem. Při modelaci roční spotřeby vycházející z průběhu denních teplot pro danou lokalitu za posledních několik let je úspora spotřeby elektrické energie více než 45% při trvalém maximálním chladícím výkonu proti systému bez aktivovaného freecoolingu (v praxi bude menší, neboť běžný provoz nebude pravděpodobně využívat plný chladící výkon jednotek)
- dva kompresory s celkovým příkonem 17,8 kW
- maximální chladící výkon 64,4 kW při 35°C
- maximální příkon 21,2 kW při 35°c
- jeden výparník
- dvě kondensační chladící smyčky a dvě smyčky pro freecooling
- jeden společný větrák pro všechny chladící smyčky
- pracují s teplotním spádem 14/18°C
- Akustický výkon $L_{WA\,max}$ 75 dB na jednotku



Oběhová čerpadla pro rozvod a oběh chladicí kapaliny

- čerpadlo Grundfos typ TPE 65-230/2 S
- 3-fázový 2-pólový motor
- elektronicky kontrolované otáčky motoru
- maximální příkon 3,3 kW
- jmenovitý tlak čerpadla PN16
- maximální průtok čerpadla (Q) 39 m^3 /h



1.1.2. ŘEŠENÍ NAPÁJENÍ S ENERGETICKÝMI OMEZENÍMI

Při návrhu systému napájení jsme vycházeli z následujících informací uvedených v zadání.

- celková spotřeba infrastruktury včetně chlazení může dosahovat max. 300 kW
- maximální výkon motorgenerátoru je 200kW
- možnost omezení příkonů jednotlivých ICT systému řízeným procesem power cappingu

Návrh zapojení jednotlivých komponent Mobilního datového centra a „Systému pro náročné výpočty" bude vytvořen především s ohledem na rovnoměrné zatížení fází napájení. Pro instalaci „Specializovaného systému" předpokládáme, že mu bude předcházet nejprve konzultace vhodného připojení k rozvodu napájení.

Při návrhu bereme rovněž v potaz:

- dočasné dodatečné teplotní přírůstky (dočasné otevření dveří, přírůstky způsobené prací obsluhy, vlivy okolního prostředí)
- latentní výkon klimatizace

Komponenty pro bezvýpadkové řešení napájení

Záložní zdroj EATON 9390 – napájí hlavně ICT systémy a jejich podpůrné systémy (chlazení) v rámci datového kontejneru Mobull, dále napájí oběhová čerpadla. UPS pracuje trvale v režimu ESM (režim vysoké účinnosti) s tím, že v případě přerušení napájení z veřejné elektrické sítě odebírá energii ze záložních baterií. Tím je zajištěno, že jsou ICT systémy izolovány od výpadku napájecího napětí ve veřejné elektrické síti a dále od většiny rušivých vlivů na externím napájení. Režim ESM je vhodný do prostředí, ve kterém se provozují ICT systémy, pro které poskytuje dostatečně kvalitní a rychle reagující ochranu a zároveň se vyznačuje vysokou účinností, což hraje významnou roli při úspoře nákladů v době, kdy se ceny energií neustále zvyšují. Funkcí UPS v navrženém uspořádání je odfiltrovat krátkodobé výpadky (v řádu minuty) externího napájení a tím omezit činnost motorgenerátoru jen na dobu, kdy dojde k dlouhodobému výpadku. Tím je zachován určitý provozní komfort a také se omezuje negativní hygienický a ekologický vliv na okolní zástavbu. Pokud dojde k výpadku jedné z redundantních jednotek UPS, napájení zajistí dva zbývající moduly UPS, které jsou dimenzovány tak, aby pokryly napájení i pro stav s nejvyšším uvažovaným příkonem části připojené na UPS – to je vnější teplotu 35°C a zároveň oba ICT systémy budou pracovat s maximálním příkonem [redacted]

200kW motorgenerátor (možnost přetížení o 10% až 500 hodin ročně) – poskytuje zdroj záložního napájení v době dlouhodobého výpadku. Bezobslužný automatický start a krátká startovací doba (řádu jednotek sekund) a automatická regulace výkonu generálu podle aktuální úrovně zátěže umožňuje rychlou reakci na výpadek hlavního napájení. Součástí motorgenerátoru je i integrovaná nádrž na 500l nafty – z důvodu efektivnosti bude nafta doplňována vždy, když objem nafty poklesne pod 400l (vlivem spotřeby během výpadku napájení nebo v průběhu automatických testů motorgenerátoru). Tím je garantována minimální rezerva pohonných hmot určená pro 7,5 hodin provozu motorgenerátoru při maximálním zatížení (200kW).

Ústředna ATC spolu s měřícími transformátory – monitoruje, zaznamenává a vyhodnocuje stav napájení celého komplexu. Obousměrně komunikuje s UPS i motorgenerátorem, vydává pokyn ke startu motorgenerátoru, k zahájení přechodu ICT systémů do power camping modu a v případě

výpadku napájení z veřejné sítě odpojuje nebo přepojuje některé systémy dle dále uvedených pravidel.

Výpadek napájení z veřejné elektrické sítě je detekován UPS (tím je zabezpečena i odolnost systému proti poruchám v rámci rozvaděčů před UPS) a zpráva je odeslána na ústřednu ATC, která řídí další procesy. Zpráva o výpadku je dále odeslána prostřednictvím SMS a emailu na předem určená čísla a adresy. Obnovení napájení bude registrováno ústřednou ATS a stejným způsobem oznámeno (SMS, email). Obě události budou zaznamenány do logů a to systémem monitoringu spotřeby i IPS software pro UPS.

Základní principy navrženého řešení napájení.

Navržené řešení napájení a průběh v době výpadku napájení z veřejné elektrické sítě slouží především jako demonstrace robustnosti a odolnosti navrženého komplexu a ověření, že splňuje požadovanou funkčnost. Některé informace (detailní popis a některé parametry „Specializovaného systému" při přechodu do režimu se sníženou spotřebou) nejsou v době přípravy nabídky známy a je možné, že nebudou známy ani v tobě instalace a akceptace Mobilního datového centra. Proto tento návrh nemusí být konečný a jediný možný a poskytneme Zadavateli v rámci naší implementace (případně v době instalace „Specializovaného systému") příslušnou součinnost pro zapracování případných připomínek, návrhů a požadavků na změnu například časového průběhu apod. (pokud to kapacita instalovaných zařízení umožní).

V době výpadku napájení z veřejné elektrické sítě uvažujeme s omezením maximálního výkonu „Specializovaný systém" na 50 kW (definováno Zadavatelem) a „Systému pro náročné výpočty" z [redacted] Obdobných kalkulací lze udělat řadu, a použité SW nástroje poskytují dostatečnou flexibilitu a umožňují kombinaci různých opatření s ohledem na rozdělení uzlů do skupin podle určení, priorit, nebo podle jiných pravidel. Přesná strategie rozdělení uzlů do skupin a nastavení detailních pravidel pro power capping bude konzultována v době implementace Sytému na místě instalace nebo případně i dříve – dle volby Zadavatele. Za rozumné považujeme provést příslušné konzultace v okamžiku po absolvování příslušných školení administrátory systému, kdy budou mít detailnější představu o vlastnostech management software. Pro dosažení „bezpečnějších" výsledků používáme při kalkulacích pro účinnost UPS hodnotu 96% (a nikoliv maximální hodnotu 99%).

[redacted]

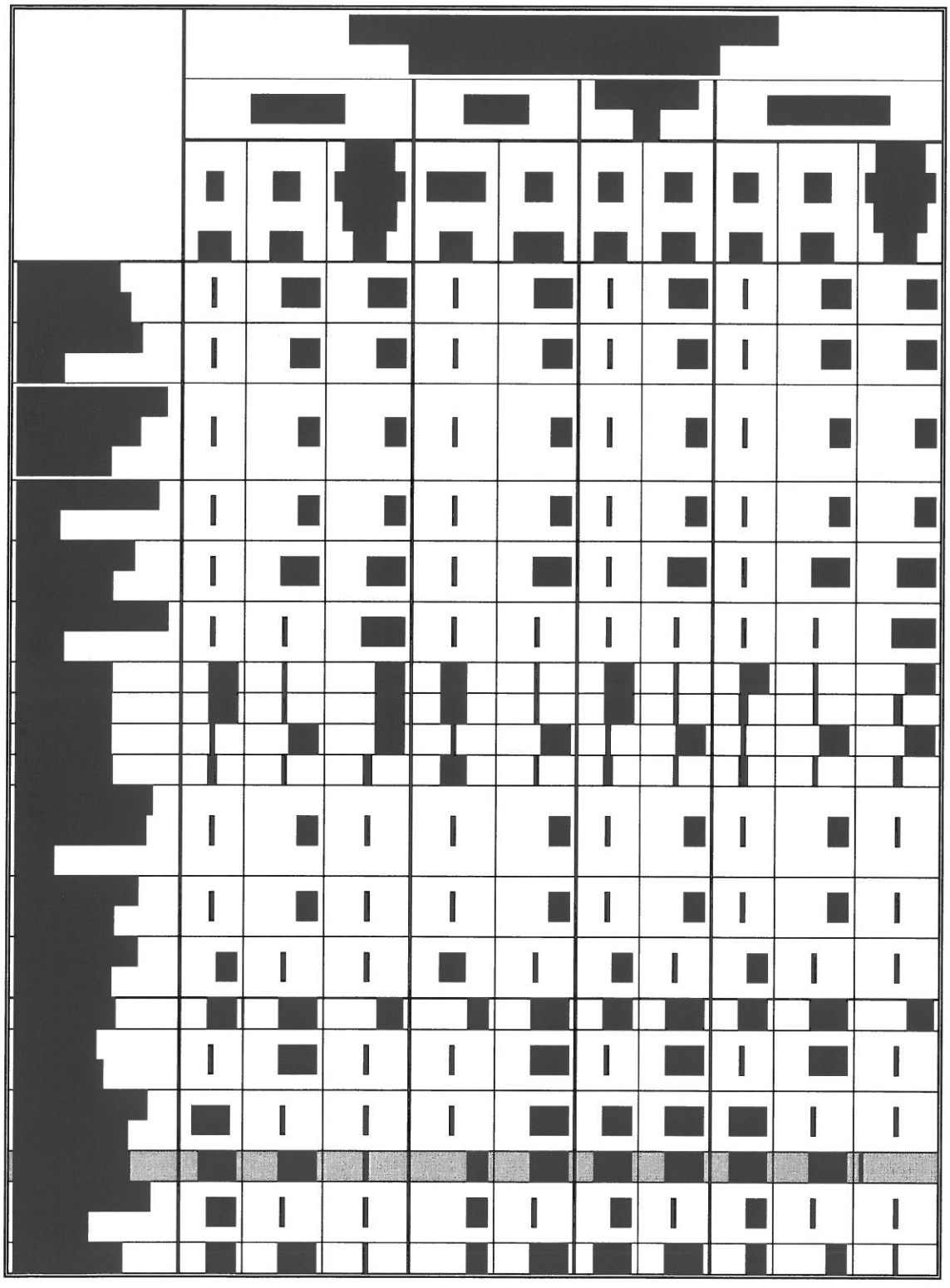
[redacted]

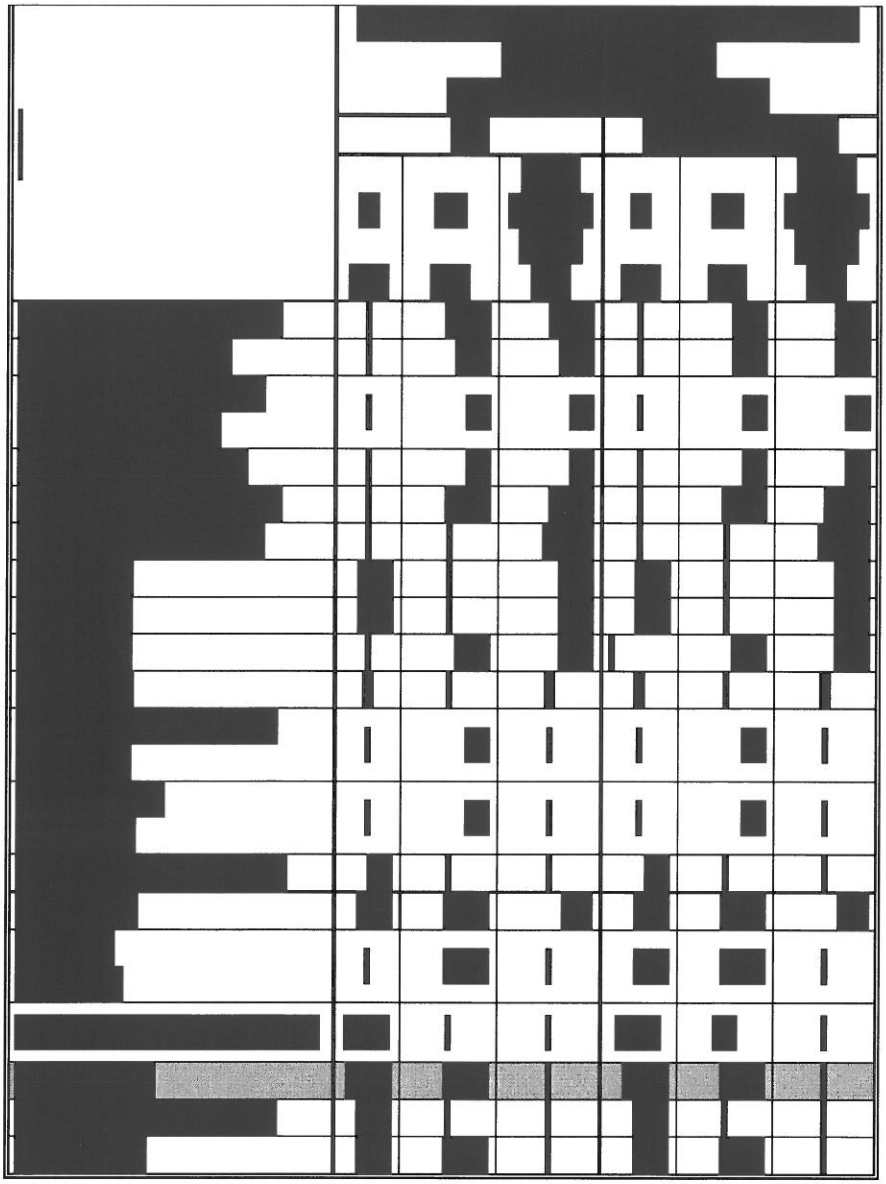
[redacted]

[redacted]

[redacted]

Následující tabulky ukazují energetickou (a chladící) bilanci pro jednotlivé fáze a různé vnější podmínky





1.1.3. ŘEŠENÍ TESTOVÁNÍ NÁHRADNÍCH ZDROJŮ NAPÁJENÍ

Testování náhradních zdrojů napájení lze rozdělit na statickou a dynamickou metodu.

Statické testování: Měření parametrů jednotlivých komponent, zejména měření stavu baterií.

Dynamické testování: Během testování budou simulovány všechny provozní stavy zařízení. Reálný výpadek bude nahrazen odpojením přívodního napájení. Testování proběhne částečně na zátěži jednotlivých zařízení nebo na IT zařízení, v datacentrech bude spuštěn výpočetní algoritmus, který zatíží tato zařízení na maximální možný výkon. Délka doby testování je 30 minut od výpadku hlavního přívodního deon. Základní testování proběhne v rámci akceptačních testů. Testování v době provozu Mobilního datového centra bude podrobně popsáno v Detailní technické specifikaci, včetně způsobu simulace zátěže „Specializovaným systémem" (formou odporové zátěže), pokud by nebyl v době akceptačních testů ještě nainstalován.

1.1.4. CHLAZENÍ

Chladicí systém pro kontejner datového centra (Mobull) je tvořen čtyřmi paralelně řazenými blokovými chladícími jednotkami pracujícími v režimu redundance 3+1. V navrženém řešení je jedna jednotka vždy připojena na UPS a ostatní tři přímo na napájení z veřejné elektrické sítě (nebo motorgenerátor v případě výpadku napájení). Ze tří jednotek připojených na síť jsou vždy dvě aktivní a třetí rezervní ve stanby režimu – role si automaticky vždy po týdnu mění (lze nastavit i jinou periodu). Čtvrtá chladící jednotka je připojena na UPS a je trvale aktivní – vždy jednou za 3 měsíce v rámci pravidelné údržby chladících jednotek bude na UPS přepojena jednotka jiná – tím bude zajištěno, že jednotky budou zatížené rovnoměrně.

Účel připojení jedné jednotky na UPS je zajistit větší flexibilitu a rezervu chladícího výkonu v době výpadku napájení, neboť chladící jednotky potřebují v takové situaci určitou dobu na zotavení a obnovení činnosti v plné kapacitě chladícího [redacted]

Kapacita zásobníků s chladicí kapalinou je volena právě s ohledem na tuto dobu tak, aby umožnila překlenout dobu [redacted] i v případě, že jsou odpojeny všechny chladící jednotky (k výpadku dojde v době, kdy je shodou okolností nefunkční chladící jednotka připojená na UPS) – s jednou funkční jednotkou se tato doba prodlužuje.

Blokové chladící jednotky jsou vybaveny funkcí tzv. freecoolingu, kdy je chladící medium před vstupem do výparníku přesměrováno pomocí trojcestného ventilu na přídavnou (freecoilingovou) chladící smyčku. Tím je umožněna úspora energie i v době, kdy není venkovní teplota dostatečně nízká, aby zajistila kompletní chlazení a kompresor dochlazuje jen zbývající část tepelné zátěže. V našem případě jsou kondenzační i freecoolingové chladící smyčky umístěny za sebou a chlazeny stejným větrákem. Tím je dosaženo nejen úspory místa, ale i dále zvýšena účinnost chladícího cyklu. Pro navržený chladící systém se freecooling začne uplatňovat již od venkovní teploty 18°C a při průběhu teplot během kalendářního roku vycházejících z průměrných dat pro Ostravu za několik posledních let je simulovaná úspora energie více [redacted] maximálním tepelném výkonu jednotky. V praxi bude úspora o něco nižší, neboť nelze předpokládat, že by ICT systémy pracovaly po celou uvažovanou dobu na maximální příkon.

V servisním kontejneru bude zabudováno nezávislé chlazení – dvě podstropní chladící jednotky.

Kontejner – kapota motorgenerátoru je určena pro venkovní provoz a proto nemá zabudováno žádné chlazení.

Chlazení vlastních ICT systémů je založeno na principu udržování konstantní teploty v rámci celého datového kontejneru. Vodou chlazené dveře jsou instalovány v zadní části racku a jejich součástí je tepelný výměník ochlazovaný proudícím chladícím mediem. Sada větráčků nasává vzduch z racku přes výměník a řídící jednotka na základě informací z teplotních čidel reguluje průtok chladicí kapaliny (prostřednictví trojcestného ventilu) a otáčky větráčků tak, aby teplota výstupního vzduchu byla shodná s teplotou nastavenou pro vnitřek kontejneru. V případě přerušení dodávky chladicí kapaliny (snížení tlaku kapaliny vlivem úniku, porucha ventilu apod.) řídící jednotka zvýší otáčky větráčků tak, aby byl rack stále chlazen. Toto uspořádání chlazení jednak eliminuje jev studené a teplé uličky a tím, že chladí přímo na místě vzniku tepla má jednu z nejvyšších účinností ze všech uvažovaných systémů.

Vestavěný nástroj „Cool Door Console" je dostupný z Internetového prohlížeče prostřednictvím nastavené IP adresy. Tento nástroj umožňuje vzdáleně nastavovat teplotu výstupního vzduchu monitorovat aktuální hodnoty parametrů (otevřené/zavřené dveře, teplotu, tlak vzduchu, stav napájecích zdrojů, chladícího subsystému, kapalinového ventilu, teplotních senzorů, detektoru úniku kapaliny a rychlost otáček jednotlivých větráčků). Pomocí tohoto nástroje lze aktivovat a konfigurovat systém zasílání varovných a informačních hlášení (SNMP trapy, maily, filtry rozesílaných alarmů apod.).

[redacted]

[redacted]

[redacted]

[redacted]

1.1.5. ZABEZPEČENÍ

Zabezpečení Mobilního datového centra lze rozdělit na dvě části – vnější a vnitřní.

Vnější zabezpečení pokrývá prostor vymezený plotem se vstupní brankou, vše připravené Zadavatelem dle podmínek Zadávací dokumentace. Tento vymezený prostor bude nepřetržitě

monitorován kamerovým systémem CCTV tvořeným čtyřmi kamerami ve venkovním provedení a jednotkou pro nahrávání kamerového záznamu s dobou záznamu 24 hod, vybavenou ethernetovým portem pro vzdálený dohled a přehrávání záznamu – vše lze propojit na centrální pult ochrany Zadavatele.

Vstupy do všech kontejnerů i kapotovaného motorgenerátoru budou zajištěn uzamykatelnými dveřmi. Součástí Servisního kontejneru je ústředna ESZ, na kterou budou napojena PIR čidla a čidla pro detekci kouře/ohně jak pro servisní kontejner, tak pro kapotovaný motorgenerátor. V případě narušení prostoru nebo detekci požáru je vyhlášen alarm a prostřednictvím emailu a SMS odeslána zpráva na předem určená čísla a adresy.

Kontejner datového centra Mobull je vybaven vlastním monitorovacím/zabezpečovacím, kamerovým a protipožárním systémem. Protipožární systém v případě detekce kouře aktivuje velmi účinný hasící systém a to distribucí plynu z tlakové lahve – Zadavatel si může vybrat ze dvou typů plynu – inertní plyn (hašení probíhá na principu snížení obsahu kyslíku pod specifickou limitní hodnotu potřebnou pro hoření – pro běžná prostředí se obvykle udává hodnota cca 13,8%) nebo plyn zhášecí (kromě efektu snížení obsahu kyslíku také specifická složka plynu absorbuje teplo plamenů fyzikálním působením případně chemickou reakcí). Vzhledem k tomu, že podmínkou užití některých plynů je okamžitá evakuace prostoru (např. vysoká koncentrace CO2 je člověku nebezpečná), je systém vybaven akustickým a optickým alarmem, který se aktivuje spolu s hasícím systémem.

Zabezpečovací systém kontejneru datového centra (Mobull) je integrován s monitorovacím systémem clusteru. Kromě údajů potřebných pro kontrolu a administraci clusteru (tlak a teplota chladicí kapaliny, její únik, teplota a vlhkost uvnitř kontejneru, spotřeba jednotlivých racků, otevření vodou chlazených dveří apod.) shromažďuje i další informace a alarmy: detekci kouře, vyhlášení požárního alarmu, spuštění hasícího systému atd.. Součástí je i evidence vstupu osob do prostoru kontejneru. Zabezpečovací systém zároveň rozesílá alarmy a informační zprávy dle předem nastavených filtrů.

Vnitřek kontejneru je monitorován kamerovým systémem (dvě kamery), který umožňuje vzdálenou vizuální kontrolu vnitřního prostoru.

1.1.6. ENERGETICKÁ NÁROČNOST ŘEŠENÍ

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

Hodnota „roční PUE Mobilního datového centra při provozu Systému pro náročné výpočty" je 1,237.

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

1.1.7. SOUČINNOST SPECIALIZOVANÝ SYSTÉM

Součástí dodávky je součinnost dodavatele se zadavatelem a budoucím dodavatelem Specializovaného systému na realizaci Specializovaného systému v Mobilním datovém centru.

Tato součinnost je podmíněna včasným předáním technické dokumentace a harmonogramu instalačních prací Specializovaného systému zadavatelem (respektive budoucím dodavatelem Specializovaného systému) společnosti Bull ke konzultaci a připomínkám. Technická dokumentace by měla obsahovat zejména návrh rozmístění jednotlivých systémů do racků, jejich připojení na napájení (typ, ke kterým PDU), maximální příkon, požadavky na chlazení, způsob připojení k Ethernet a Infiniband sítím, případně další specifické požadavky ze strany budoucího dodavatele Specializovaného systému. Společnost Bull provede kontrolu návrhu z hlediska používání infrastruktury včetně rovnoměrnosti zatížení jednotlivých fází napájecí soustavy) Mobilního datového centra, případně navrhne úpravy ať už návrhu nově instalovaného systému nebo systému původního.

Vzhledem k požadavku odstavce 10.10 Závazného vzoru smlouvy, ve kterém Zadavatel stanovil, že povinnost Dodavatele odstranit vadu Díla nebo Mobilního datového centra nezaniká v případě, kdy tuto vadu způsobil nesprávným užíváním či jiným způsobem Objednatel nebo třetí osoba, je požadavek na předložení znalosti harmonogramu instalace (především fyzické) Specializovaného systému pro dodavatele Bull důležitý a jeho splnění mu umožní zejména:

- zvýšit preventivně na dobu instalace a implementace úroveň dohledu nad instalovaným systémem.
- účastnit se na místě instalace fyzické instalace systému a instruovat pracovníky dodavatele Specializovaného systému o rozmístění a účelu jednotlivých PDU, o správné manipulaci s chladícími dveřmi, případně dalšími částmi instalované infrastruktury
- účastnit se prvotního připojení na napájení a testovacího spuštění jednotlivých částí Specializovaného systému

Proto navrhujeme, aby Zadavatel, jakmile určí dodavatele Specializovaného systému, svolal s dostatečným předstihem koordinační schůzku všech tři stran, na které se projednají předběžné termíny a základní detaily vzájemné součinnosti.

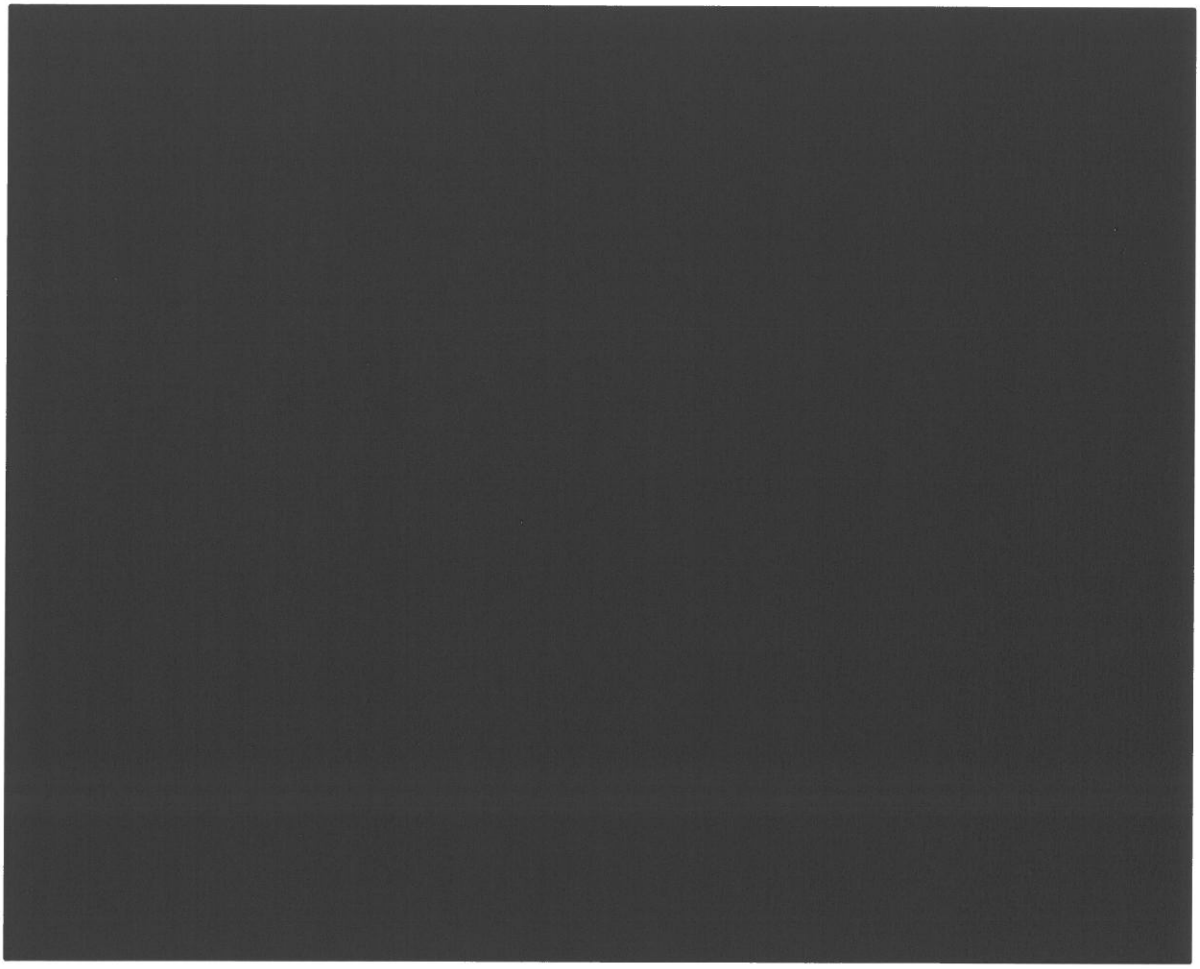
Obdobným způsobem navrhujeme zorganizovat schůzku všech tří stran před naplánováním stěhování ICT systémů do Stacionárního datového centra. Vzhledem k tomu, že ke stěhování budou použity racky z Mobilního datového centra, předpokládáme velmi úzkou spolupráci s dodavatelem Specializovaného systému a pečlivé detailní naplánování celé akce tak, aby činnosti jednotlivých dodavatelů na sebe časově vhodně navazovaly.

1.2. Návrh Systému pro náročné výpočty

Předmětem dodávky je komplexní řešení Systému pro náročné výpočty tj. komplex výpočetních, úložných síťových a dalších systémů a softwarového řešení včetně implementace.

Navrhované řešení je založeno na nové řadě bullx B510, 3. generaci architektury INCA (INtegrated Cluster Architecture) která je vybavena procesory Intel® Xeon® SandyBridge-EP. Výpočetní uzly

malého clusteru jsou propojeny sítí s vysokou propustností a nízkou latencí Infiniband QDR. Řešení rovněž obsahuje vysoce výkonná úložiště pro dočasná data výpočtu (SCRATCH) a domovská data uživatelů (HOME). Součástí dodávky je i HW infrastruktura pro hostování virtuálních serverů, databázové a další servery, datové úložiště s blokovým přístupem včetně SAN sítě, která k němu zprostředkuje připojení, a řešení pro zálohování dat.



Obr. 1 – Globální architektura Systému pro náročné výpočty

1.2.1. NÁVRH VÝPOČETNÍCH, ÚLOŽNÝCH A SÍŤOVÝCH SYSTÉMŮ

1.2.1.1. VÝPOČETNÍ SYSTÉMY

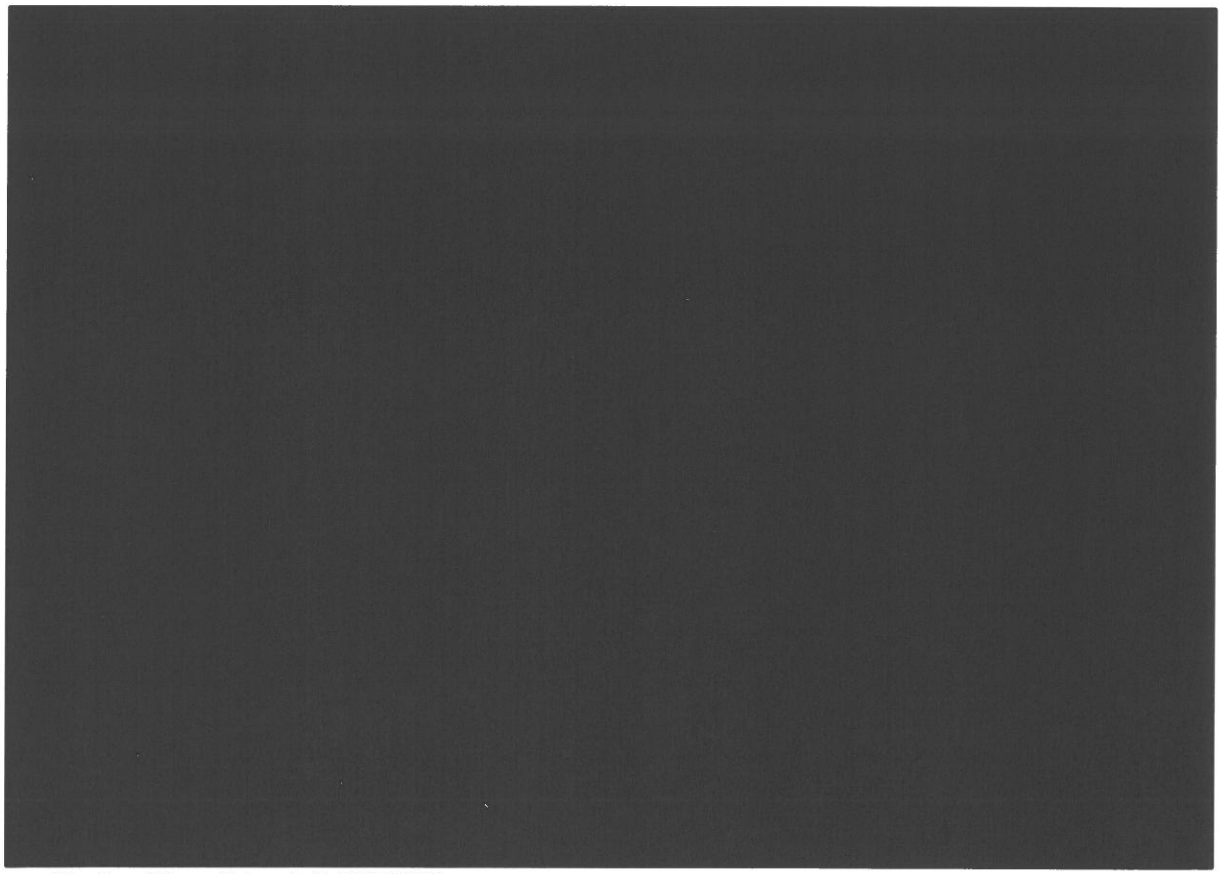
Výpočetní uzly Systému pro náročné výpočty (dále SNV) tvoří celkem 13 bullx blade chassis obsahujících celkem 207 výpočetních uzlů, z toho 27 uzlů je vybaveno jedním výpočetním akcelerátorem. Z těchto 27 uzlů jsou celkem 4 uzly vybaveny akcelerátorem typu Intel Xeon Phi architektury Many Integrated Cores, 23 uzlů je vybaveno GPU akcelerátorem technologie CUDA. Uzly bez akcelerace jsou realizovány žiletkovými servery typu bullx B510, jejichž provedení umožňuje provoz dvou samostatných výpočetních uzlů v jedné z celkem 9 pozic serverového chassis. Akcelerované uzly jsou tvořeny modelem bullx B515, každý takovýto uzel zabírá jednu z 9 pozic chassis.

Řízení a provoz výpočetních uzlů SNV zajišťuje obslužná infrastruktura, která je dimenzována tak, aby zajistila spolehlivý, bezpečný, rychlý a efektivní provoz SNV. Infrastruktura je realizována především dvěma administračními servery (admin server) tvořící vysoce dostupný pár se sdíleným úložištěm technologie Fibre channel 8Gbps. Tento vysoce dostupný pár pracuje v režimu active-active. Pro práci uživatelů SNV je cluster vybaven dvěma aktivními přístupovými servery (login node), které oba současně obsluhují přistupující uživatele SNV.

1.2.1.2. ÚLOŽNÉ SYSTÉMY

Souborové datové úložiště SCRATCH je řešeno pomocí produktu bullx PFS, které je založeno na klastrovém souborovém systému LUSTRE. Úložiště tvoří jedna vysoce dostupná buňka pro přístup k metadatům souborového systému (MDT – MetaData Target) a jedna vysoce dostupná buňka pro přístup k souborovým datovým objektům (OST – Object Storage Target). MDT buňka je tvořena dvěma metadata servery (MDS) s jedním sdíleným diskovým polem NetAppE2600. OST buňka je tvořena dvěma object storage servery (OSS) a dvěma sdílenými diskovými poli NetApp E5400. Výpočetní a obslužné uzly provozované v prostředí operačního system Linux přistupují k souborovému úložišti SCRATCH přímo pomocí nativního klienta souborového systému Lustre, pro přístup uzlů provozovaných v prostředí operačního systému Windows je k dispozici SCRATCH gateway server, který re-exportuje datový prostor úložiště SCRATCH pomocí protokolu CIFS.

Řešení úložiště HOME je založeno na diskovém poli NetApp E5400 se dvěma head-nody zprostředkujícími k němu přístup pomocí síťového souborového systému NFS. [redacted]



Obr. 2 - schéma přístupu k úložišti HOME

[redacted]

Výpočetní a obslužné uzly provozované v prostředí operačního systému Linux přistupují k souborovému úložišti HOME přímo pomocí síťového souborového systému NFS, pro přístup uzlů provozovaných v prostředí operačního systému Windows je k dispozici HOME gateway server, který re-exportuje datový prostor úložiště HOME pomocí protokolu CIFS. HOME gateway server dále umožňuje interní přenosy pomocí protokolů HTTPS, SCP/SFTP a FTPS. Tento server rovněž disponuje 10gbps Ethernet konektivitou do LAN/WAN routeru pro realizaci externích přenosů dat z a do HOME storage všemi výše zmíněnými protokoly, aniž by k tomu bylo potřeba využívat přístupových serverů, které mohou být momentálně vytíženy prací uživatelů a přenosy jejich pracovních dat.

Řešení úložiště s blokovým přístupem je založeno na diskovém poli NetApp E2600

[redacted]

1.2.1.3. SÍŤOVÉ SYSTÉMY

Architektura vzájemného propojení uzlů výpočetní sítí je plně neblokující fat-tree topologie realizovaná technologií Infiniband QDR s přenosovou rychlostí 40Gbps. Okrajové switche propojovací sítě pro připojení výpočetních uzlů jsou tvořeny 36-portovými QDR switch moduly bullx blade chassis. Okrajové switche pro připojení obslužné, virtualizační a databázové infrastruktury, a dále páteřní switche propojovací sítě tvoří 36-portové switche Mellanox IS5030Q. Oba typy InfiniBand switchů jsou založeny na technologii Mellanox InfiniScale IV, což zaručuje bezproblémovou součinnost zařízení. Všechny porty používají 4X agregaci.

Lokální síť LAN určená pro uživatelskou komunikaci, Admin síť určená pro administrační komunikaci jednotlivých komponent Systému pro náročné výpočty, a tzv. IPMI síť sloužící ke vzdálenému managementu serverů jsou realizovány technologií Ethernet 1000BASE-T. Použity jsou switche řady CISCO Catalyst C3750.

SAN síť určená pro propojení Úložiště s blokovým přístupem (BLOCK storage) s virtualizační infrastrukturou a databázovými a dalšími servery je tvořena dvěma 24-portovými switchy Brocade 300 technologie Fibre Channel s přenosovou rychlostí 8 Gbps. SAN síť je rovněž použita pro propojení

Calypso Media serveru Zálohovacího system s páskovou knihovnou Quantum Scalar i500, která obsahuje 10 2-portových mechanik LTO-5 s Fibre Channel konektivitou. Pro snížení potřebného počtu portů a kabelů pro připojení mechanik jsou použity 4 Fibre Channel I/O blades.

1.2.2. NÁVRH SOFTWAROVÉHO ŘEŠENÍ

Jako systémový software nabízíme bullx supercomputer suite, který se používá např. ke správě Tera100, jednoho z největších superpočítačů na světě patřící výzkumné organizaci CEA. bullx supercomputer suite je komplexní softwarový balík, který pokrývá všechny oblasti správy superpočítače, jeho dat a aplikací. To zahrnuje jednak všechny komponenty potřebné k instalaci, nasazení, správě a monitoringu superpočítače, pak rovněž software pro vývoj, provoz a ladění HPC aplikací. Tato sada je založena na 64-bitovém Linuxu v kombinaci s "Best Of Breed" open source softwarem a předními otevřenými standardy s přidanou hodnotu vlastního vývoje společnosti Bull. Součásti bullx supercomputer suite jsou integrovány do jednotného celku se společným databázovým úložištěm konfigurací a stavových informací jednotlivých komponent a aplikací celého řešení. To umožňuje efektivněji využívat prostředky systému s vysokou mírou spolehlivosti běhu Vašich aplikací.

- **bullx Management Center** poskytuje nástroje pro systémovou administraci a provoz superpočítače. Skládá se z 3 hlavních částí
  - **Infrastructure Manager**
  - **Software manager**
  - **Monitoring & Control Manager**

- **Application management**
  - **PBS Pro**
  - **bullxMPI**

- **Data management**
  - bullx PFS – Lustre

Pro účely jednotné správy uživatelských účtů v obou prostředích operačních systémů Linux a Windows a pro účely přístupu ke sdíleným datovým úložištím z obou těchto prostředí bude subcluster v režimu operačního systému Windows provozován buď pod centrální adresářovou službou LDAP, nebo bude vyvinut mechanismus pro synchronizaci dat z centrálního LDAP adresáře do separátního Active Directory adresáře, který bude k dispozici uzlům subclusteru v režimu OS Windows.

1.2.3. VÝKONOVÉ A KAPACITNÍ PARAMETRY ŘEŠENÍ

Výpočetní uzly bez GPU akcelerace

	Jádro	Socket	Uzel	Celkem
Počet	2880	360	180	-
Špičkový výkon CPU	19,2 GFLOPs	153,6 GFLOPs	307,2 GFLOPs	55,3 TFLOPs
RAM	4 GB	32 GB	64 GB	11,520 TB

Výpočetní uzly s GPU akcelerací

	Jádro	Socket/GPU	Uzel	Celkem
Počet	432	54	27	-
Špičkový výkon CPU	18,4 GFLOPs	147,2 GFLOPs	294,4 GFLOPs	7,94 TFLOPs
Špičkový výkon MIC (orientační hodnoty)	-	1000 GFLOPs	1000 GFLOPs	4 TFLOPs
Špičkový výkon GPU	-	665 GFLOPs	665 GFLOPs	15,3 TFLOPs
RAM	4 GB	32 GB	64 GB	1,728 TB

Efektivita HP LINPACK benchmarku při měření výkonu pouze CPU dosahuje 87%.
Celkový teoretický špičkový výkon Rpeak všech CPU je 63,24 TFLOPs, celkový Rmax je 55,024 TFLOPs

Efektivita HP LINPACK benchmarku při měření výkonu pouze GPU akcelerátoru dosahuje 55%.
Efektivita HP LINPACK benchmarku při měření výkonu pouze MIC akcelerátoru dosahuje 70%.

Čistá kapacita úložiště SCRATCH činí 135 TiB po odpočtu režie zabezpečení RAID a očekávané 7% režije souborového systému Lustre.

Dlouhodobě udržitelné výkonnostní parametry jsou
6GiB/s read@512kiB block
6GiB/s write@512kiB block
18,03 kIOPs 75/25% read/write@4kiB block

Čistá kapacita úložiště HOME činí 304 TiB po odpočtu režije zabezpečení RAID a očekávané 5% režije podkladového souborového systému ext3.

Dlouhodobě udržitelné výkonnostní parametry jsou
2GiB/s read@512kiB block
2GiB/s write@512kiB block
42,07 kIOPs 75/25% read/write@4kiB block

Čistá kapacita úložiště BLOCK činí 45,8 TiB po odpočtu režije zabezpečení RAID.

Dlouhodobě	udržitelné	výkonnostní	parametry	jsou
2.5GiB/s		read@512kiB		block
1GiB/s		write@512kiB		block
11,4 kIOPs 75/25% read/write@4kiB block				

Zálohovací řešení má k dispozici 12TB deduplikovaného diskového prostoru pro staging záloh.

Nekomprimovaná kapacita páskové knihovny činí 544,5 TB.
Rychlost zápisu při použití všech 10 LTO-5 mechanik je 5,04 TB/h

Podrobné výkonnostní a kapacitní parametry jsou uvedeny v souboru „VŠB_Malý cluster_Příloha č. 6 ZD - Technické parametry nabídky.xlsx", který je nedílnou součástí nabídky.

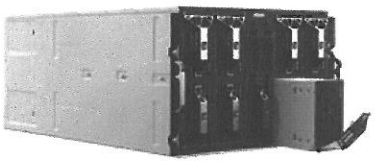
Hodnoty uváděné pro akcelerátory typu MIC jsou pouze orientační. Bull má k dispozici bližší údaje od výrobce a výsledky vlastních měření, tyto informace však podléhají striktním pravidlům, proto je nelze poskytnout, přestože je uzavřena dohoda NDA. Detailní hodnoty parametrů akcelerátorů typu MIC budou zveřejněny v rámci Detailní technické specifikace.

1.3. Jednotlivé systémy

Tato kapitola obsahuje popis funkcionality a vlastností jednotlivých komponent řešení.

1.3.1. VÝPOČETNÍ CLUSTER

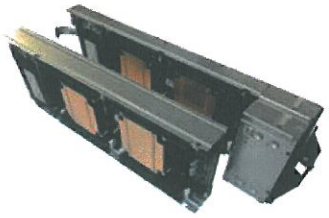
13x bullx Blade chassis



bullx Blade Chassis

- 9 double width blade bullx B510 including 2 bi-sockets nodes
- 1 InfiniBand Switch QDR 36 ports 40 Gbps
- 1 management module, CMM
- 1 Ethernet switch Gbits 24 ports
- 2 FANS
- 1 LCD Unit
- 7U

90x bullx B510 double blade



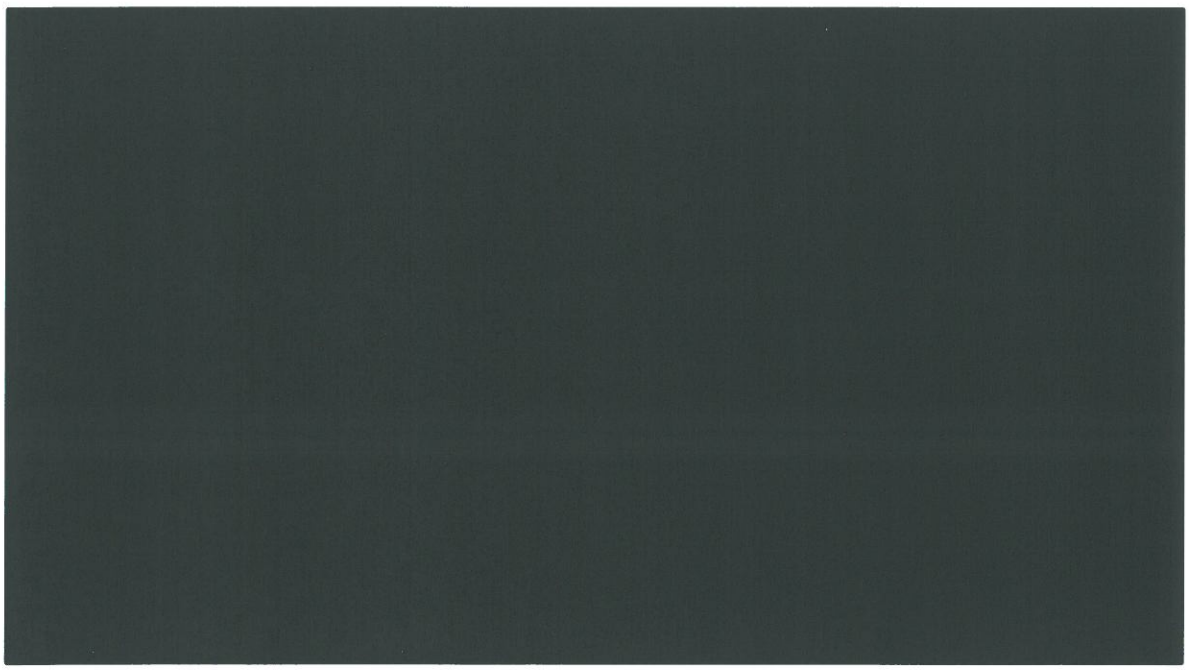
bullx B510

bullx B510

- 2 SandyBridge CPU Boards (SCB) – dual socket
- 2 common redundant large fans
- 2 SATA hard disk drives 2.5"

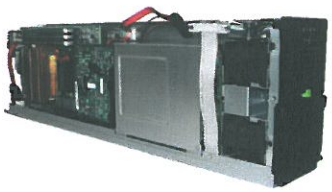
SCB

- 2 Intel Sandy Bridge EP E5-2665, 8c/2.4GHz/20Mo/8GT/s
- 64 GB ECC SDRAM (8 x 8 GB DDR3 DIMM 1600 MHz)
- 1 x 300 GB SATA 2,5" 7,2 kRPM HDD
- 1 x ethernet port 1 Gbits/s
- 1 single port IB HCA, QDR support
- 1 integrated BMC
- 1 Dual port 1Gb Ethernet controller



Obr. 3 - Blokové schéma bullx B510 SCB

27x bullx B515 - 23x NVIDIA M2090, 4x Intel Xeon Phi



bullx B515

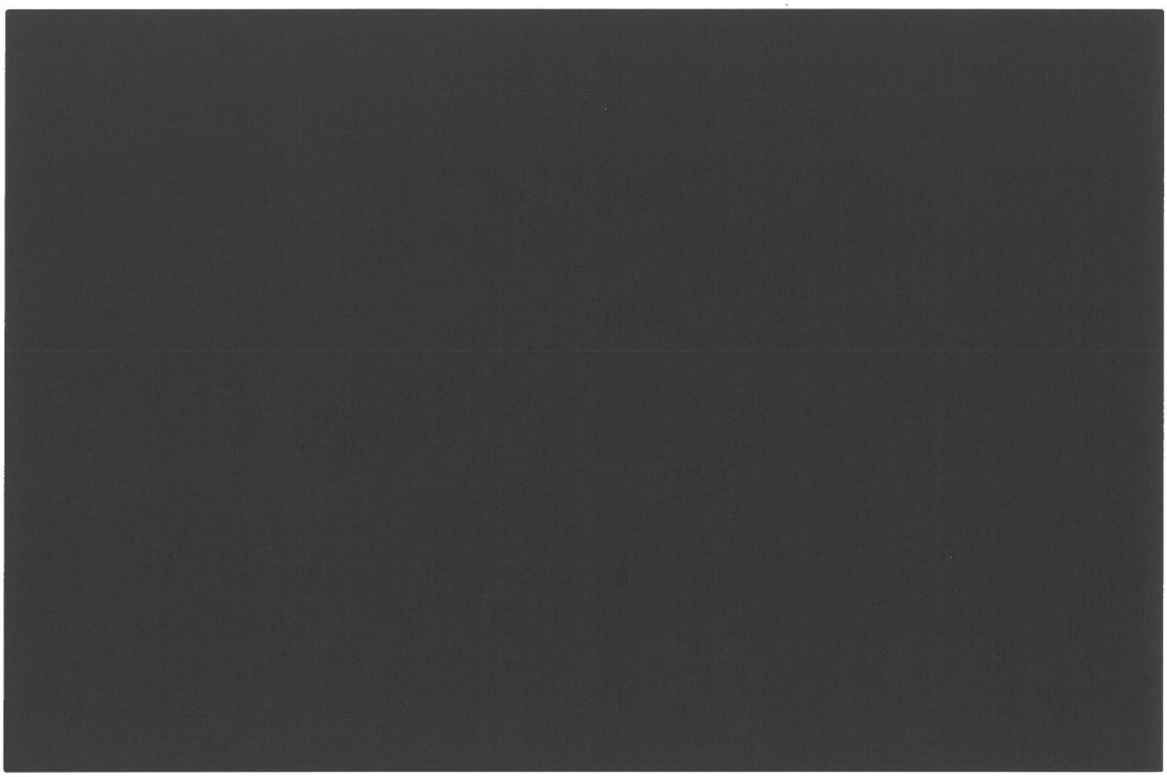
bullx B515

- 1 SandyBridge CPU Board (SCB) – dual socket
- 2common large fans
- 1 SATA hard disk drive 2.5"

SCB

- 2 Intel Sandy Bridge EP E5-2470, 8c/2.3GHz/20Mo/8GT/s
- 64 GB ECC SDRAM (4 x 16 GB DDR3 DIMM 1600 MHz)
- 1 x 300 GB SATA 2,5" 7,2 kRPM HDD
- 1 x Intel Xeon Phi / NVIDIA M2090

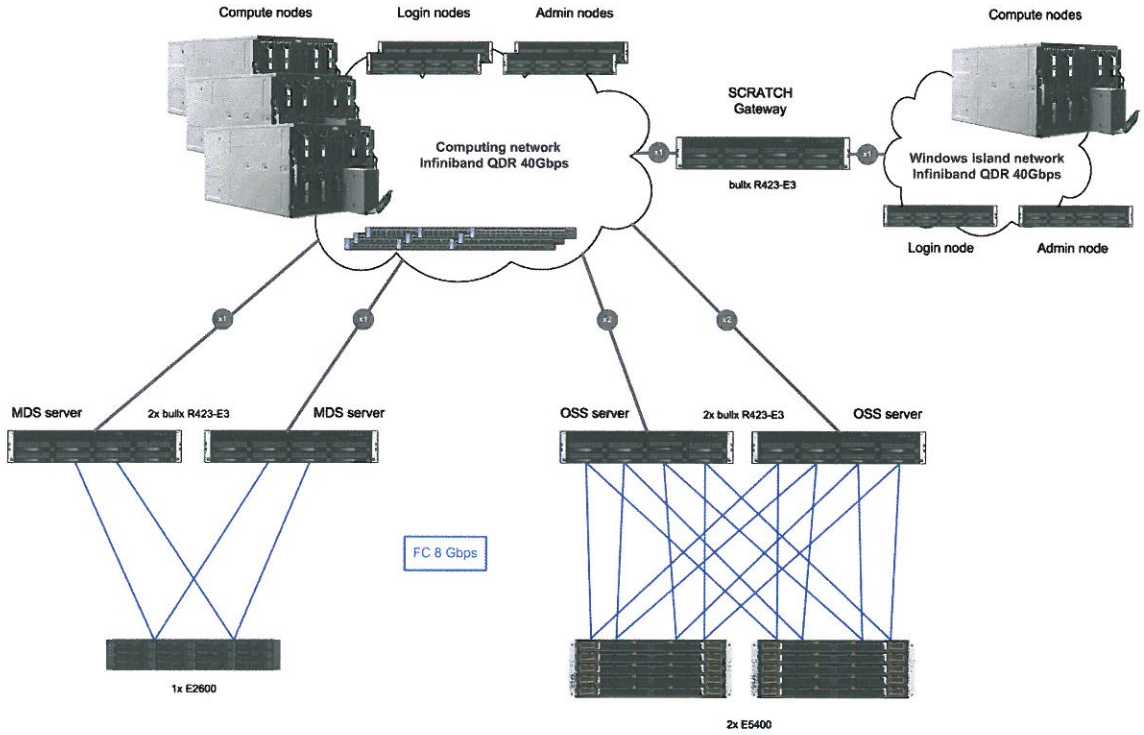
- 1 x ethernet port 1 Gbits/s
- 1 single port IB HCA, QDR support
- 1 integrated BMC
- 1 Dual port 1Gb Ethernet controller



Obr. 4 - blokové schéma bullx B515 SCB

1.3.2. DATOVÁ ÚLOŽIŠTĚ

1.3.2.1. ÚLOŽIŠTĚ SCRATCH



Obr. 5 - blokové schéma úložiště SCRATCH

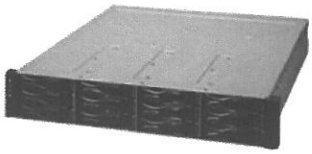
2x MDS server

bullx R423-E3

bullx R423-E3

- 2 Intel Sandy Bridge EP E5-2650, 8c/2.0GHz/20Mo/8GT/s
- 64 GB ECC SDRAM (8 x 8 GB DDR3 DIMM 1600 MHz)
- 2 x 300 GB SATA 3,5" 15 kRPM HDD (RAID1)
- 2 x ethernet port 1 Gbits/s
- 1 ConnectX-2 single port 4x QDR
- 1 LPe12002 dual port 8Gbps FC
- 1 integrated BMC
- 2U

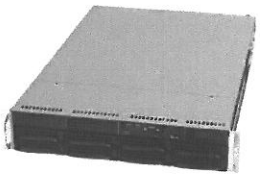
1x MDS Storage



NetApp CDE2680-12

NetApp E2600
- 1 CDE2680-12 enclosure, 2x RAID controller
- 4 GB cache, 8 8Gbps FC host ports, 4 6 Gbps SAS backend ports
- 12 300GB SAS 3.5" 15,7kRPM HDD 2xRAID5(4+1) + 2HS
- Redundant power supplies
- 2U

2x OSS server



bullx R423-E3

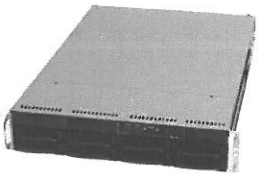
bullx R423-E3
- 2 Intel Sandy Bridge EP E5-2650, 8c/2.0GHz/20Mo/8GT/s
- 32 GB ECC SDRAM (8 x 4 GB DDR3 DIMM 1600 MHz)
- 2 x 300 GB SATA 3,5" 15 kRPM HDD (RAID1)
- 2 x ethernet port 1 Gbits/s
- 2 ConnectX-2 single port 4x QDR
- 4 LPe12002 dual port 8Gbps FC
- 1 integrated BMC
- 2U

2x OSS Storage

NetApp CDE5400-60

NetApp E5400
- CDE5400-60 enclosure, 2x RAID controller
- 24 GB cache, 8 8Gbps FC host ports, 4 6 Gbps SAS backend ports
- 53 2TB NL-SAS 3.5" 7,2kRPM HDD 5xRAID6(8+2) + 3HS
- Redundant power supplies
- 4U

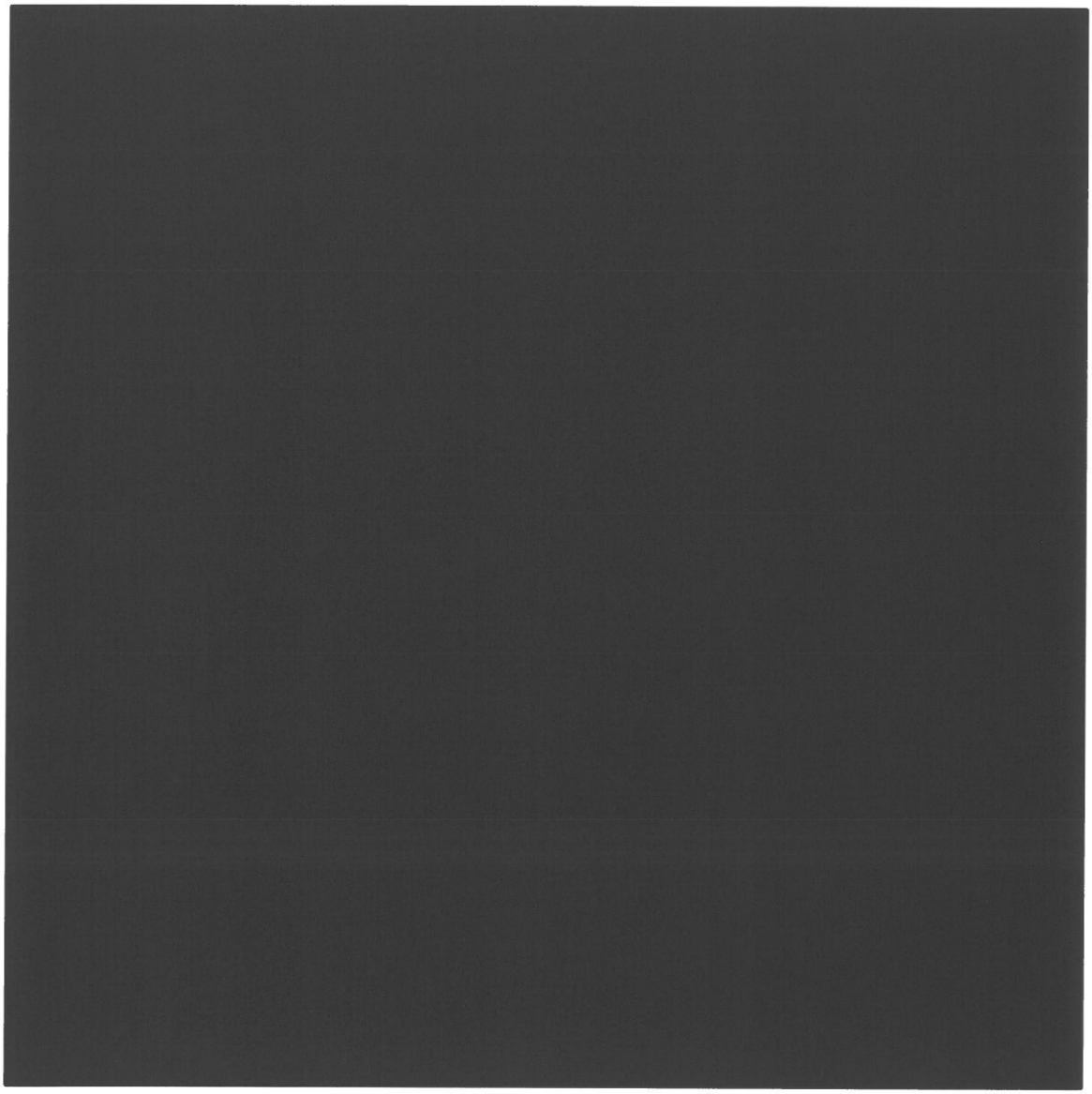
1x SCRATCH gateway server



bullx R423-E3

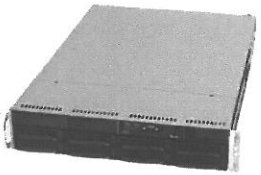
bullx R423-E3
- 2 Intel Sandy Bridge EP E5-2670, 8c/2.6GHz/20Mo/8GT/s
- 32 GB ECC SDRAM (8 x 4 GB DDR3 DIMM 1600 MHz)
- 2 x 300 GB SATA 3,5" 15 kRPM HDD (RAID1)
- 2 x ethernet port 1 Gbits/s
- 2 ConnectX-2 single port 4x QDR
- 1 integrated BMC
- 2U

1.3.2.2. ÚLOŽIŠTĚ HOME



Obr. 6 - Blokové schéma úložiště HOME

2x HOME NFS server



bullx R423-E3

bullx R423-E3
- 2 Intel Sandy Bridge EP E5-2650, 8c/2.0GHz/20Mo/8GT/s
- 64 GB ECC SDRAM (8 x 8 GB DDR3 DIMM 1600 MHz)
- 2 x 300 GB SATA 3,5" 15 kRPM HDD (RAID1)
- 2 x ethernet port 1 Gbits/s
- 1 ConnectX-2 single port 4x QDR
- 2 LPe12002 dual port 8Gbps FC
- 1 integrated BMC
- 2U

1x HOME storage

NetApp CDE5400-60
NetApp DE66000

NetApp E5400
- CDE5400-60, 3 DE6600 eclosures, 2x RAID controller
- 24 GB cache, 8 8Gbps FC host ports, 4 6 Gbps SAS backend ports
- 227 2TB NL-SAS 3.5" 7,2kRPM HDD 22xRAID6(8+2) + 7HS
- Redundant power supplies
- 16U

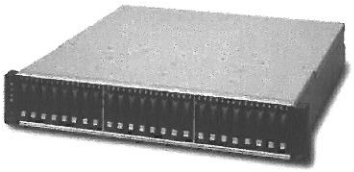
1x HOME gateway

bullx R423-E3

bullx R423-E3
- 2 Intel Sandy Bridge EP E5-2650, 8c/2.0GHz/20Mo/8GT/s
- 64 GB ECC SDRAM (16 x 4 GB DDR3 DIMM 1600 MHz)
- 4 x 300 GB SATA 3,5" 15 kRPM HDD (RAID1)
- 2 x ethernet port 1 Gbits/s
- 1 10G single port ethernet SFP+
- 2 ConnectX-2 single port 4x QDR
- 1 integrated BMC
- 2U

1.3.2.3. ÚLOŽIŠTĚ S BLOKOVÝM PŘÍSTUPEM

1x BLOCK storage



NetApp CDE2680-24

NetApp DE66000

NetApp E2600
- 1 CDE2680-24, 1DE6600 enclosures, 2x RAID controller
- 4 GB cache, 8 8Gbps FC host ports, 4 6 Gbps SAS backend ports
- 73 900GB SAS 2.5" 10kRPM HDD 7xRAID6(8+2) + 3HS
- Redundant power supplies
- 6U

1.3.3. ZÁLOHOVÁNÍ DAT

Funkci zálohování a automatické archivace bude poskytovat software Calypso.
Systém je navržen tak, aby pokryl následující oblasti:

- Zálohování klíčových serverů, dat a konfigurací řešení systému Malý cluster
- Zálohování souborového datového úložiště HOME – rozsah 100TiB
- Zálohování dalších serverů zadavatele implementovaných v systému Malý cluster

Calypso Qsnap umožňuje snapshot na úrovni filesystému, je tedy možné vytvářet snapshoty jednotlivých souborů.

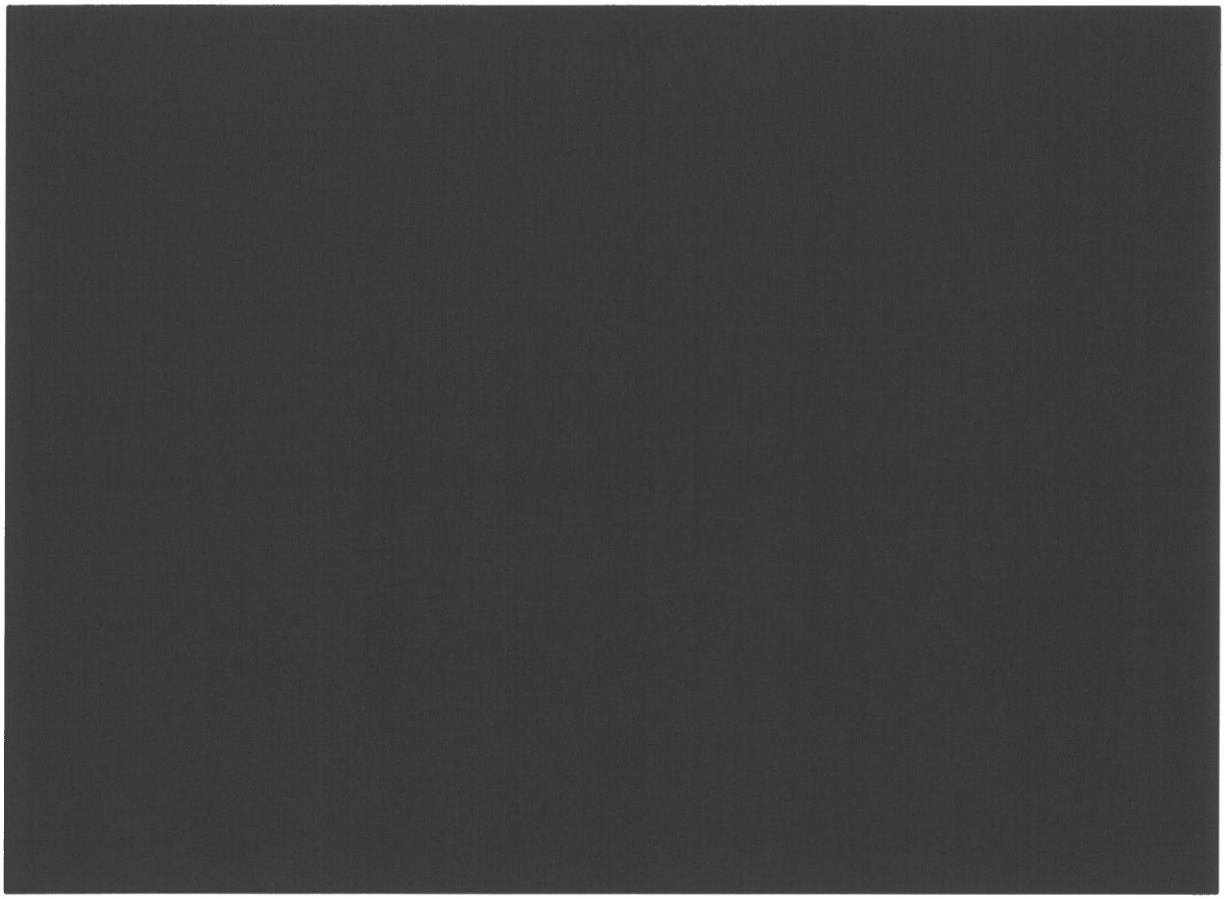
[redacted]

[redacted]

[redacted]

Software pro zálohování a obnovu dat splňuje následující základní vlastnosti:

a. zálohování a obnova v prostředí Linux a Windows
b. podpora knihoven fyzických i virtuálních, zálohování na disk
c. podpora SAN připojených mechanik
d. integrace se zvolenou serverovou virtualizací pro vysoký výkon a efektivitu
e. granulární obnova individuálních souborů a složek
f. obnova vlastníků, práv a atributů souborů a složek
g. možnost obnovit zálohu fyzického stroje do virtuálního prostředí



Obr. 7 - blokové schéma systému pro zálohování dat

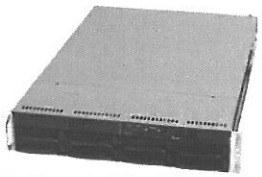
Licence SW Calypso pokrývají: 4 klienty windows pro obslužnou infrastrukturu + zvlášť 30 požadovaných klientů windows, 14 linuxových klientů pro obslužnou infrastrukturu + zvlášť 30 požadovaných klientů linux, 1 klient PostgreSQL pro zálohu administračního úložiště clusterDB, 1 klient MS-SQL pro zálohu VMWare vSphere vCenter databáze, 12 TB deduplikovaného prostoru pro D2D2T funkcionalitu, 10 zálohovacích mechanik.

Calypso ComServ server

bullx R423-E3

bullx R423-E3
- 2 Intel Sandy Bridge EP E5-2650, 8c/2.0GHz/20Mo/8GT/s
- 8 GB ECC SDRAM (2 x 4 GB DDR3 DIMM 1600 MHz)
- 2 x 500 GB SATA 3,5" 7,2 kRPM HDD (RAID1)
- 2 x ethernet port 1 Gbits/s
- 1 integrated BMC
- 2U

Calypso Media server



bullx R423-E3

bullx R423-E3
- 2 Intel Sandy Bridge EP E5-2650, 8c/2.0GHz/20Mo/8GT/s
- 32 GB ECC SDRAM (8 x 4 GB DDR3 DIMM 1600 MHz)
- 2 x 500 GB SATA 3,5" 7,2 kRPM HDD (RAID1)
- 4 x 2TB SATA 3,5" 7,2 kRPM HDD
- 2 x ethernet port 1 Gbits/s
- 2 LPe1250 single port 8Gbps FC

- 1 ConnectX-2 single port 4x QDR
- 1 integrated BMC
- 2U

1.3.4. PÁSKOVÁ KNIHOVNA

Pásková knihovna

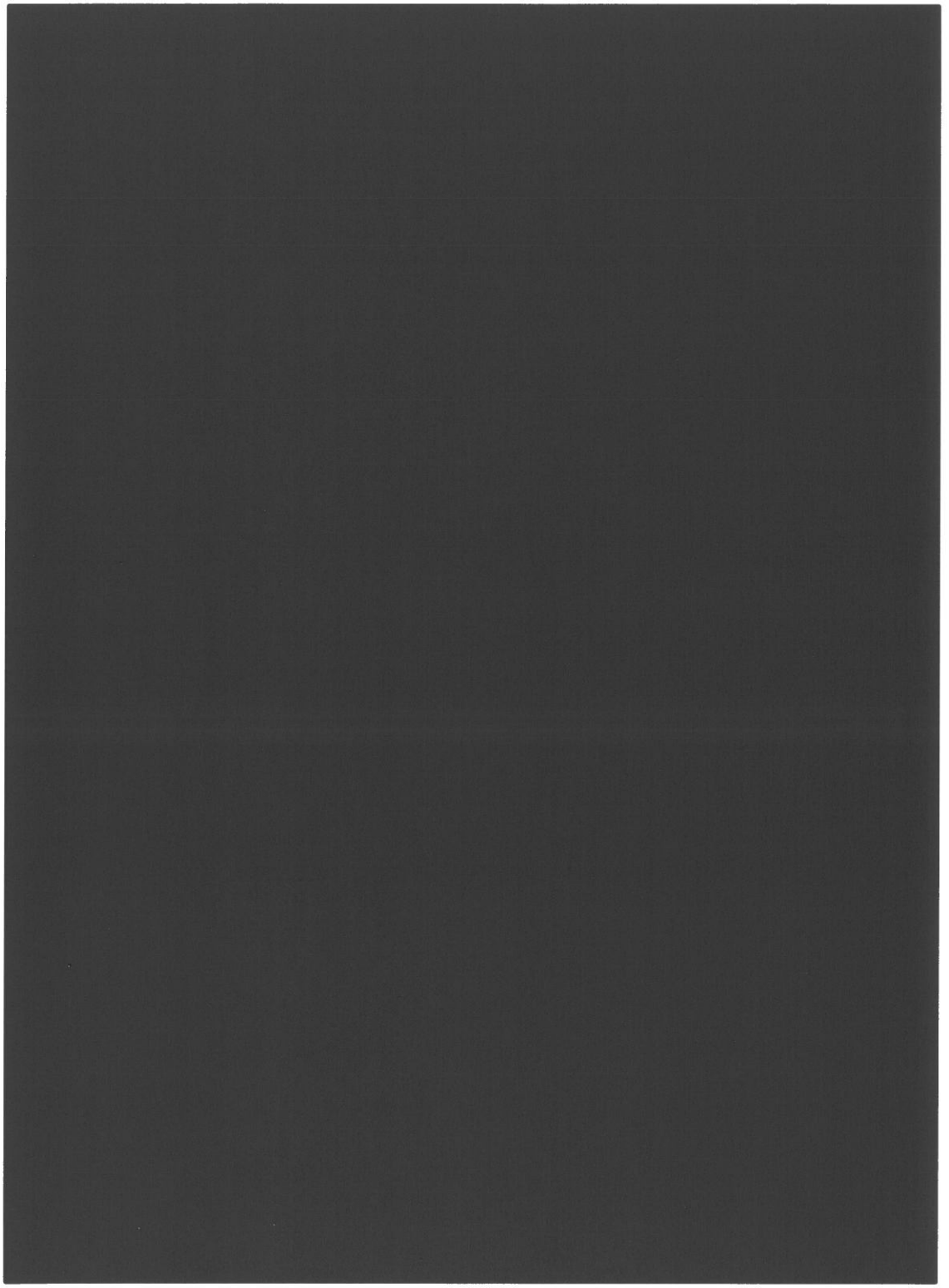
Quantum Scalar i500

Quantum Scalar i500
- 363 slots
- 10 drives LTO-5 , FC 8GB
- 370 LTO5 and 15 cleaning cartridges
- 544,5 TB uncompressed capacity
- 5,04TB/h backup rate
- 41U

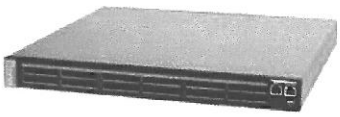
Kapacita zálohovacích médií je 555 TB

1.3.5. SÍTĚ

1.3.5.1. VÝPOČETNÍ SÍŤ



Obr. 8 - blokové schéma výpočetní sítě



Mellanox IS5030Q

Mellanox IS5030Q

- managed QDR 40 Gbps InfiniBand switch
- 36 QSFP 4x QDR ports
- Mellanox InfiniScale IV architecture
- 2,88 Tbps switching capacity
- port-to-port latency <100ns
- IBTA 1.21 compliant
- Quality of Service enforcement
- 9 Virtual lanes: 8 data + 1 management
- Adaptive routing, Congestion control, Port mirroring
- 48K entry linear forwarding data base
- Redundant power supplies
- 1U

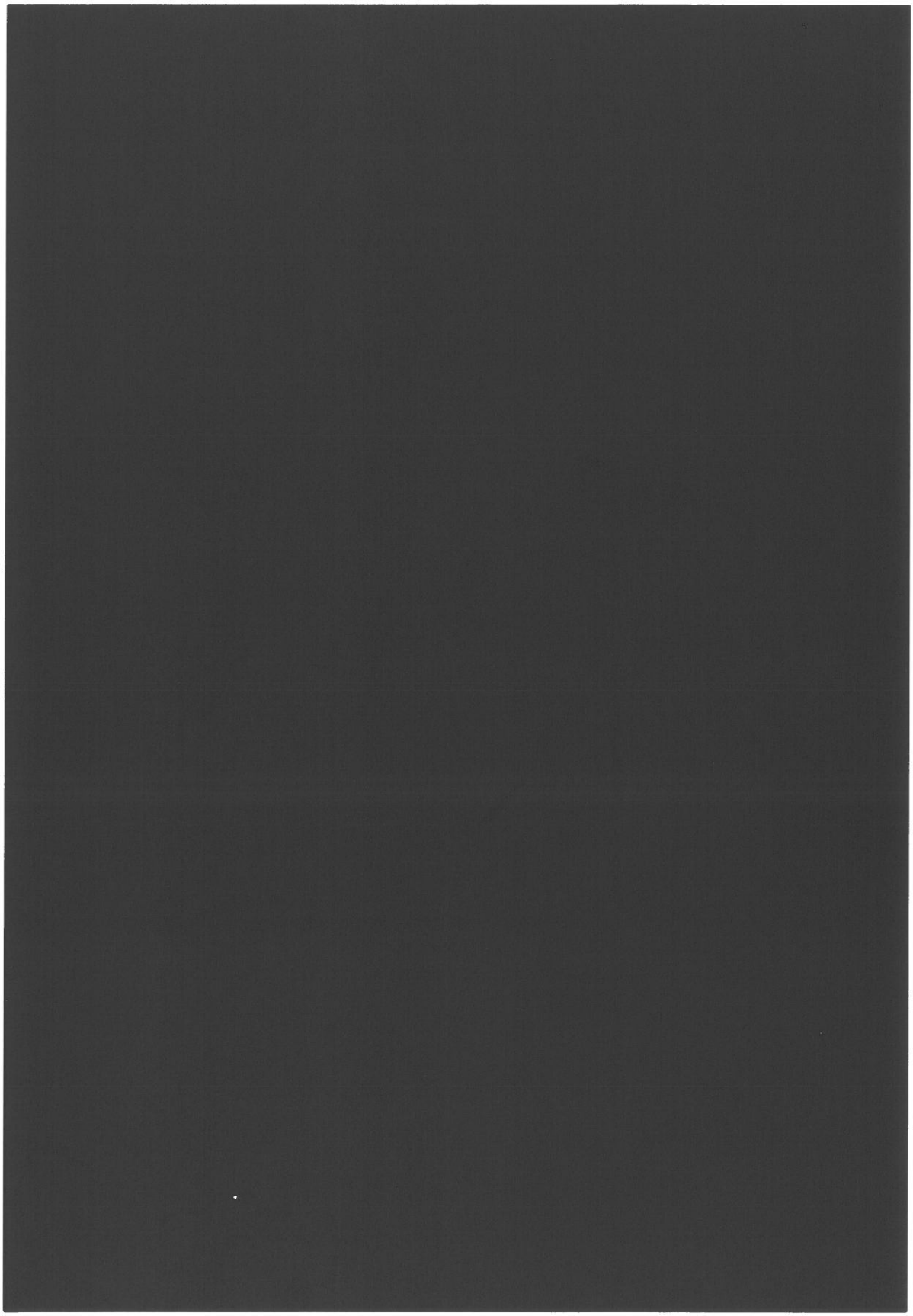
1.3.5.2. SERVISNÍ SÍŤOVÁ INFRASTRUKTURA A WAN KONEKTIVITA

Servisní síťová infrastruktura je tvořena 3 stohovanými switchi Cisco C3750 pro hlavní část SNV a samostatným switchem Cisco C3750 pro část systému vyčleněnou pro provoz prostředí operačního systému Windows, tzv. subcluster.

Uživatelský síťový provoz v rámci „LAN" sítě je z důvodu bezpečnosti oddělen od administračního provozu pomocí 802.1Q tagged VLAN.

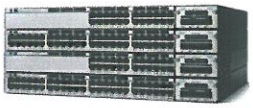
Administrační síťový provoz probíhá v defaultní netagované síti VLAN ID 1 což vylučuje možné problémy s konfigurací 802.1Q VLAN na servisních portech propojených infrastruktur.

Pro potřeby out-of-band managementu je k dispozici „IPMI" síť tvořená samostatným switchem Cisco C3750, do něhož jsou zapojeny IPMI management porty všech serverů řešení, umožňující vzdálený přístup ke konzolím a ovládání napájení. Tato síť je rovněž používána pro zajištění vysoké dostupnosti administračních serverů.



Obr. 9 - Blokové schéma zapojení LAN, Administrační a IPMI sítě

3x Cisco C3750-48, stacked, 10gigE uplinks, LAN/Admin



Cisco C3750 Series

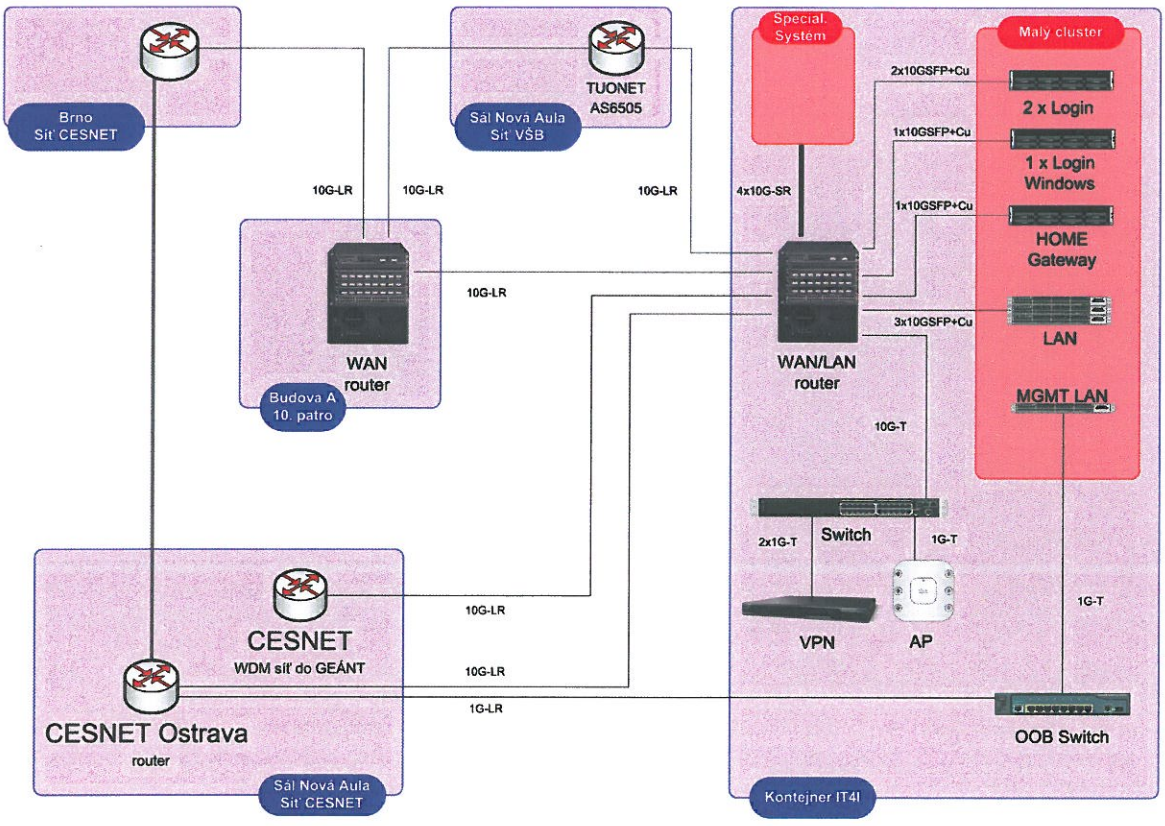
CISCO Catalyst C3570X-48TS
- 48-port 1Gbps switch
- 2x 10Gbps Ethernet SFP/ SFP+ uplinks
- CISCO Stackwise stacking
- 1U/switch

Cisco C3750-24 Windows LAN/Admin

CISCO Catalyst C3570X-24TS
- 48-port 1Gbps switch
- 1U

Cisco C3750-48 IPMI síť

CISCO Catalyst C3570X-48TS
- 48-port 1Gbps switch
- 1U



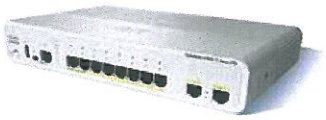
Obr. 10 - blokové schéma zapojení WAN

WAN/LAN směrovač (4 plně osazené sloty, 5U)

Cisco C6504E

Cisco Catalyst 6504-E
- Sup2T (VS-S2T-10G) supervisor, 2GB RAM
- 2 8-port 10GE karty s podporou TrustSec
- redundantní AC zdroje 2.7kW
- 5U

Přepínač pro OOB mgmt. (out of band management)



Cisco C3560-C

Cisco Catalyst 3560-C
- 8 1Gbps ports, 2 SFP uplinks
- 1U

WiFi AP

Cisco Aironet 3502I-E-K9

Cisco Aironet 3502I
- 802.11a/g/n
- PoE powered

VPN koncentrátor

Cisco ASA5550

Cisco ASA 5550
- 5000 user VPN licenses
- up to 1.2 Gbps throughput
- 1U

Přepínač „Switch"

Cisco Catalyst C2960S-24TD-L
- 24 1Gbps ports, 2 10Gbps SFP+ uplinks
- 1U

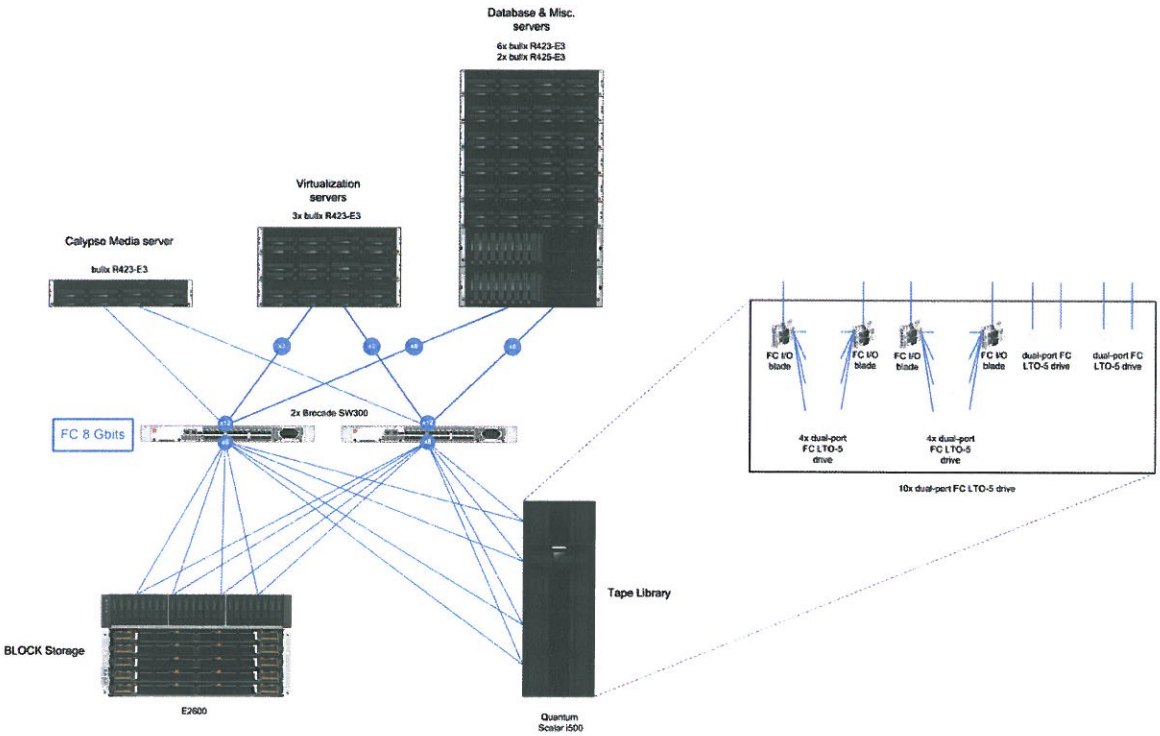
Cisco C2960S

WAN směrovač

Cisco C6504E

Cisco Catalyst 6504-E
- Sup2T (VS-S2T-10G) supervisor, 2GB RAM
- 8-port 10GE karta s podporou TrustSec
- redundantní AC zdroje 2.7kW
- 5U

1.3.5.3. SAN INFRASTRUKTURA



Obr. 11 - blokové schéma SAN sítě

2x Brocade 300 SAN switch

Brocade B300

Brocade B300
- 24 8Gbps Fibre channel ports, licensed
- 192 Gbps bandwidth
- 700ns latency cut-through switching
-1U

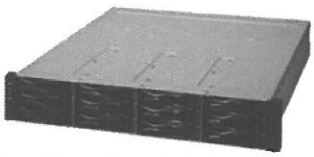
1.3.6. SERVISNÍ A ADMINISTRAČNÍ SERVEROVÁ INFRASTRUKTURA

2x Admin server

bullx R423-E3

bullx R423-E3
- 2 Intel Sandy Bridge EP E5-2650, 8c/2.0GHz/20Mo/8GT/s
- 32 GB ECC SDRAM (8 x 4 GB DDR3 DIMM 1600 MHz)
- 2 x 300GB SAS 3,5"15kRPM HDD (RAID1)
- 2 x ethernet port 1 Gbits/s
- 1 LPe12002 dual port 8Gbps FC
- 1 ConnectX-2 single port 4x QDR
- 1 integrated BMC
- 2U

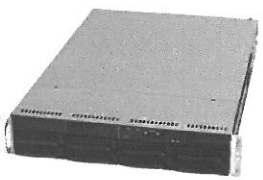
Admin storage



NetApp CDE2680-12

NetApp E2600
- 1 CDE2680-12 enclosure, 2x RAID controller
- 4 GB cache, 8 8Gbps FC host ports, 4 6 Gbps SAS backend ports
- 6 450GB SAS 3.5" 15,7kRPM HDD RAID5(4+1) + 1HS
- Redundant power supplies
- 2U

2x Login server

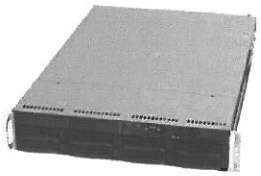


bullx R423-E3

bullx R423-E3
- 2 Intel Sandy Bridge EP E5-2670, 8c/2.6GHz/20Mo/8GT/s
- 128 GB ECC SDRAM (16 x 8 GB DDR3 DIMM 1600 MHz)
- 2 x 300GB SAS 3,5"15kRPM HDD (RAID1)
- 2 x ethernet port 1 Gbits/s
- 1 10G single port ethernet SFP+

- 1 ConnectX-2 single port 4x QDR
- 1 integrated BMC
- 2U

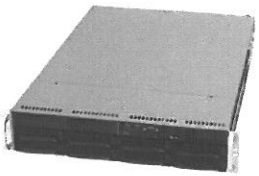
Windows Admin server



bullx R423-E3

bullx R423-E3
- 2 Intel Sandy Bridge EP E5-2650, 8c/2.0GHz/20Mo/8GT/s
- 32 GB ECC SDRAM (8 x 4 GB DDR3 DIMM 1600 MHz)
- 6 x 300GB SAS 3,5"15kRPM HDD (RAID1)
- 2 x ethernet port 1 Gbits/s
- 1 ConnectX-2 single port 4x QDR
- 1 integrated BMC
- 2U

Windows Login server

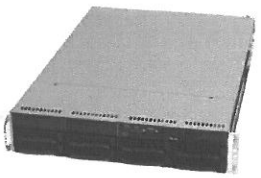


bullx R423-E3

bullx R423-E3
- 2 Intel Sandy Bridge EP E5-2670, 8c/2.6GHz/20Mo/8GT/s
- 128 GB ECC SDRAM (16 x 8 GB DDR3 DIMM 1600 MHz)
- 2 x 300GB SAS 3,5"15kRPM HDD (RAID1)
- 2 x ethernet port 1 Gbits/s
- 1 10G single port ethernet SFP+
- 1 ConnectX-2 single port 4x QDR
- 1 integrated BMC
- 2U

1.3.7. DATABÁZOVÉ A JINÉ SERVERY

2x DB server typ A



bullx R423-E3

bullx R423-E3
- 2 Intel Sandy Bridge EP E5-2665, 8c/2.4GHz/20Mo/8GT/s
- 64 GB ECC SDRAM (16 x 4 GB DDR3 DIMM 1600 MHz)
- 2 x 300GB SAS 3,5"15kRPM HDD (RAID1)
- 4 x ethernet port 1 Gbits/s
- 2 LPe1250 single port 8Gbps FC
- 1 ConnectX-2 single port 4x QDR
- 1 integrated BMC
- 2U

2x DB server typ B

bullx R423-E3

bullx R423-E3
- 2 Intel Sandy Bridge EP E5-2665, 8c/2.4GHz/20Mo/8GT/s
- 96 GB ECC SDRAM (12 x 8 GB DDR3 DIMM 1600 MHz)
- 2 x 300GB SAS 3,5"15kRPM HDD (RAID1)
- 4 x ethernet port 1 Gbits/s
- 2 LPe1250 single port 8Gbps FC
- 1 ConnectX-2 single port 4x QDR
- 1 integrated BMC
- 2U

2x DB server typ C

bullx R423-E3

bullx R423-E3
- 2 Intel Sandy Bridge EP E5-2665, 8c/2.4GHz/20Mo/8GT/s
- 512 GB ECC SDRAM (16 x 32 GB DDR3 DIMM 1600 MHz)
- 2 x 300GB SAS 3,5"15kRPM HDD (RAID1)
- 2 x 128GB SATA 2,5" SSD
- SAS2/SATA3 RAID controller 512MB cache, BBU
- 4 x ethernet port 1 Gbits/s
- 2 LPe1250 single port 8Gbps FC
- 1 ConnectX-2 single port 4x QDR
- 1 integrated BMC
- 2U

2x DB server typ D

bullx R423-E3

bullx R425-E3
- 2 Intel Sandy Bridge EP E5-2670, 8c/2.6GHz/20Mo/8GT/s
- 64 GB ECC SDRAM (16 x 4 GB DDR3 DIMM 1600 MHz)
- 2 x 300GB SAS 3,5"15kRPM HDD (RAID1)
-1 NVIDIA QUadro 4000 GPU
- 4 x ethernet port 1 Gbits/s
- 2 LPe1250 single port 8Gbps FC
- 1 ConnectX-2 single port 4x QDR
- 1 integrated BMC
- 4U

1.3.8. VIRTUALIZAČNÍ INFRASTRUKTURA

3x Virtualization server

bullx R423-E3

bullx R423-E3
- 2 Intel Sandy Bridge EP E5-2665, 8c/2.4GHz/20Mo/8GT/s
- 128 GB ECC SDRAM (16 x 8 GB DDR3 DIMM 1600 MHz)
- 2 x 500GB SATA 3,5"7,2kRPM HDD (RAID1)
- 4 x ethernet port 1 Gbits/s
- 2 LPe1250 single port 8Gbps FC
- 1 ConnectX-2 single port 4x QDR
- 1 integrated BMC
- 2U

Virtualization management server (vSphere vCenter)

bullx R423-E3

bullx R423-E3
- 2 Intel Sandy Bridge EP E5-2620, 8c/2.0GHz/15Mo/8GT/s
- 8 GB ECC SDRAM (2 x 4 GB DDR3 DIMM 1600 MHz)
- 2 x 500GB SATA 3,5"7,2kRPM HDD (RAID1)
- 2 x ethernet port 1 Gbits/s
-- 1 integrated BMC
- 2U

1.3.9. NÁVRH SOFTWAROVÝCH ŘEŠENÍ

Součástí nabídky jsou tyto licence:

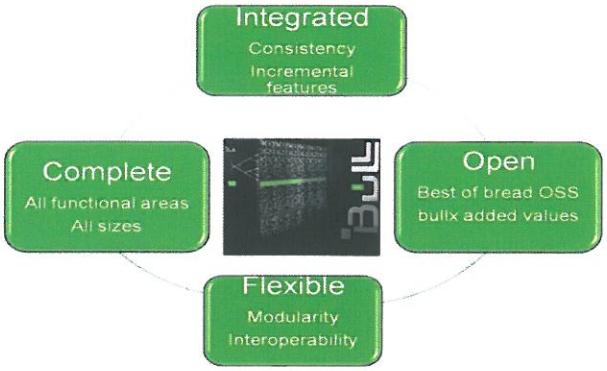
bullx Linux – permanent license and 3y support	436
bullx scs Advanced Edition AE2.x 1 to 144 sockets – 3Y support per socket , permanent license	144
bullx scs Advanced Edition AE2.x 145 to 432 sockets – 3Y support per socket,permanent license	288
bullx Parallel File System - permanent license , 3Y support per OSS	4
Calypso Backup SW - permanent license and 3y support	1
Job Scheduler PBS-Pro <433 Sockets - 3Y support	430
Academic VMware vSphere 5 Enterprise for 1 processor and 64GB RAM + 3y tech. support	6
Academic VMware vCenter Server 5 Standard for vSphere 5 + 3y tech.support	1
OS Microsoft® Windows Server® 2008 R2 x64 Standard Edition, 5 users English	2

bullx SCS - Supercomputer Suite

bullx supercomputer suite je kompletní sada specializovaného HPC software, který pomáhá zákazníkům implementovat, spravovat a plně provozovat jejich HPC clustery jednoduchým, spolehlivým a a efektivním způsobem. Suita se zaměřuje na potřeby správy superpočítače, aplikací a dat.

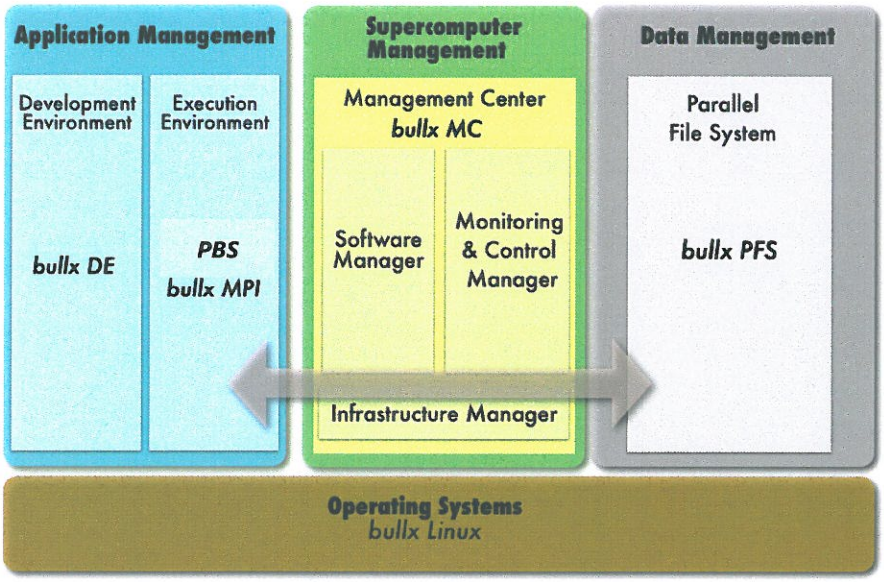
Suita nástrojů je plně otevřená, jelikož je založena na tom nejlepším open source software a používá předních open standards vylepšených o další vlastnosti představující přidanou hodnotu společnosti Bull.

Jednotlivé komponenty jsou integrovány dohromady za použití centrálního úložiště informací o výpočetním systému, tzv. Supercomputer Information Repository, které umožňuje centralizovaný a konzistentní pohled na superpočítač, a stav na něm provozovaných aplikací.



Comprehensive, Open, Integrated and Flexible solution

bullx supercomputer suite je navržen modulárně. Uživatelé mají plnou volnost k použití dalších



bullx supercomputer suite

předních softwarových řešení třetích stran souběžně s bullx SCS. Takováto řešení jsou neustále testována a validována pro použití s bullx SCS.

1.3.9.1. STRUKTURA PRODUKTU

Supercomputer management

bullx Management Center poskytuje administrátorům bohatou a výkonnou sadu nástrojů pro nasazení a provoz jejich HPC clusterů efektivním způsobem. Skládá se ze tří složek:

- **Infrastructure Manager**
  Na rozdíl od jiných seskupení poskytovatelů HPC Bull používá sofistikované a automatizované nástroje pro budování HPC clusterů. Návrh clusteru a infrastruktury zahrnuje optimalizovanou a značenou strukturovanou kabeláž, bere v úvahu fyzické vlastnosti datového centra a efektivitu napájení a chlazení. **IM** poskytuje základní funkce pro nastavení a konfiguraci clusteru a jeho infrastruktury:

  - Instalace IM s přednastavenou konfigurací umožňuje implementaci rozsáhlých výpočetních clusterů účinným a spolehlivým způsobem, což také výrazně zlepšuje jejich udržovatelnost.
    Pre-load konfigurace umožňuje identifikaci jednotlivých zařízení včetně jejich fyzického umístění podle informací v databázi clusterDB.

    Dostupnost této pre-load informace poskytuje administrátorům možnost kontroly souladu mezi nastavenou konfigurací v clusterDB a aktuálním stavem clusteru zjišťovaným pomocí automatického mechanismu.

  - Automatické vyhledávání a konfigurace většiny zařízení a síťových nastavení. Složitější konfigurace sítě (např. stohování, VLAN), lze také nastavit poloautomaticky. Na základě informací předem instalované v clusterDB IM automaticky konfiguruje všechna zařízení a sítě tak, aby byl cluster schopen provozu. I když instalace je založena na pre-load informaci, MAC adresy zařízení nemusí být zadávány (ruční zadávání adres MAC je velmi těžkopádné a náchylné k chybám) Mechanismus

automatického zjišťování shromáždí MAC adresy všech připojených zařízení (za předpokladu, že switche umožňují správu).

- Automatické zjišťování nově přidaného vybavení (switche, výpočetní uzly) při rozšiřování clusteru. V případě výměny nebo rozšíření zařízení, může správce provádět cílenou automatickou identifikaci a konfiguraci nových zařízení, a aktualizovat clusterDB o jejich nové informace.

- **Software Manager**

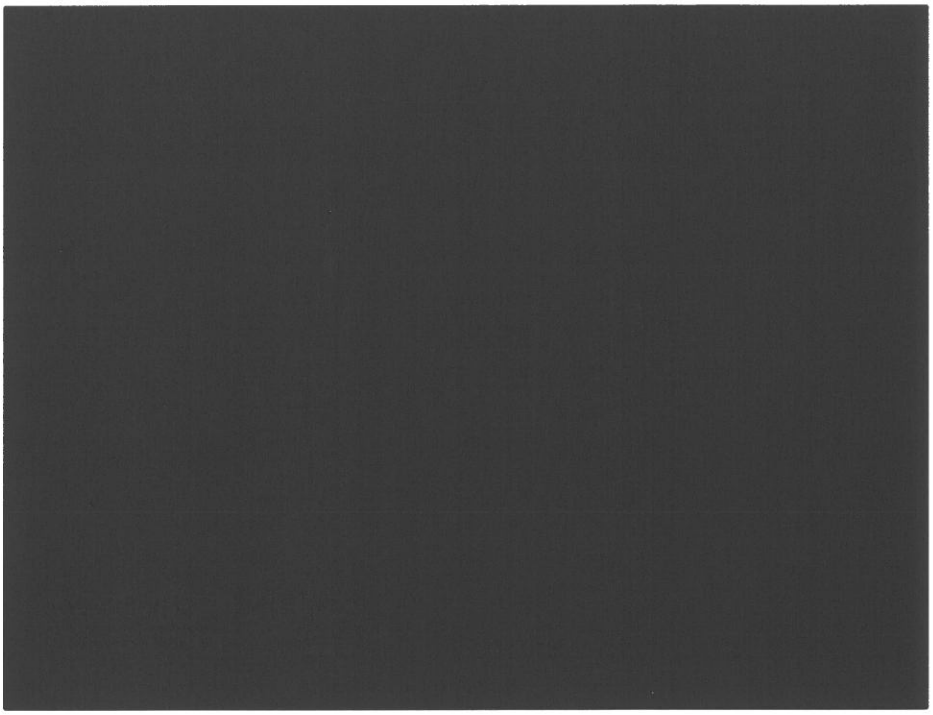
SM je zodpovědný za instalaci softwaru, nasazení, konfigurace a aktualizace. Software Management je důležitým úkolem při správě velmi velkých clusterů. Jako takový vyžaduje vysokou úroveň automatizace procesů nezbytných k zajištění spolehlivosti a integrity clusteru z hlediska instalovaného software. Kromě toho, škálovatelnost a výkon operací se softwarem je rozhodující pro snížení doby potřebné k nasazení do provozu a zvýšení využití clusteru. Hlavní funkce SM jsou:

- Správa SW profilů uzlů pomocí předkonfigurovaných nastavení. Každý uzel clusteru má jednu nebo více rolí. Role je funkce, kterou uzel v clusteru plní (např. servisní, výpočetní, I/O uzel, apod.). Role určuje soubor služeb, které poběží na daném uzlu Profil určuje seznam balíčků, které budou na uzlu instalovány. K dispozici je sada předdefinovaných profilů (např. servisní, výpočetní, akcelerovaný server, Lustre OSS, Lustre MDS, atd.). Pro každou roli je definován profil.

  Výběr balíčků spojený s profilem je pečlivě navržen tak, aby obsahoval pouze potřebné funkce a služby pro provoz každého uzlu bez zvýšené režie. Nicméně administrátoři mohou i měnit a přizpůsobovat přednastavení jak uznají za vhodné.

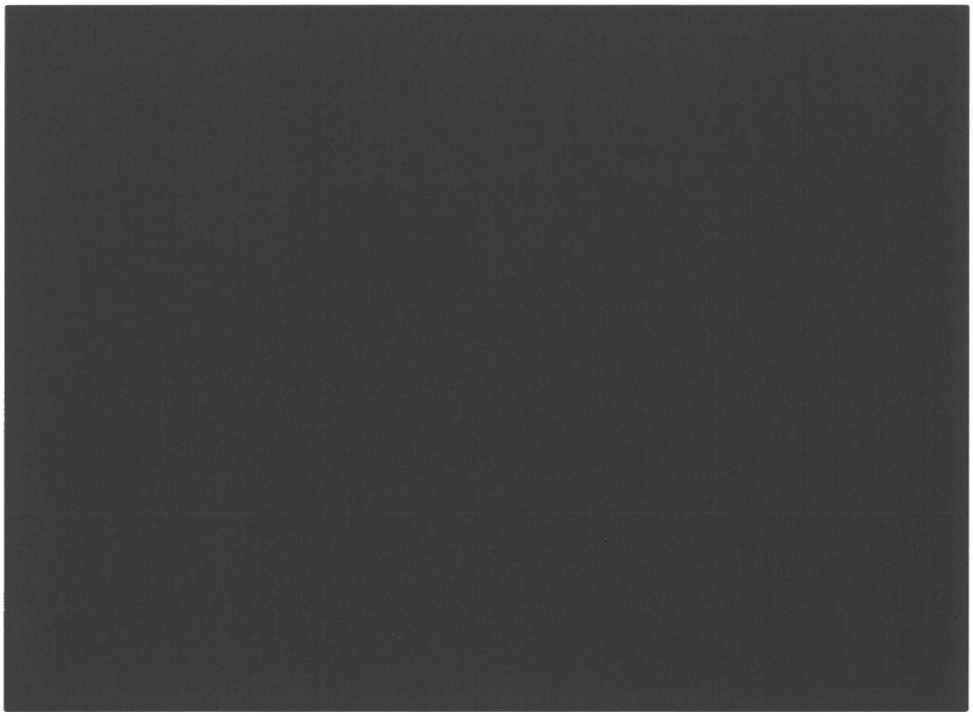
  SRI nabízí multi-OS, paralelní, profilově orientovanou automatizovanou instalaci. Tento nástroj pomáhá správcům clusteru s instalací softwaru, umožňuje vzdáleně instalovat referenční uzly. Referenční uzel je uzel, který se používá pro tvorbu konkrétního profilu (výpočetní uzel, akcelerovaný výpočetní uzel, uzel Lustre, atd.).

  Tuto instalaci je pak možné replikovat (pomocí image provisioning) na všechny uzly, které mají plnit stejnou roli. Instalační balíčky na referenčním uzlu jsou automaticky voleny podle jeho profilu. Poskytuje také možnost paralelního provádění instalačních skriptů na cílových uzlech. Pro mechanismy instalace se používá proprietární nástroj Yabix – Yet Another Bull Installation tool for eXtreme computing, který využívá YUM repositories.



- Image based provisioning pro kompletní a flexibilní správu systémových obrazů. SM obsahuje kompletní sadu funkcí, které umožňují tvorbu, úpravu a mazání systém image, aplikaci patchů a upgrade SW komponent. Tyto obrazy se používají k nasazení uzlů clusteru v nastavovací fázi a během pozdějších aktualizací (zavádění nové aplikace atd.) v běžném provozu. Systémové obrazy jsou šířeny a nasazovány paralelně na cílové uzly (viz níže) a umožňují rozdílové nasazení image (Delta, přenáší se pouze rozdíl mezi oběma systémovými obrazy). Využívá se proprietární nástroj ksis.

- Rychlý, spolehlivý a škálovatelný mechanismus šíření systém image. Vlastní směrovací a přenosový protokol je implementován pro dosažení rychlé a spolehlivé propagace systémových image. Je založen na myšlence vytvoření cesty o stromové topologii přes všechny uzly. Image je potom rozdělena na části, která se potom šíří „pipeline" mechanismem touto cestou do všech uzlů. Tímto způsoben je dosaženo hromadného nasazení systémových image v O (log n) čase.

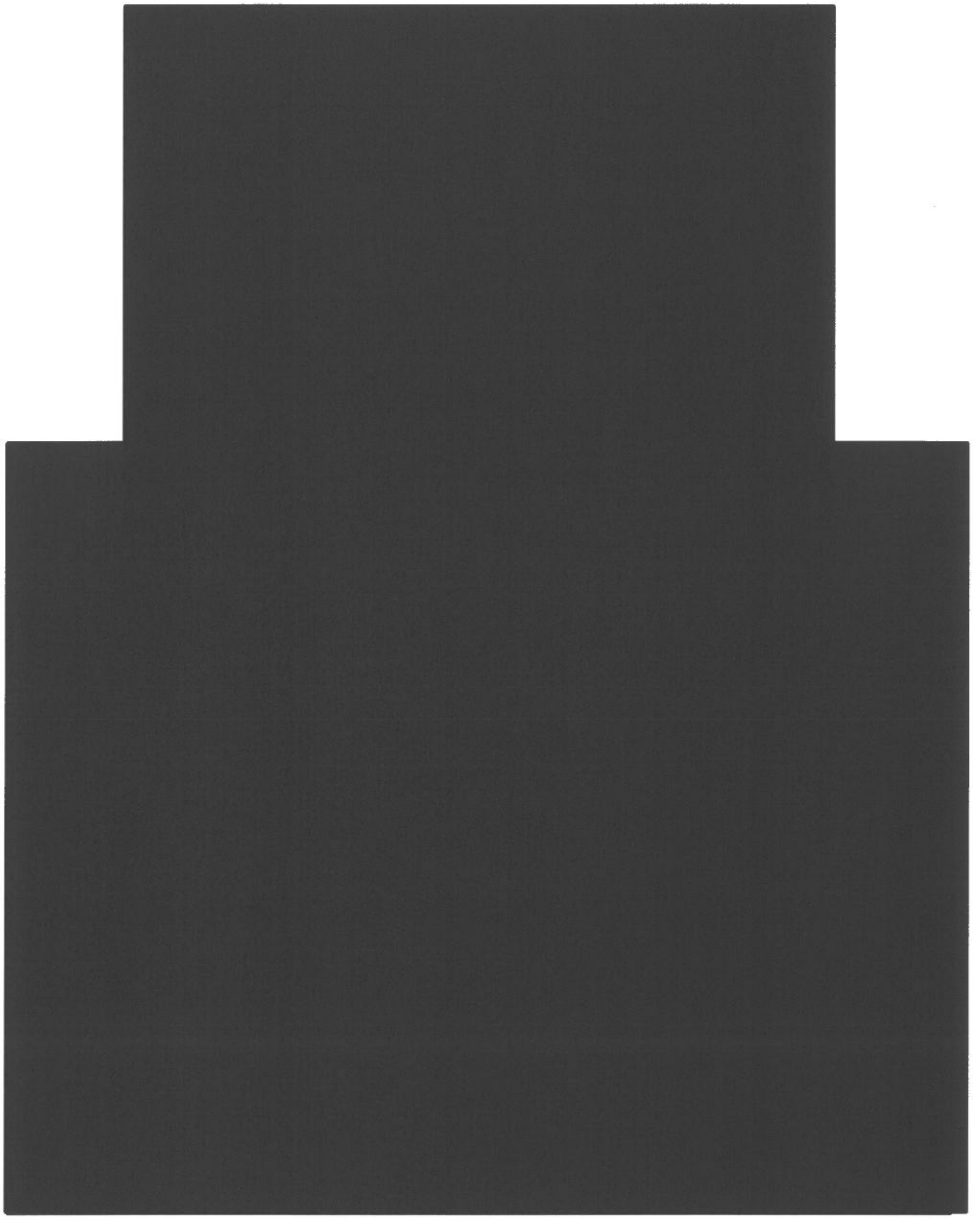


- Kompletní klient/server nástroj pro správu konfigurací. Na rozdíl od ad-hoc a ruční aktualizace softwaru a konfigurace, poskytuje automatizační nástroj 2 zásadní výhody, které jsou kritické pro administraci velkých clusterů: a) systematické logování a verzování aktualizací konfigurace, b) bezpečnost (zabezpečení SSL certifikáty, autentizace klientů) a spolehlivost procesu aktualizace (tj. správce je informován o tom, zda operace aktualizace uspěje nebo selže). Používá se proprietární nástroj kconf který je založen na paralelním open source automatizačním nástroji puppet.

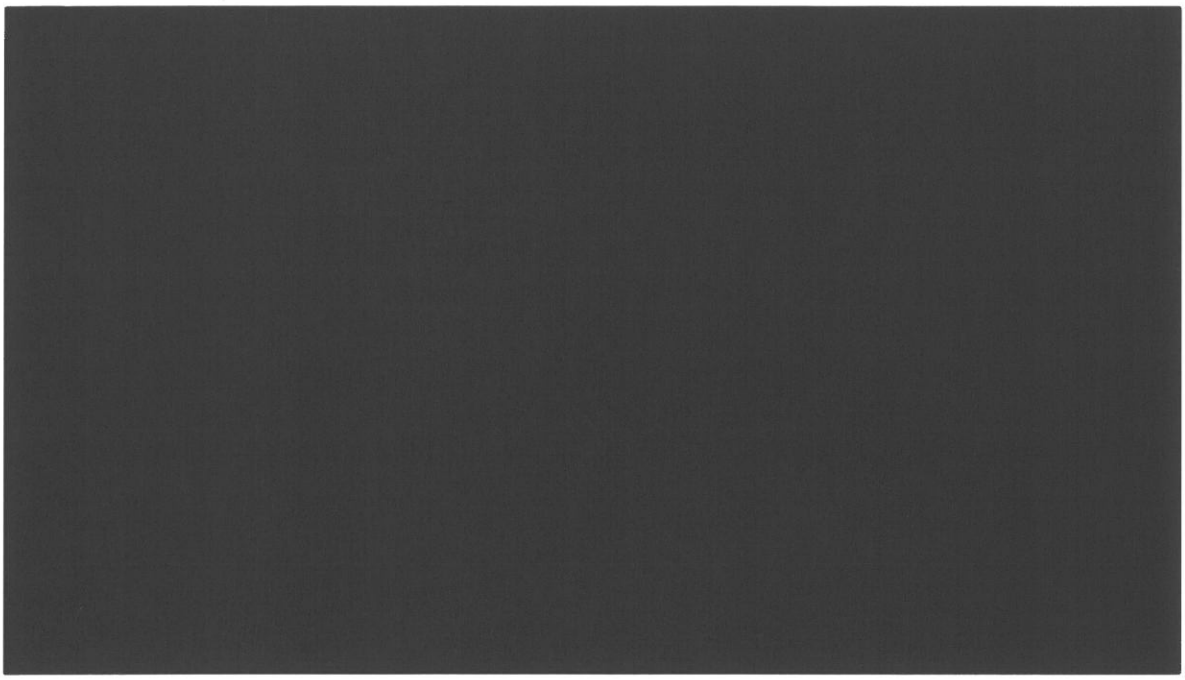
- Globální správa bezpečnostních nastavení clusteru.

- **Monitoring & Control Manager**

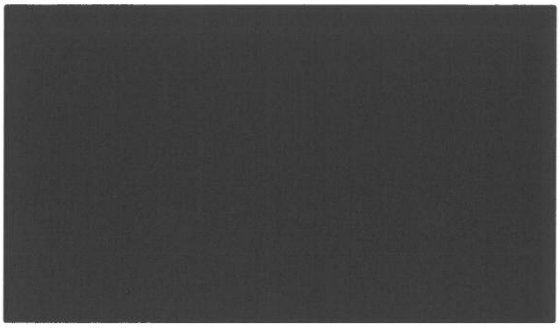
  Cílem MCM je poskytnout centralizované a aktuální zobrazení stavu běžícího clusteru spolu s nástroji pro škálovatelné provádění operací pro správu. MCM monitoring je založen na produktu BSM – Bull System Manager, jedná se o vysoce kustomizovaný Nagios s rozšířením Ganglia.



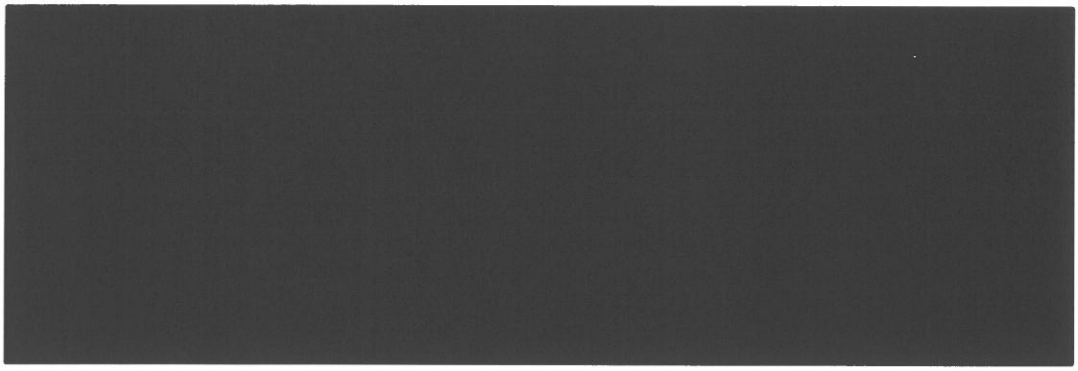
- Centralizované webové rozhraní s pokročilým grafickým rozhraním pro dozor clusteru. Monitoring GUI poskytuje grafický způsob sledování všech druhů zařízení (např. uzly, switche, napájecí zdroje, atd.). Obsahuje předdefinovanou sadu služeb, jako je měření teploty, měření spotřeby energie, momentální stav systému, atd., které mohou být sledovány pro každý typ zařízení v nastavených pravidelných intervalech. Pro každou službu, jsou různé úrovně prahu (např. critical, warning, OK, apod.) definovány podle měřené hodnoty. Stav jednotlivých služeb je uveden graficky (barevný systém), takže administrátor může mít současně i globální a přesnou představu o celkovém zdraví clusteru. Administrátor může definovat, podle potřeby, nové služby spojené s konkrétními zařízeními nebo jejich skupinami (např. podle typu zařízení).



- Víceúrovňový dohled
  GUI umožňuje administrátorovi a globální pohled na cluster, nicméně správce může přiblížit na konkrétní části clusteru (např. rack, blade šasi, atd.) a až na konkrétní zařízení (např. blade, napájecí zdroj, atd.). Podobné informace (např. o stavu služby, hodnoty metrik, logy, atd.), jsou zobrazeny konzistentním způsobem pro všechny různé druhy zařízení.



- Monitoring a vizualizace metrik, které lze v MCM definovat, pro zadanou související službu. Specifikace služby zahrnují, mimo jiné, i skript, který sbírá hodnoty metrik, stejně jako interval, který se používá pro sběr dat (spouštění skriptu), k dispozici jsou opět sady předdefinovaných služeb a metrik, ale lze definovat i vlastní. Zvláštní nástroj se používá k zobrazení vývoje metrických hodnot pomocí grafů a křivek.



- Rule-based engine pro správu chyb a závad systému, který provádí automatickou detekci sad událostí pro preventivní údržbu. Engine umožňuje monitoring údajů pro každé zařízení v clusteru. Některé z těchto údajů jsou zobrazovány v GUI, ale celkové množství shromažďované informace je velmi velké, přesahující možnosti lidské kontroly. Tato nová technologie umožňuje správci definovat sady pravidel, která se aplikují na údaje shromážděné z monitoringu. Engine takto umožňuje detekci vzorů chování systému (např. chyby paměti), která spouští předdefinované akce jako např. zaslání zprávy na centrum podpory. Engine je založen na nástroji SEC- Same Event Correlator.

- Rule-based engine pro detekci vzorů chování pro řízení spotřeby elektrické energie. Na základě vzorů chování důležitých ukazatelů, jako je teplota nebo spotřeba, jejichž hodnoty jsou periodicky shromažďovány, jsou spouštěny administrátorem definované akce jako vypnutí, snížení frekvence, přechod do úsporného režimu, zasláním e-mailu s hlášením apod. K dispozici je rovněž sada předdefinovaných pravidel pro správu napájení. Engine je založen na nástroji SEC- Same Event Correlator.

- Pro sběr systémových logů se využívá syslog-ng.

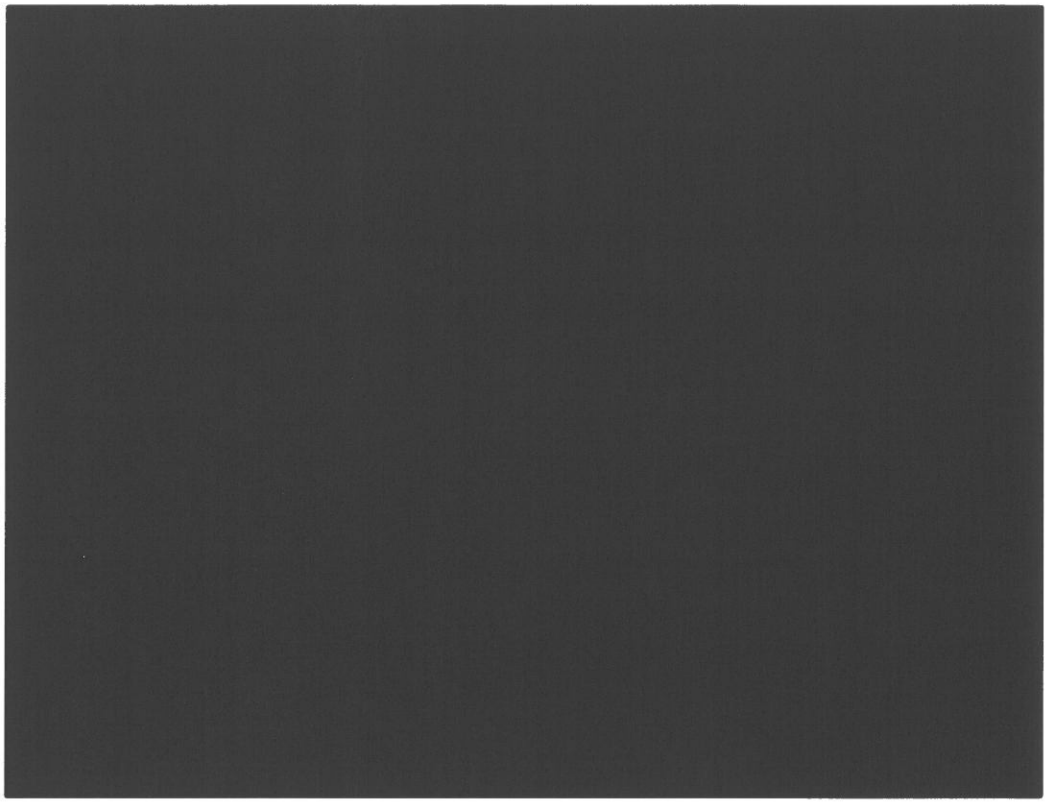
- Sofistikovaný a komplexní nástroj pro správu, který poskytuje prostředky pro realizaci a měření SLA.

- Nástroj pro paralelní spouštění příkazů na clusteru.

- Vzdálený přístup k systémové konzoli napříč celým clusterem.

- **clusterDB:**
  Za účelem dosažení maximální flexibility a rozšiřitelnosti, si jednotlivé součásti systému vyměňují informace prostřednictvím centrálního databázového úložiště zvaného clusterDB. Jak statické informace o konfiguraci hardwaru a softwaru clusteru, tak i dynamické a provozní informace jsou uloženy v clusterDB. Kromě toho, správci mají přístup k tomuto úložišti, což jim umožňuje využívat

bohaté informace uložené v clusterDB k vybudování vlastních administrátorských nástrojů a řešení. ClusterDB je založen na open source databázi PostgreSQL.



Správa aplikací

- **bullx MPI**

bullx MPI je knihovna založená na OpenMPI, obohacená o řadu funkcí, které vylepšují OpenMPI ve třech směrech: a) masivní zvýšení škálovatelnosti, výkonu a spolehlivosti aplikací využívajících MPI, b) poskytuje těsnou integraci s dalšími složkami bullx supercomputer suite c) zvyšuje robustnost produktů s více reaktivními procesy pro ladění a opravu chyb a silnější procesy validace pro nově integrované funkce.
Tato vylepšení jsou implementována způsobem, který zaručuje kompatibilitu s OpenMPI. Jakýkoliv program, který běží s OpenMPI bude fungovat se stejnou nebo lepší úrovní výkonu s bullx MPI.
Hlavní rysy bullx MPI jsou:

- Kompatibilní s MPI 2.1. MPI je stále se vyvíjející standard.
- Afinita procesů s jemnou granularitou technologie umisťování procesů prostřednictvím integrace se správci úloh a výpočetních prostředků (LSF, PBSpro a Slurm).

- Kernel-based data mover (k dispozici pouze pro bullx Linux) který optimalizuje MPI komunikaci v rámci jednoho uzlu. Modul jádra MDM byl vyvinut jako zero-copy mechanismus, z něhož

profitují procesy běžící na stejném uzlu díky optimálnímu využívání šířky pásma MPI přístupu do paměti.

- Vyladění náročných skupinových operací (alltoall, allreduce, atd.). bullx MPI používá pro jejich implementaci kromě toho v rámci Open MPI standartního i vlastní framework pro kolektivní algoritmy, který se zaměřuje na škálovatelnost kolektivních operací pro systémy řádu peta-scale, a který využívá informace o konkrétní topologii.

- Vyladění MPI-IO operací (k dispozici pro bullx PFS). Použitý paralelní filesystem Lustre pracuje s těsnou vazbou na MPI-IO. To umožňuje aplikacím benefitovat z optimalizovaného mapování mezi paralelní MPI I/O funkcemi a jejich přístupem k paralelnímu souborovému systému.

- MPI Multipathing a failover.
  MPI není původně navženo jako řešení odolné proti závadám. To nutí uživatele k vývoji zavádění dodatečných nástrojů, aby ochránili běh svých aplikací a předcházeli hodinovým ztrátám využití strojového času při manuálních reakcích na závady. bullx MPI umožňuje běhu aplikací přežít pád primární komunikační sítě automatickým transparentním restartem komunikace po sekundární, nebo po nouzové síti pokud je k dispozici.

- Detekce vzorů abnormálního chování MPI
  Během životního cyklu MPI aplikace hlídá bullx MPI komunikace mezi procesy a jejich partnery. Při zamrznutí komunikace v důsledku chyb a dead-locků pak aktivně ukončuje stagnující komunikace. To umožňuje uvolňovat komunikující procesy, které pak přecházejí do stavů snížené spotřeby, což vede k úspoře energie.

- Detekce špatného využívání propojovací sítě.
  bullx MPI zavádí řadu monitorovacích mechanismů. Tento monitoring s nízkou režií umožňuje detekci situací, kdy jsou parametry propojení sítě nevyhovující pro její správné využití aplikacemi. Tyto informace jsou poskytovány prostřednictvím různých mechanismů (stderr, syslog ...).

- bullx MPI profilovací analyzátor, nástroje pro kontrolu MPI a diagnostiku (součástí bullx DE).
  bullx MPI obsahuje profilovací knihovny, které lze linkovat s aplikacemi, a které pak poskytují různé metriky MPI a komunikační matice. Záměrem tohoto mechanismu je detekce porušování standardu MPI (špatná MPI sémantika), neobvyklého chování (špatné chování MPI) a dead-locků (chyby MPI programování). Diagnostické nástroje jsou nezbytné pro interpretaci dat z defektního běhu aplikace. bullx Management Center může v reakci na takováto zjištění víceúrovňově automaticky spouštět skripty které mohou konkrétní situace řešit. Získané informace rovněž významnou měrou usnadňují uživateli ladění vlastního běhu aplikace.
- Knihovna bullx MPI je validována pro použití s paralelními debuggery třetích stran TotalView a Allinea DDT

- **PBS Pro Job Manager**

Výkonný a časem prověřený správce úloh PBS (Portable Batch System) Professional (PBS Pro™) funguje na principu SOA – Service Oriented Architecture. Je používán na více než 1400 instalacích po celém světě. Je schopen efektivně řídit výpočetní zátěže superpočítačových clusterů, SMP systémů a hybridních systémů, škáluje do tisíců procesorů.

Správa výpočetní zátěže
- špičková škálovatelnost, pokročilé flexibilní fronty
- komplexní jazyk pro odesílání úloh
- job arrays
- suspend/resume úloh
- checkpoint/restart na aplikační úrovni
- checkpoint/restart na úrovni OS
- závislost a zřetězení úloh
- automatický staging souborů
- komplexní logy pro accounting
- pokročilé bezpečnostní funkce a ACL
- správa dávkových i interaktivních úloh
- konzistentní uživatelské a správcovské rozhraní
- v souladu s POSIX standardem pro dávky

Pokročilé plánovací algoritmy
- plánování podle využití prostředků
- optimalizované možnosti backfillingu
- plánování na základě znalosti topologie
- virtualizace uzlů
- pokročilé umísťování úloh
- maximalizace využití HW a aplikačních licencí
- preemptivní plánování pro okamžité spouštění prioritních úloh
- plánování s politikou fair-share
- rezervace prostředků
- řízení více superpočítačů současně
- znalost síťové topologie

Spolehlivost odolnost škálovatelnost
- automatický failover plánovacího serveru
- autonomní monitoring systémů
- automatické obnovení úloh
- pokročilá správa procesů díky těsnější integraci s MPI knihovnami
- automatické disaster recovery do vzdálených lokací
- zaručení exkluzivního přístupu k uzlům
- ověřená správa více jak 500,000úloh denně
- nasazeno na několika superpočítačíchs více jak 10,000 CPU jádry

Data Management

- **bullx PFS**

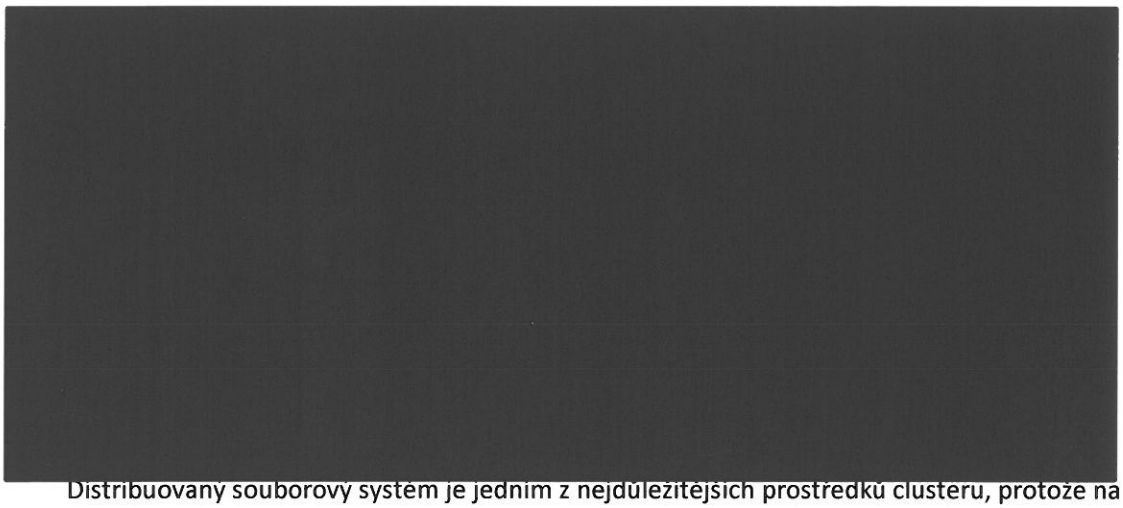
bullx PFS se opírá o Lustre Distributed File System rozšířený o řadu funkcí, které zlepšují Lustre ve třech směrech: a) zvyšují škálovatelnost, výkon a spolehlivost Lustre, b) poskytují těsnou integraci s dalšími složkami bullx Supercomputing Suite c) zvýšují robustnost systému s více reaktivními procesy pro opravy chyb a silnější validaci pro nově integrované funkce.

Hlavní rysy bullx PFS, jsou:

- Ladění pro zvýšení škálovatelnosti a výkonu
  Tuning se týká jednak jádra OS, jednak parametrů bullx PFS. Požadované úpravy linuxového jádra jsou integrovány do bullx Linux. OS Tři hlavní funkce jsou ovlivněny: Umisťování a lokalita procesů Lustre, parametry I/O operací (kompletní I/O stack, včetně ovladačů disků a Fibre Channel), a multipathingu.
- Funkce vysoké dostupnosti (I/O cells)
  bullx PFS poskytuje potřebné nástroje pro konfiguraci serveru Lustre odolnou proti chybám, tzv. I/O cells tvořené až 4-mi uzly. Každé cílové zařízení úložiště lze v případě závady transparentně zpřístupnit přes alternativní server v rámci I/O cell. To výrazně omezuje vliv selhání serveru, protože jeho pracovní zátěž může být automaticky rozložena na zbývající servery.
- Předem připravený profil pro komponenty Lustre v bullx Management Center pro snadné nasazení a konfiguraci prostředí Lustre. Odpovídající speciálně upravená instalace bullx Linux se v rámci profilu automaticky nainstaluje na příslušné Lustre servery.

- Integrace centrálního nástroje pro správu Lustre „Shine", na jehož vývoji se společnost Bull podílí, s bullx Supercomputing Suite, umožňuje využití předdefinované konfigurace souborového systému, monitoring provozu filesystémů a sdílení této informace přes centrální úložiště clusterDB, a centrální správu souborových systémů.

- Konkrétní monitorovací funkce pro Lustre jsou k dispozici přímo v bullx Management Center.



Distribuovaný souborový systém je jedním z nejdůležitějších prostředků clusteru, protože na něm závisí činnost každého uzlu. Z tohoto důvodu je provoz Lustre pečlivě sledován dvěma nezávislými způsoby:

- Každý klient Lustre je pravidelně proaktivně kontrolován zda je schopen komunikovat se servery. Tato kontrola nijak dodatečně nezatěžuje souborové systémy.
- Pravidelně je spouštěn test zápis/čtení/porovnání na 2 náhodně vybraných uzlech pokrývající všechny systémy souborů a všechna fyzická úložiště.

Tento sofistikovaný monitorovací mechanismus zajišťuje, že distribuovaný souborový systém je stále pod pečlivou kontrolou, a že všechny události, jež mohou mít vliv na celý cluster, jsou odhaleny včas.

Současná verze Lustre použitá v bullx PFS je verze 2.0. V Q3 2012, bullx PFS bude postaveno na Lustre 2.1 a bude schopno přímého propojení s Hierarchical Storage Management (HSM) řešením s podporou rozšířených atributů pro HSM funkce.

Komponenty a principy Lustre

Lustre filesystem odpovídá normě POSIX

Konfigurace lustre zahrnuje:

- 1 nebo 2 MetaData servery (MDS)
- 1 nebo více Object Storage Serverů (OSS) spojených do Object Storage Targets (OST)

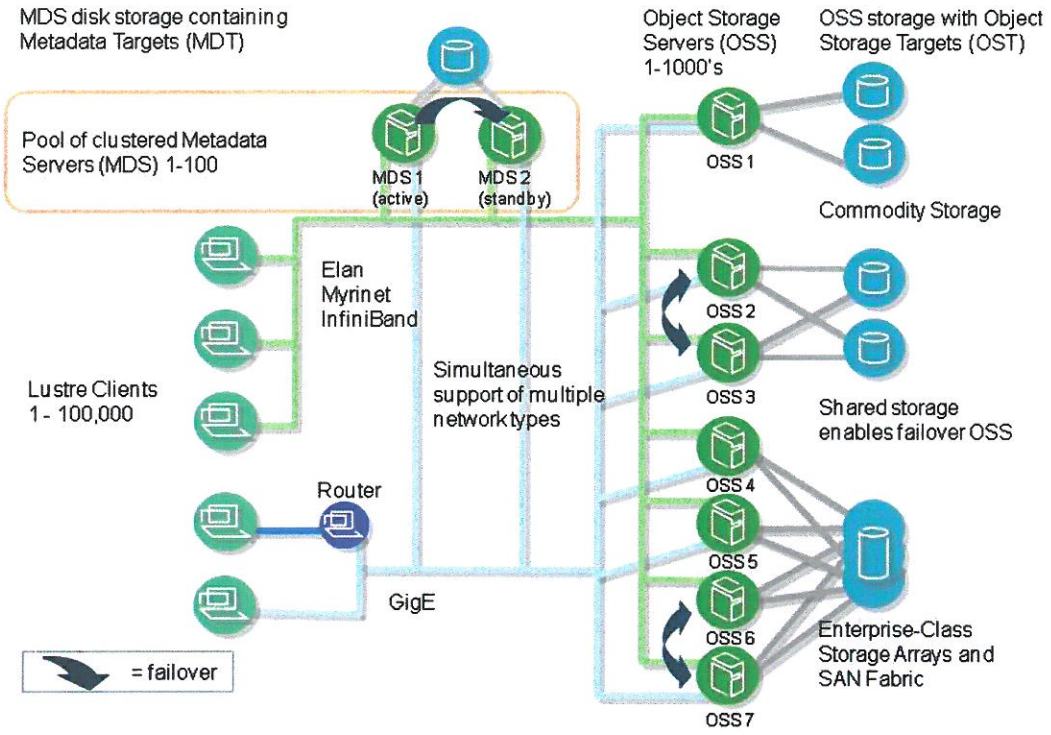
MDS server ukládá metadata ma jeden nebo více Metadata Targets (MDTs). Každý MDT ukládá metadata (inodes/attributes = owner, filename, directory, permission, lustre stripping layout). Každý soubor využívá jeden MDT, který může být k dispozici pro více MDS, přičemž ho používá vždy jen jeden MDS (primary MDS). V případě havárie primárního MDS, přebírá jeden ze zbývajících MDS roli primárního MDS. Obecně má každá instance Lustre jeden primary MDS a jeden failover MDS. Oba musí mít přístup ke společným MDT, což vyžaduje přístup ke sdílenému úložišti.

OSS poskytuje I/O služby pro ukládání a čtení souborů, a komunikuje po síti s jedním nebo vice Object Storage Target (OST). Každý OSS nemusí mít přístup ke všem OST, ale pouze k těm OST, které řídí. Pro zabezpečení provozu, jeden OSS může být nastaven jako primární pro příslušný OST se záložními OSS. V takovém případě musí být příslušný OST zpřístupněn všem takovým OSS.

OST ukládá souborová data (chunks) jako datové objekty na jeden nebo více OSS. OST je tzv.logical unit volume formátovaný jako ext3 nebo ext4 filesystem .Jediný Lustre filesystem může mít vice OST, každý sloužící podmnožině dat souborů kdy vztah mezi souborem a OST není nezbytně 1:1. Kvůli optimalizaci výkonu může být soubor rozprostřen přes mnoho OST pomocí mechanismu file strippingu který je řízen pomocí Logical Object Volume (LOV). Maximální velikost OST závisí jak na použitém úložném RAID subsystému (LUNy E5400), jednak na operačním systému (bullx Linux), ale take na omezeních použitého filesystému ext3 nebo ext4. V současné době je to buď 8 nebo16 TB.

Lustre klienty jsou výpočetní uzly. Ty mohou přistupovat k lustre filesystému skrz různé heterogenní site jako Infiniband (FDR, QDR, DDR), TCP (GigE, 10 GigE) nebo jiné proprietární sítě (Cray, Elan, Myrinet) skrz komponentu LND (Lustre Networking Driver) API LNET (Lustre networking). LNET podporuje protokol RDMA pokud ji podporuje použitá podkladová síť (např. IB).

Jeden globální Namespace: Klienti, kteří mountují lustre filesystem, vidí jediný koherentní sdílený jmenný protstor. Více klientů může zapisovat najednou do různých částí jediného souboru, zatímco ostatní klienti mohou soubor zároveň číst. The původní použitý mount point nemusí být nutně stejný na všech uzlech a lze ho libovolně měnit.



Lustre General Architecture

Podporované operační systémy

bullx supercomputer suite komponenty jsou k dispozici na následujících operačních systémech:

- Redhat Enterprise Linux 6
- bullx Linux6

bullx Linux

[redacted]

[redacted]

[redacted]

[redacted]

Statické umísťování stránek nestačí v případě běhu vice aplikací, které navíc mění během svého života způsob přístupu do paměti.
Linux umožňuje standartní migrační mechanismus, který za běhu přesouvá stránky procesu, který cestuje mezi CPU. Tento mechanismus však není optimální, protože přesouvá vice než potřebné minimum stránek.

[redacted]

[redacted]

1.3.9.2. SYSTÉM PRO SPRÁVU INCIDENTŮ A POŽADAVKŮ

Pro potřeby správy incidentů a požadavků nabídka obsahuje softwarovou komponentu Argos, která zajišťuje:

- Incident processing, management & follow-ups
- Historie oprav a servisních zásahů
- Měření dostupnosti systému
- GUI & CLI rozhraní
- Víceuživatelské prostředí
- Správa náhradních dílů

[redacted]

1.3.9.3. CALYPSO

Pro potřeby zálohování a automatické archivace souborů je do nabídky zařazen Software Calypso. Popis automatické archivace a obnovy souborů je popsán v kapitole 5.3.3

1.3.9.4. PODPORA OS WINDOWS

[redacted]

[redacted]

1.3.10. SPOLEHLIVOST A DOSTUPNOST ŘEŠENÍ

Díky spolehlivému hardwarovému a softwarovému prostředí s redundancí všech kritických komponent zajišťuje architektura Malého clusteru bezproblémový chod uživatelských aplikací následujícími prostředky:

- [redacted] Pro administraci (nejkritičtější bod superpočítače) je odolnost zajištěna redundancí komponent: [redacted]

[redacted]

Použité softwarové řešení pro vysokou dostupnost je založeno [redacted]

HA [redacted]

redundantní s funkcionalitou multipathing.

- Správce úloh PBS Pro podporuje nativně překlopení služby job scheduleru na záložní server v případě výpadku
- 2 přístupové servery, v případě výpadku jednoho serveru je obsluha uživatelů zajištěna zbývajícím serverem.
- Servery Bullx R423-E3 (Admin/Login): Zabezpečení klíčových komponent: redundant power supply, Fiber Channel adapter, ECC DDR3 memory, OS disky mirrorovány.
- I/O služby:
  - **SCRATCH STORAGE**
    - 2 servery pro zaručení dostupnosti metadat (MDS)
    - 2 OSS servery s 1 + 1 redundancí
    - Redundantní řadiče diskového pole E5400
    - Zabezpečení RAID6 , hotspare disky
    - Redundantní Fibre Channel připojení k serverům, multipath
  - **HOME STORAGE**
    - 2 NFS servery pro zaručení dostupnosti dat
    - Redundantní řadiče diskového pole E5400
    - Zabezpečení RAID6 , hotspare disky
    - Redundantní Fibre Channel připojení k serverům
  - **BLOCK STORAGE**
    - Redundantní řadiče diskového pole E2600
    - Zabezpečení RAID6, hotspare disky
    - Připojení přes redundantní SAN

1.4. *Technické parametry nabídky*

Všechny technické parametry nabídky jsou uloženy v souboru „VŠB_Malý cluster_Příloha č. 6 ZD - Technické parametry nabídky.xlsx“, který je nedílnou součástí nabídky.