

MetaCentrum & BÚ

Souhrn informací pro instalaci clusteru [MetaCentra](#) v [Botanickém ústavu AV](#). Přehled pro všechny zúčastněné. Upravujte dle potřeby.

Co zbývá dodělat...

Fix Me!

-  Každý uzel musí mít [veřejnou IPv4 adresu](#) (bez NATu, nesmí být za firewallem) a [přidělené hostname](#).
-  Stroje se budou [instalovat po síti](#), budou bootovat přes PXE.
-  Připojení diskového pole.
-  Nastavení OPA.
-  Vytvoření [virtuálních strojů](#) a [přístup k jejich správě](#).
-  Dořešit klimatizaci celé serverovny.
-  Vyčistit servery a serverovnu.
- Zkontrolovat PDU.
- Doinstalovat `carax1`, `draba1`.
- Doinstalovat čelní uzel `tilia`.
- Dodělat `tilia` `nfs` (`Perun`, export přes NFS po IP a `Omni-path`, aliasy a připojení do zbytku `MetaCentra`).
- Předělání `/home/clusteru` ze `/storage/praha1` na `/storage/pruhonice1-ibot`.

Fix Me!

Hardware a jeho určení

Stručný přehled HW (uzlů) a [jeho fotky](#)... Víceméně vše od SuperMicro.

HPC uzly (3x)

- 4x Intel Xeon Gold 6230, 48x 32GB DDR4 RAM, 2X Samsung SSD 960GB NVMe 2,5"
- Určeny pro [Apache Spark](#), [Apache Hadoop](#) a další libovolné úlohy `MetaCentra`
- SSD by měly být v RAID 0 — na 1. disku `/`, na 2. `swap`, na zbytku přes oba disky `/scratch`

- Spark a Hadoop se bude spouštět přes [Singularity](#) (viz [níže](#))

"Standardní" výpočetní uzly (6x)

- 1x AMD EPYC Naples 7261, 8x 64GB DDR4 RAM, 2X Samsung SSD 960GB NVMe 2,5"
- Zcela standardní uzly MetaCentra
- SSD by měly být v RAID 0 — na 1. disku /, na 2. swap, na zbytku přes oba disky /scratch

Virtuální stroje

- MicroBlade 314E-220, osazen 7 žiletkami (MicroBlade modul 6128R-T2X), 7 pozic pro žiletky volných.
- Každá žiletka má 2x Intel Xeon E5-2640v4, 4x 32GB DDR4 RAM, 2x Intel SSD 120GB SATA3.
- Budou se [instalovat stejně jako ostatní přes PXE](#).

Potřebujeme tři stroje "zvláštního určení" (čelní uzel, správa a databázový server, viz níže). Pro tyhle servery platí stejné požadavky jako pro fyzické stroje, tedy hostname, veřejná IPv4 adresa nastavený [boot přes PXE](#). Adresy jsou v [přehledové tabulce](#).

- **Čelní uzel**
 - 4 vCPU, 8 GB RAM, min 50 GB disk.
 - Bude sloužit jako standardní [čelní uzel MetaCentra](#) — pro přípravu a spouštění úloh.
- **Databázový server** s databázemi pro HPC uzly (Spark, Hadoop)
 - 8 vCPU, 32 GB RAM, min 60 GB disk.
 - NoSQL databáze [Neo4j](#) a [MongoDB](#); a [MySQL/MariaDB](#) a [PostgreSQL](#)
 - Bude v naší správě (Vojtěch, Yann) a poběží na [openSUSE Leap](#).
 - Databáze budou přístupné normálně přes IP adresu, jiné speciální nastavení není nutné.
 - Nemělo by to být ve stejném adresním rozsahu, podsíti jako ostatní servery.
- **Managementovací server**
 - 2 vCPU, 2-4 GB RAM, min 12 GB disk.
 - Bude sloužit jako "brána" pro správu managovací sítě serverů (ikvm serverů, mgmt porty diskového pole, switche, atd.).
 - Bude mít přístup k IPMI, pro MetaCentrum bude sloužit ke správě serverů průhonického clusteru — aby to šlo podle potřeby restartovat a tak.
 - Možná na něm poběží mail server, možná ještě něco dalšího používaného ke správě clusteru.

Podpora hyperthreadingu

Všech 9 výpočetních strojů podporuje hyperthreading, má tedy k dispozici teoreticky 2x více vláken než jader. Na všech strojích má MetaCentrum nastaveno v PBS počet fyzických jader a úlohy si říkají o fyzická jádra. V současnosti nelze v PBS nějak míchat požadavky na fyzická a HT jádra, většina aplikací a uživatelů stejně počítá, že dostane plnotučné jádro.

Aplikace, které dokáží HT efektivně využít, lze si rezervovat celý stroj a použít libovolný počet jader — PBS v takovém případě nehlídá překročení požadavků na CPU.

Obecně zatím PBS neumí automaticky při požadavku na exkluzivní uzel natáhnout parametry úlohy na všechny zdroje uzlu, který plánovač přidělí, uživatel musí při použití hyperthreadingu apod. sám

přizpůsobit výpočet na konkrétní stroj, předem nebo za běhu. Automaticky se předává (např. jako default pro MPI utility integrované s PBS) požadované minimum počtu procesorů podle zadání.

Souborový server (1x)

- V tuto chvíli máme jeden, rádi bychom výhledově dokoupili ještě jeden.
- 1x AMD EPYC Naples 7401P, 8x 16GB DDR4 RAM, 2X Samsung SSD 960GB NVMe 2,5".
- V rámci MetaCentra je jako storage-pruhonice1-ibot.metacentrum.cz.

Diskové pole

- QSAN XCubeSAN, 21x 14TB 7200RPM SAS3
- Součástí pole je i SSD cache (SW Qcache "COQ SSD-C" a 2x SSD "PAH WUSTR6440ASS200")
- Kapacita je 179 TB

Použití diskového subsystému

Hodláme na něj zálohovat naše data, nicméně všichni uživatelé budou zároveň uživateli MetaCentra, s vhodně nastavenými kvótami (mohou tam být domovské adresáře všech uživatelů, v případě potřeby se jednotlivcům zvedne kvóta). Bude to vyřešené standardně přes NFS, úložiště bude zároveň dělat domovské adresáře průhonických strojů a bude přístupné ze všech uzlů MetaCentra. Diskové pole je zálohované na DÚ CESNETu. Budou tam uložena data pro Spark (v domovských adresářích jednotlivých uživatelů) — úložiště musí být přístupné ze Singularity obrazu Sparku.

Přehled strojů

Typ stroje	Hostname	IP adresa	MAC adresy	IPMI Management	Poznámky
SG350X switch	X	192.168.160.52	04:eb:40:e0:15:ad	X	X
SMP 1	carex1.ibot.cas.cz	147.231.111.21	ac:1f:6b:ad:2b:e8, ac:1f:6b:ad:2b:e9	ac:1f:6b:aa:c1:fb, 192.168.160.21	X
SMP 2	carex2.ibot.cas.cz	147.231.111.22	ac:1f:6b:ad:2b:72, ac:1f:6b:ad:2b:73	ac:1f:6b:aa:c1:bd, 192.168.160.22	X
SMP 3	carex3.ibot.cas.cz	147.231.111.23	ac:1f:6b:ad:2b:78, ac:1f:6b:ad:2b:79	ac:1f:6b:aa:c1:c0, 192.168.160.23	X
SMP 4	carex4.ibot.cas.cz	147.231.111.24	ac:1f:6b:ad:2b:6c, ac:1f:6b:ad:2b:6d	ac:1f:6b:aa:c1:bb, 192.168.160.24	X
SMP 5	carex5.ibot.cas.cz	147.231.111.25	ac:1f:6b:ad:2c:84, ac:1f:6b:ad:2c:85	ac:1f:6b:aa:c2:48, 192.168.160.25	X
SMP 6	carex6.ibot.cas.cz	147.231.111.26	ac:1f:6b:ad:2c:8c, ac:1f:6b:ad:2c:8d	ac:1f:6b:aa:c2:34, 192.168.160.26	X
HPC 1	draba1.ibot.cas.cz	147.231.111.31	ac:1f:6b:c1:b8:a6, ac:1f:6b:c1:b8:a7, ac:1f:6b:c1:b8:a8, ac:1f:6b:c1:b8:a9	ac:1f:6b:a1:44:9b, 192.168.160.27	X
HPC 2	draba2.ibot.cas.cz	147.231.111.32	ac:1f:6b:c1:b4:26, ac:1f:6b:c1:b4:27, ac:1f:6b:c1:b4:28, ac:1f:6b:c1:b4:29	ac:1f:6b:a1:51:28, 192.168.160.28	X
HPC 3	draba3.ibot.cas.cz	147.231.111.33	ac:1f:6b:c1:b8:82, ac:1f:6b:c1:b8:83, ac:1f:6b:c1:b8:84, ac:1f:6b:c1:b8:85	ac:1f:6b:a1:44:b2, 192.168.160.29	X
Čelní uzel	tilia.ibot.cas.cz	147.231.111.10	X	X	VM

Typ stroje	Hostname	IP adresa	MAC adresy	IPMI Management	Poznámky
Úložiště	tilia-nfs.ibot.cas.cz	147.231.111.20	X	ac:1f:6b:aa:c1:c6, 192.168.160.20	X
Diskové pole	X	192.168.1.1, 192.168.11.1	??	192.168.160.53	X
Databázový server	sorbus.ibot.cas.cz	147.231.111.11	X	X	VM, pod správou BÚ
Managementovací server	tilia-mng.ibot.cas.cz	147.231.111.12	X	X	VM
PDU 1	X	X	X	00:20:85:e0:5f:8f, 192.168.160.50	PDU
PDU 2	X	X	X	00:20:85:e0:58:f6, 192.168.160.51	PDU

Výpis HW podle dodavatele

- Spotřeba W celkem 7347
- U Celkem 33

Uzel SMP 6ks

- .a-2610Q-H11S 2U-4U S-SP3,2GbE,8/16sATA,2NVMe,3 PCI-E16,3PCI-E8,8DDR4-2666
- CPA E7261 Amd EPYC Naples (SPA3 LGA) 7261 — 2,5GHz, 8core/16thread, 64MB L3, 170W/155W, 1P/2P, WOF 1 Ks
- MBS H11SSL-NC-B H11SSL-NC S-SP3, ATX, 2GbE, 8DDR4-2666, 3PCI-E16g3,3-E8, 2NVMe, 8SAS3(LSI3008), 8sATA,M.2,IPMI 1 Ks
- PA4 64EL-2666-4R 64GB 2666MHz DDR4 ECC LoadReduced 4R, LP(31mm), Samsung 8 Ks
- LAI OPA100-1-E8LP Intel Singl port OPA100 (QSFP28) PCI-E8 LP 1 Ks
- LA# I210AT G2 Síťová karta Intel i210AT 2× GbE na zákl. desce 1 Ks
- VG# AST2500 Aspeed AST2500 1 Ks
- PAS PM983 0960 25 Samsung SSD PM983 960GB NVMe 2,5" 2 Ks
- CO# IPMI XX IPMI 2.0 modul s KVM-over-LAN na základní desce 1 Ks
- KAS CBL-SAST-0934-12 OCuLink NVME SFF-8611 -> OCuLink NVME SFF-8611 — 50cm 2 Ks
- CAS MCP-220-82619-0N Přídavný NVMe box pro 2×2,5" hotswap do SC826B/SC216B/846X/847B/417B/226S na zadní straně (oculink konektor) 1 Ks
- CAS SC826TS-R700LPB SC826TS-R700LP noBPN,CD 2U eATX13,12sATA/SAS, rPS 700W, 7LP, černé 1 Ks
- CZ# 2 Redundantní zdroj 1 Ks
- CHS SNK-P0063AP4 SNK-P0063AP4 Aktivní 2U heatsink pro AMD SP3 LGA 1 Ks
- SLA ZAS NBD OSTAT Zásah 3r NBD on-site — cena po individuálním posouzení 1 Ks
- Spotřeba W 1ks 409
- Spotřeba W celkem 2454
- U Celkem 12

Uzel HPC 3ks

- CPI XC6230-TR Intel Xeon Gold 6230 — 2,1GHz@10,40GT 27,50MB cache 20core,HT,125W,FCLGA3647,2P/4P,2933MHz tray 4 Ks
- PCS 2049U-TR4 SuperServer 2049U-TR4 2U 4S-P, 4GbE,20SFF+4SFF/NVMe, IPMI, 48DDR4, 6PCI-E16,8-E8, rPS (80+ TITANIUM) 1 Ks
- SLA ZAS NBD OSTAT Zásah 3r NBD on-site — cena po individuálním posouzení 1 Ks

- PA4 32ER-2933-2R 32GB 2933MHz DDR4 ECC Registered 2R, LP(31mm), Samsung 48 Ks
- LAI OPA100-1-E8LP Intel Singl port OPA100 (QSFP28) PCI-E8 LP 1 Ks
- COS AOC-SLG3-4E4T Supermicro AOC-SLG3-4E4T — 4×NVMe (Oculink SFF-8611) HBA, PCI-E16 g3 LP (X11) 1 Ks
- KAS CBL-SAST-0929 OCuLink NVME SFF-8611 (AOC) -> SFF-8643 NVME (backplane) — 57cm 4 Ks
- CAS MCP-220-00127-0B 2,5" NVMe rámeček pro SC113/116/119/213/216/219 (gen3) 2 Ks
- PAS PM983 0960 25 Samsung SSD PM983 960GB NVMe 2,5" 2 Ks
- Spotřeba W 1ks 1210
- Spotřeba W celkem 3630
- U Celkem 6

Souborový server

- .a-2610Q-H11S 2U-4U S-SP3,2GbE,8/16sATA,2NVMe,3 PCI-E16,3PCI-E8,8DDR4-2666 1ks
- CPA E7401P Amd EPYC Naples (SPA3 LGA) 7401P — 2,0GHz, 24core/48thread, 64MB L3, 170W/155W, 1P pouze, WOF 1 Ks
- MBS H11SSL-NC-B H11SSL-NC S-SP3, ATX, 2GbE, 8DDR4-2666, 3PCI-E16g3,3-E8, 2NVMe, 8SAS3(LSI3008), 8sATA,M.2,IPMI 1 Ks
- PA4 16ER-2666×1R 16GB 2666MHz DDR4 ECC Registered 1R, LP(31mm), Samsung 8 Ks
- LAI OPA100-1-E8LP Intel Singl port OPA100 (QSFP28) PCI-E8 LP 1 Ks
- LAS AOC-STGN-I2S Supermicro STGN-i2S — Dual port 10GbE (SFP+) PCI-E8 (g2) LP, odpovídá X520A-DA2 1 Ks
- LAS AOC-STGS-I1T Supermicro STGS-I1T — Singl port 10GbE (10GbE-T) PCI-E4 (g3) LP, odpovídá X550AT 2 Ks
- LAS AOC-S40G-I1Q Supermicro S40G-I1Q — Singl port 40GbE (QSFP+) PCI-E8 (g3) LP, odpovídá XL710 2 Ks
- LA# I210AT G2 Síťová karta Intel i210AT 2× GbE na zákl. desce 1 Ks
- VG# AST2500 Aspeed AST2500 1Ks
- PAS PM983 0960 25 Samsung SSD PM983 960GB NVMe 2,5" 2Ks
- CO# IPMI XX IPMI 2.0 modul s KVM-over-LAN na základní desce 1Ks
- KAS CBL-SAST-0934-12 OCuLink NVME SFF-8611 -> OCuLink NVME SFF-8611 — 50cm 2Ks
- CAS MCP-220-82619-0N Přídavný NVMe box pro 2×2,5" hotswap do SC826B/SC216B/846X/847B/417B/226S na zadní straně (oculink konektor) 1Ks
- CAS SC826TS-R700LPB SC826TS-R700LP noBPN,CD 2U eATX13,12sATA/SAS, rPS 700W, 7LP, černé 1Ks
- CZ# 2 Redundantní zdroj 1Ks
- CHS SNK-P0063AP4 SNK-P0063AP4 Aktivní 2U heatsink pro AMD SP3 LGA 1Ks
- SLA ZAS NBD OSTAT Zásah 3r NBD on-site — cena po individuálním posouzení 1Ks
- Spotřeba W celkem 433
- U Celkem 1

Diskové pole 1ks

- QXS16 QSAN 3U XeonD/max.128GB,iSCSI SAN,16×SAS3, 4×10GbE-T,volitelně:16×FC/10GbE, DualCntrl,rPS
- COQ XS3224-D QSAN XCubeSAN XS3224 4U,iSCSI SAN,24×SAS3, 4×10GbE-T,volitelně:16×FC/10GbE, DualCntrl,rPS 1Ks
- COQ C2F-SP256G QSAN XCubeSAN XS1200/3200/5200/7200 — SuperCap + M.2 flash

- (max.256GB RAM) 1Ks
- COQ XS-08GB QSAN XCubeSAN XS1200/3200/5200 — DDR4 8GB 1Ks
- HDH WUH721414AL4204 14TB WDC Ultrastar He14/DC530 — 7200rpm, SAS3, 4kn, 512MB, (SE) 3,5" 21Ks
- LPS C6-018 Cat6 RJ-45 (10GbE-T) — 180cm, žlutý 2Ks
- COQ SLR-RM3640 QSAN XCubeSAN/DAS XS3200/5200/5300/7200 — kolejnice 1Ks
- KAN 220V 1A 16A Value Kabel síťový prodlužovací, IEC320 C14 — C13, 1m, černý 2Ks
- SLA OP NBD P500Q Oprava 3r NBD on-site pro QSAN P500Q, Příjem hlášení 8-22 hod. PO-PA 1Ks
- Spotřeba W celkem 430
- U Celkem 3

Infrastruktura

- KAM QSFP-030 2Ks
 - Mellanox MC2210128-003 QSFP+-QSFP+ 40GbE kabel, 3m
- KAS CBL-0892-OPC20 10Ks
 - Supermicro QSFP28 OmniPath (OPA100) pasivní DAC kabel 2m
- LAC SG220-50-K9-EU 1Ks
 - Cisco SG220-50 50-Port Gigabit Smart Switch
- LAS SSH-C48QM 1Ks
 - INTEL® Omnipath 100G (OPA100) ToR switch 48 QSFP portů 2PS
- LPX RTA-47-L81-CAX-A1-MAA + RAX-VC-X42-X2 1Ks
 - 19" rozvaděč stojanový 47U/800×1000, typ RT-A3, dv. síto 80%-6mm,RAL7035
- SLA ZAS NBD OSTAT 1Ks
 - Zásah 3r NBD on-site — cena po individuálním posouzení
- UPE 9SX11KiRT 1Ks
 - UPS 1/1fáze, 11kVA — 9SX 11000i RT6U
- UPE 9SX5KiRT 5KV 3U 1Ks
 - UPS 1/1fáze, 5kVA — 9SX 5000i RT3U
- UPE EFLX12I 4Ks
 - FLEXPDU 12 IEC/2X 6 IEC 10A 1X 16A
- UPE INSTAL 1Ks
 - Standardní instalace na připravené přívody k UPS 7-15kVA
- UPE KK MS 5P5PX 2Ks
 - Komunikační karta — MS Web/SNMP Gigabit (pro 5P, 5PX,
- Síťová kabeláž 1Ks
 - Kabeláž dle potřeby
- Spotřeba W celkem 400
- U Celkem 11

Změny oproti výpisu výše

1. Dvě UPS typu "UPE 9SX11KiRT" a "UPE 9SX5KiRT 5KV 3U" budou nahrazeny jednou "UPE EAZC122440110000" včetně komunikačních karet.
2. 4x pdu rozvody typu "UPE EFLX12I" budou nahrazeny 2x řízenými pdu typu "UPE EMIB20"
3. Server typu "Diskové pole" bude doplněn o SW Qcache "COQ SSD-C" a 2x SSD "PAH WUSTR6440ASS200"
4. Lan switch "LAC SG220-50-K9-EU" bude nahrazen switchem "LAC SG350X-48-K9-EU"

Kontakty

Zodpovědní za nákup HW a jeho vědecké využití:

- [Yann JK Bertrand](#), 271015472 — šéf bioinformatiky
- [Vojtěch Zeiek](#), 271015173, 608944101, <https://trapa.cz/cs> — hlavní pracovní síla bioinformatiky



Správa IT v BÚ --- přístup do serverovny

Zmáčknout reset, nastavovat něco v BIOSu, občas vytáhnout nějaký kabel.

- [Jana Vojířová](#) (vedoucí IT), 271015210, 602300401
- [Jan Panoch](#) (zodpovědný za síť)
- [Jan Král](#), 271015210

Apache Spark a Hadoop

[Apache Spark](#), [Apache Hadoop](#) — toto se týká hlavně Yanna

Asi nejjednodušší bude spouštět Spark v Singularity. Existují [obrazy pro Docker](#), které [by měly být snadno převoditelné na Singularity](#), případně můžeme [připravit obraz pro Singularity](#). Pak se podle

nás bude Spark dát používat v intencích [dokumentace MetaCentra](#).  Předpokládáme, že bude jednodušší mít Spark v Singularity, než jej [instalovat ručně](#), ač ani to by neměl být zásadní problém. Root účet každopádně nepotřebujeme.

Singularity Spark kontejner dáme k dispozici ostatním uživatelům. Vystavíme jej v určitém adresáři (na docker/singularity hubu, AFS, NFS, ...) a na [wiki](#) popíšeme, jak si jej zkopírovat, spustit a případně upravit.

Na uzlech, kde Spark poběží nebude třeba žádná speciální trvale běžící služba, vše se bude spouštět standardně přes qsub. Do obrazu Sparku bude připojeno naše úložiště (takže třeba něco jako `/storage/pruhonice2-archive/...` připojené třeba do `/mnt`, něco jako `singularity -B /storage -B /auto -B /mnt`). Úlohy se tedy budou spouštět na vyžádání, stáhnou si data z

centrálního úložiště a na konci po sobě uklidí (vyčistí se `/scratch`). Prostě jako obvykle.  Úlohy mohou běžet velmi dlouho (q_2w apod.), nicméně volná kapacita uzlů bude k dispozici pro libovolné jiné úlohy.

Z kontejneru Singularity budou také přístupné databáze na našem databázovém VM (viz výše) standardně přes IP adresu.

Síťové servery v BÚ

- **DHCP server:** 192.168.1.1 (nebo 192.168.160.1 na samostatné vln pro výpočetní cluster)
- **DNS server:** 192.168.3.93 a 192.168.3.94
- **Mail server:** smtp.ibot.cas.cz (192.168.3.214)

Fotky clusteru

Rack s clusterem



Datové úložiště



HPC uzly



Standardní výpočetní uzly



Virtuální servery



Kabeláž



Prioritní přístup BÚ k průhonickému clusteru

- Lze omezit maximální délku úloh ostatních uživatelů spouštěných na průhonickém uzlu.
 - Typicky jsou alespoň z počátku povoleny úlohy “cizích” uživatelů jen do q_1d včetně. Podle zkušeností to lze upravit.
- V [Perunovi](#) je skupina `ibot`, kterou spravuje VZ a která zajišťuje přístup do speciální fronty zajišťující prioritní spuštění úloh na průhonickém clusteru (`qsub -q ibot ...`) — není nutné

specifikovat konkrétní stroj/cluster (nicméně to lze kombinovat). Skupina iBot zároveň zajišťuje kvótu 2 TB na průhonickém datovém úložišti (lze ji navýšit).

- Na úložišti jsou adresáře všech uživatelů, ale všichni kromě členů určitých skupin nebo vybraných jednotlivců mají limit 10 GB. S kvótami pro uživatele/skupiny lze libovolně manipulovat.

Správa skupiny iBot

- Registrační formulář není k dispozici, správa skupiny je možná přímo v [Perunovi](#) (správcem je VZ). Registrační formulář by se vyplatil při cca 50 a více uživatelích. Bylo by potřeba vědět, co by na přihlášce noví uživatelé měli vyplňovat, třeba nějaké zdůvodnění atp. A jestli schvalovat ručně, správcem, nebo automaticky. Také je potřeba vědět, jestli členství v iBot bude expirovat nějak jinak, než v Meta.
- K dispozici je [E-mailová konference](#), jejímiž členy jsou automaticky všichni členové skupiny iBot.

Sdílené adresáře na úložišti

Vhodné pro:

1. Zálohování dat nějaké laborky, přístroje, apod., aby to nebylo vázáno na konkrétního uživatele, ale aby tam mohlo zapisovat víc lidí a/nebo aby existoval nějaký servisní účet.
2. Sdílení dat mezi více lidmi pracujícími na jednom projektu.

Podle potřeby se budou přidávat skupiny zajišťující oddělený přístup do sdílených adresářů. Možností je i vytvoření servisního uživatele v nějaké skupině, do které zároveň zařadíme uživatele, kteří mají mít možnost čtení a/nebo zápisu v daném sdíleném adresáři.

Dotazy na MetaCentrum

1. Odpovídají [vybraná doménová jména uzlů](#) tradicím a stylu MetaCentra? 
2. Instalují se [virtuální servery](#) jinak, než fyzické? Vyžadují nějaké zvláštní nastavení?
3. Součástí [diskového pole](#) bude SSD cache. Bude to vyžadovat nějakou speciální instalaci?
 - Možná se ukáže, že to jako cache moc efektivně nefunguje a použijeme to nějak jinak...
4. Bude [databázový server](#) pod naší správou vyžadovat nějaké speciální nastavení, anebo to prostě bude běžný linuxový server?
5. Jakým způsobem dáme "náš" Singularity [Spark](#) kontejner k dispozici ostatním uživatelům. Vystavíme jej v určitém adresáři a na wiki popíšeme, jak si jej zkopírovat, spustit a případně upravit?
6. Z kontejneru [Spark](#) Singularity musí být možný přístup k databázím na našem databázovém VM. Nepředpokládáme, že by toto byl problém. Vyžaduje to nějaké speciální nastavení?
7. Z kontejneru [Spark](#) Singularity musí být přístupné úložiště `/storage/pruhonice2-archive/...`
8. Jak bude vypadat ta instalace? Budete se chtít domluvit na nějakém přesném čase? Anebo to jen nabootuje přes PXE a Vy se toho automaticky ujmete? Anebo budete nejdříve chtít přístup k IPMI, abyste si to mohli restartovat podle potřeby?
9. Předpokládám, že Vám budu posílat nějaké přístupové údaje. Jak to uděláme bezpečně? Nemáte

třeba k pracovní adrese GPG?

10. Ohledně sdílení s ostatními uživateli MetaCentra je naše představa taková, že naši uživatelé budou mít vysokou prioritu a jejich úlohy se na našem clusteru budou spouštět rychleji, než by se spouštěly na nějakém náhodném uzlu. Spíš bychom to zkusili bez toho, aniž bychom si nějaké uzly rezervovali výhradně pro sebe. Možná bychom mohli (zpočátku) omezit délky úloh/maximální zdroje, které na našem clusteru ostatním dovolíme spustit. Je-li tato představa

správná, jak se upravují ty limity? Tak, že Vám napíšeme mail?  Jak se poznají “naši” uživatelé? Tak, že jsou členy nějaké VO, jejichž členy si v Perunu spravujeme my sami? A při spouštění úlohy pak ke qsub přidáme parametr cluster/host/vnode?

11. Jak budeme nastavovat, kdo bude mít jakou kvótu na našem diskovém poli? Pomocí nějaké VO spravované v Perunu? Bylo by ideální, abychom Vás s každou změnou nemuseli obtěžovat, protože těchto změn může být docela dost...
12. Jaká je adresa na formulář na žádost o připojení do skupiny ibot?
13. Jak se přistupuje na datové úložiště? Jaké adresy a jaké protokoly?
14. Zapnout hyperthreading? Lscpu ukazuje na draba3 160 jader, ale plánovač jen 80.
15. Jaké zvolit nastavení limitů pro nečleny skupiny ibot?
16. Jak budou spravována hesla managementovacích adres?
17. Jak je to s PDU? Funguje správně?



Kam potřebují mít správci MetaCentra přístup

- Switch SG350X (webové rozhraní, SSH) propojující všechny fyzické uzly (HPC a SMP) a diskový server.
- IPMI na konzoli všech serverů v racku připojených switchem SG350X. Potom bude MetaCentrum moct instalovat uzly zcela samostatně. V podstatě to bude probíhat tak, že si z PXE MetaCentrum naboootuje vždy jeden uzel za každou HW konfiguraci, odzkouší, že instalace funguje správně a pak spustí instalaci na zbytku clusteru.
- Správa virtuálních serverů (čelní uzel, správcovský uzel apod.). Webové rozhraní virtualizační platformy nebo nějaký jiný způsob, jak případně zaseknutý virtuál natvrdo resetovat a u té správy virtuálních serverů potřebujeme hlavně přístup na konzoli, abychom mohli instalovat/reinstalovat případně debugovat, když se virtuál zasekne.

Správa virtuálních serverů MetaCentrem

Příprava na straně BÚ a následná správa MetaCentrem...

Pokud bude nějaký stroj instalován ze strany BÚ, stačí na něj dát SSH klíč ssh-ed25519 AAAAC3NzaC1lZDI1NTE5AAAAIA+u1/VyPP3G10Qd+ud+15KmcHG9lla69g/2y4qinJkg dexter a

do nějakého textového souboru uložit hesla. 

Instalace

- Na vsechny vypocetni stroje nainstalujeme automatickou instalaci Debian9
- Na frontend vypocetnich uzlu nainstalujeme automatickou instalaci Debian9 NEBO (pro virtualy mozna o trochu jednodussi varianta) vy nainstalujete cisty Debian9 a date mi tam ssh klic na roota
- Na frontend diskoveho pole nainstalujeme automatickou instalaci Debian10
- Na management stroj nainstalujeme automatickou instalaci Debian10 NEBO (pro virtualy mozna o trochu jednodussi varianta) vy nainstalujete cisty Debian10 a date mi tam ssh klic na roota

Jeste tam, kde je "vy nainstalujete" je mozna i varianta, ze nam date pristup na konzoli s pripojenym debian netinst iso a ja si to nainstaluji sam. Ohledne verzi zjednodusene receno - tam kde maji uzivatele moznosti spoustet vypocetni SW bude prozatim Debian9 pricemz upgrade celeho MetaCentra na Debian10 probehne behem pristiho roku. Na ostatnich uzlech bude Debian10.

Správcovská rozhraní

- Správa virtuálů (proxmox): <https://192.168.4.20:8006/>
- Správa SuperMicro clusteru: <https://192.168.160.20/>
- Switch clusteru: viz [IPMI adresy](#)

Požadavky MetaCentra na síť

Zaklad je sitovy pristup, domluvte si, ze budete mit pripravene napojeni vaseho switchne na ten centralni, pripraveny na prelomu roku, ze budete mit pripravene IP adresy apod. Pro instalaci si kolegové s vami domluvi vse potrebne (instalujeme to s DHCP+TFTP), zbytek uz by mela byt vice mene automatika.

Potrebujeme pro kazdy uzel jednu verejnou IPv4 adresu, která nebude za firewalllem. Vhodne je mit ty stroje v jednom souvislem bloku adres. Co se nazvu tyce, bud udelame 2 ruzne nazvy (HPC/SMP), nebo jeden nazev s cisly. Konkretni nazev necham na vas, napr. na MU se inspiruji J. R. R. Tolkienem (Doom, Arien, Skirit), v Olomouci pivem (Radegast, Gambrinus), na ZCU je to recka mytologie (Nympha, Minos). Kazdy uzel musi mit v DNS A a PTR zaznam.

V pripade, ze mame v jedne lokalite a v jedne dodavce nekolik HW konfiguraci, osvedcilo se nam pojmenovavat tak, aby bylo jedno spolecne jmeno pro cluster ale bylo zde rozliseni cisly, ve vasem pripade to konkretne bude (v pripade, ze se cluster bude jmenovat treba kytka):

```
# SMP:
kytka1-1.ibot.cas.cz
kytka1-2.ibot.cas.cz
..
kytka1-6.ibot.cas.cz

# HPC:
kytka2-1.ibot.cas.cz
kytka2-2.ibot.cas.cz
kytka2-3.ibot.cas.cz
```

Samozrejme pokud by se vam vic libilo mit 2 ruzna jmena, tak je to taky mozne, pak jen trochu

nebude sedet nazev frontendu a diskoveho pole, protoze se bude jeden z clusteru jmenovat jinak (coz taky samozrejme nicemu nevadi, jen kosmetika).

Kazdy uzel musi mit primo jednu verejnou IPv4 adresu bez jakychkoliv NATu a firewallu. V opacnem pripade bude problem jednak provest vzdalenou instalaci a jednak uzel provozovat. Totez bych prosil i pro -mng, frontend, frontend diskoveho pole, zkratka vseho, co chcete integrovat do MetaCentra.

Instalace --- instrukce z MetaCentra

Co se samotne instalace tyce, nejjednodussi zpusob je nechat nas cluster nainstalovat pres PXE hromadnou automatickou instalaci. K tomu staci, kdyz **bud**:

- nastavite vas DHCP server nasledovne:

nastavení DHCP serveru

```
use-host-decl-names on
next-server 147.251.9.127
if option architecture-type = 00:09 {
filename "/grub/x86_64-efi/core.efi";
} elsif option architecture-type = 00:07 {
filename "/grub/x86_64-efi/core.efi";
} else {
filename "/grub/i386-pc/core.0";
}
# a u jednotlivych stroju jeste pro jistotu
option host-name "clusterXY.domena.cz";
```

- **NEBO** nastavite dhcp-relay na sitovem prvku na 195.113.214.4 a poslete nam seznam MAC adres, IP adres a hostnamu jednotlivych stroju

Nasledne se necha jeden stroj za kazdou konfiguraci nabootovat z PXE, my si ho osahame, navrhneme rozlozeni disku a nainstalujeme cely cluster.

Pokud by vam tento zpusob nevyhovoval nebo by stroje mely problem s bootovani pres PXE, umime vytvorit MetaCentrum instalaci i z ciste nainstalovaneho Debian 9 Stretch. Neni to preferovany zpusob, ale pokud by vam to sedelo vice, je to mozne.

Dale bychom potrebovali znat **adresy DNS serveru a mail serveru**, ktere je mozne na clusteru pouzivat. Pres ten mail server budeme posilat postu jak pro administratory tak pro uzivatele. Pripadne umime oboje outsourcovat na servery CESNETu, ale lepsi je to mit co nejbliz clusteru.

From:
<https://oxalis.ibot.cas.cz/wiki/> - Lab, project and server notes

Permanent link:
https://oxalis.ibot.cas.cz/wiki/doku.php/metacentrum_uzly?rev=1573128892

Last update: 2019/11/07 13:14



