Příloha č. 1_Specifikace prací

ClusterFIT – Zadání konceptu a požadavků realizace

Cílem projektu je navrhnout postup realizace pro vytvoření globální a versatilní výpočetní platformy ClusterFIT, tj. propojení výpočetní kapacity v rámci celé infrastruktury Fakulty informačních technologií (FIT) jakými jsou učebny, servery, obecné počítače, notebooky apod. vše včetně vytvoření a nasazení Proof-Of-Conceptu řešení (PoC) a testovacího řešení. Platforma umožní uživateli přehledně využít volné výpočetní prostředky fakulty nejen pro vykonávání vlastní pracovní činnosti, ale i v rámci sebevzdělávání.

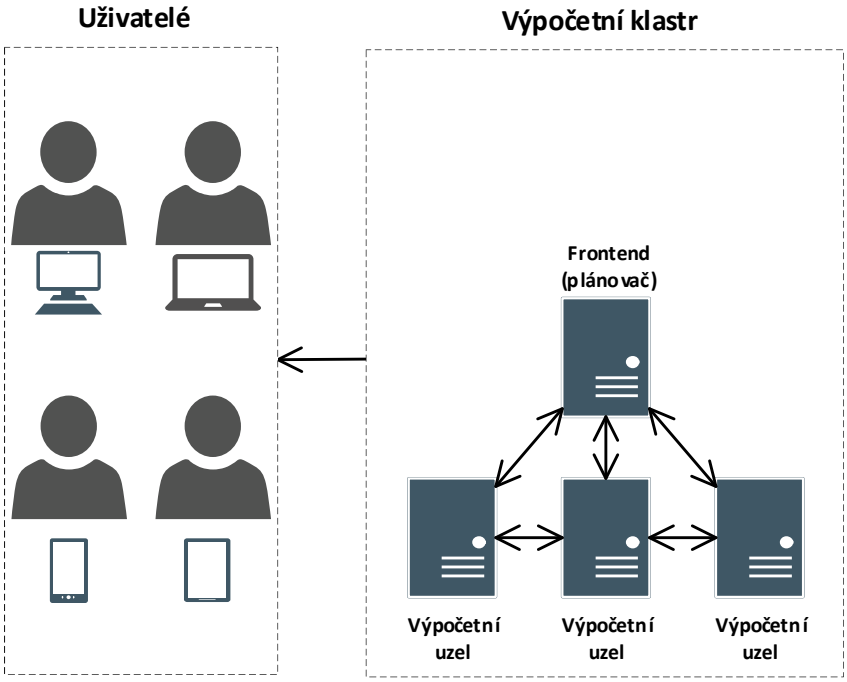
Požadavky na platformu

- Podpora CPU i GPU výpočtů,
- Podpora různých CPU architektur, min. x86-64, ppc64, ARM,
- Podpora různých architektur GPU, min. CUDA, RADEON,
- Podpora výpočtů GPGPU min. na úrovni fyzických serverů, min. v architektuře CUDA,
- Podpora paralelních HPC výpočtů přes rozhraní MPI,
- Realizace výpočetního klastru bude provedena v systému openSUSE,
- Instalace a správa výpočetního klastru bude provedena v nástroji pro automatizaci SALT, v řešení realizovaným v UYUNI (https://www.uyuni-project.org/),
- Řešení musí podporovat horizontální i vertikální škálovatelnost,
- Služba bude dostupná všem uživatelům se vztahem k fakultě způsobem kdy rozhraní ClusterFIT bude dostupné odkudkoliv pro přihlášené uživatele,
- Možnost využívat výpočetních a kapacitních prostředků v rámci:
    - Učeben, kdy počítače v učebnách mohou být kdykoli restartovány do různých OS, můžou být libovolně vypnuty / zapnuty,
    - Fyzických serverů, kdy tyto mohou být libovolně restartovány, vypnuty / zapnuty,
    - Virtuálních serverů, kdy tyto mohou být libovolně restartovány, vypnuty / zapnuty, či migrovány v rámci virtualizační platformy CloudFIT postavené nad OS Linux SLES, QEMU, KVM řízeny systémem OpenNebula s clusterovým datovým úložištěm OCFS2 propojeném na protokolu Fibre Channel 32Gbps a provozovaném nad centrálním diskovým polem typu SAN exportovaným jako standardní LUN,
    - Virtuálních serverů, kdy tyto mohou být libovolně restartovány, vypnuty / zapnuty, či migrovány v rámci virtualizační platformy VMware řízeny systémem vCenter s clusterovým datovým úložištěm VMFS propojeném na protokolu Fibre Channel 32Gbps a provozovaném nad centrálním diskovým polem typu SAN exportovaným jako standardní LUN,
    - Uživatelských stanic, kdy tyto mohou být libovolně restartovány do různých OS, dlouhodobě vypnuty / zapnuty,
- Implementace výpočetních uzlů v Linux openSUSE a s podporou pro OS s jádry Windows 10, Windows 11, Linux SUSE, RedHat, Debian a Ubuntu. Podpora jádra OS Linux Alpine výhodou,

- Řešení musí umožňovat připojení přes novou ethernetovou síť fakulty NetFIT s podporou autentizace přes 802.1x s koncepcí RBA (Role-Based-Access), centrálně řízenou politikami rolí v systému Aruba ClearPass s možností několikanásobného zanoření sítě v síti,
- Součástí řešení bude monitoring služby a všech jejích částí a bude propojitelný s monitorovacími nástroji Zabbix a NAGIOS,
- Využití výpočetních prostředků učeben bude možné i v době probíhající výuky bez negativního dopadu na její průběh díky nastavitelnému škálování maximálního zatížení uzlů,
- Řešení bude podporovat plánování úloh na základě různých kritérií, min. však dle:
    - Priority výpočtu,
    - Časového omezení,
    - Velikosti zdrojů – min. velikost RAM a počtu CPU,
- Řešení umožní spouštění interaktivních i neinteraktivních úloh,
- Řešení umožní plánování časově omezených i neomezených úloh,
- Řešení bude podporovat možnost exkluzivního přístupu ke zdrojům,
- Řešení bude podporovat možnost tzv. spravedlivého plánování úloh, kdy se všem uživatelům jejich úlohy dostanou na řadu v rámci stejného poměrového časového úseku,
- K řešení bude vytvořena dokumentace pro administrátory pro zajištění podpory provozu a nastavení řešení,
- K řešení bude vytvořena dokumentace pro uživatele, kde budou jasně popsány možnosti využití platformy pro uživatele,
- Řešení musí umožňovat rozšiřitelnost vlastností plánovače,
- Řešení musí podporovat rezervaci zdrojů v plánovaných časových úsecích pro vybranou skupinu uživatelů (např. rezervaci výpočetních zdrojů v poledne ve 13:00 na 3 hodiny),
- Řešení musí umožnit využívání externích zdrojů veřejných cloudových služeb min. pro Microsoft Azure, Amazon AWS a Google Cloud; podpora dalších výpočetních cloudů výhodou,
- Řešení musí podporovat řízení spotřeby výpočetních uzlů pro řízení vyhrazených výpočetních uzlů pomocí vypínání či snižováním spotřeby elektrické energie u uzlů, které nejsou využívány. V případě potřeby bude možné uzly znovu zapnout či jim zvýšit výkon.
- Řešení musí umožnit podporu pro oddělené provozování výpočetních klastrů, kde úlohy mohou běžet jak v rámci daného klastru anebo přes všechny výpočetní klastry najednou,
- Řešení bude podporovat textovou i grafickou vizualizaci stavu a vlastností plánovače/úloh,
- Řešení bude podporovat základní i rozšířený koncept výpočetní platformy, popsán níže,
- Součástí dodávky bude zaškolení obsluhy v rozsahu min. 5MD, vysvětlení základních řídících parametrů systému a kompletní provozní dokumentace,
- Součástí dodávky budou konzultační práce soužící k podpoře provozu systému v rozsahu min. 10MD,
- Celkové řešení musí být realizováno v max. časovém rozsahu 50MD,

Příloha č. 1_Specifikace prací

Popis základního konceptu výpočetního klastru ClusterFIT

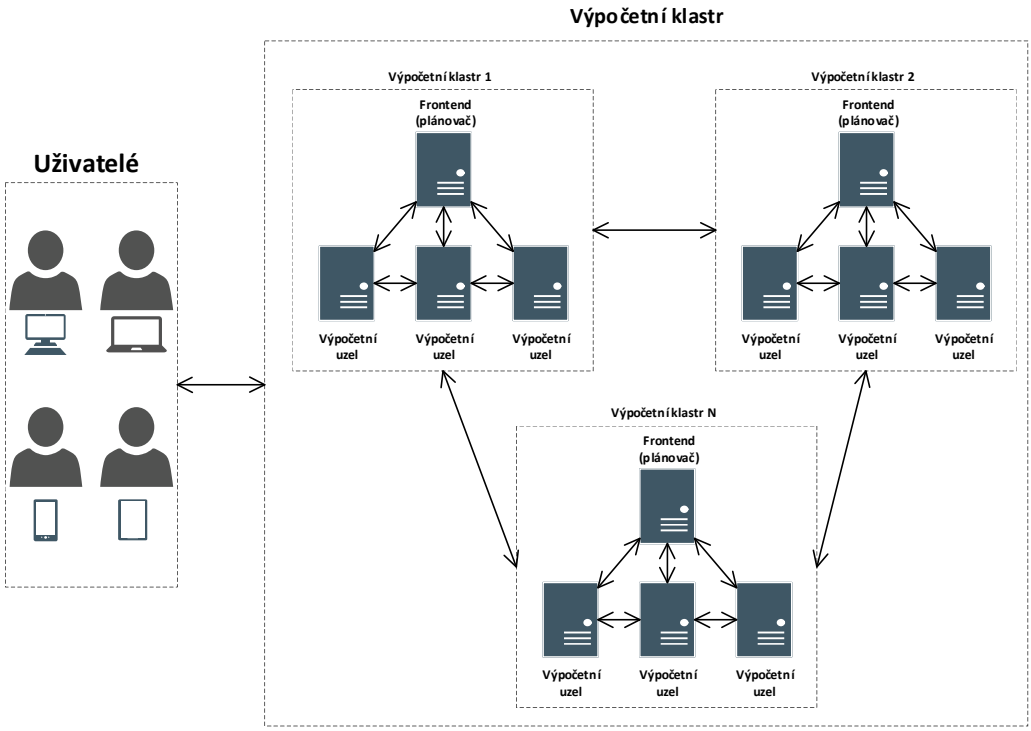


Uživatelé se přihlašují na frontend, kde může běžet i samotný plánovač. Uživatelé na frontendu připravují své programy a výpočty. S pomocí plánovače řídí své úlohy (spouští, ruší, monitorují) a čekají na výsledky svých výpočtů. Plánovač na základě parametrů naplánovaných úloh a na volných výpočetních zdrojů, které má k dispozici, vybírá jednotlivé výpočetní uzly, na kterých budou prováděny naplánované úlohy.

Výpočetní uzly mohou být vyhrazeny pro běh pouze jedné naplánované úlohy anebo mohou být sdíleny mezi více najednou běžících naplánovaných úloh. Výpočetní uzly také mohou být sdílené mezi uživateli, nebo vyhrazené pro omezenou skupinu uživatelů.

Příloha č. 1_Specifikace prací

Popis rozšířeného konceptu výpočetního klastru ClusterFIT



Výpočetní platformu ClusterFIT bude možné rozdělit do menších výpočetních clusterů podle různých kritérií anebo jejích kombinací. To závisí na konkrétních scénářích praktického využití, minimálně jej bude možné rozdělit dle kritérií:

- CPU výpočty,
- GPU výpočty,
- Sdílené výpočetní uzly,
- Vyhrazené výpočetní uzly,
- Výpočetní uzly běžící na konkrétním OS, min. pro SUSE, Debian, Ubuntu a CentOS,
- Výpočetní uzly běžící v učebně,
- Fyzické výpočetní uzly,
- Virtuální výpočetní uzly,
- Výpočetní uzly podle CPU architektur
- Výpočetní uzly podle GPU architektur

Jednotliví uživatelé mohou plánovat své úlohy v rámci konkrétního dílčího výpočetního clusteru, nebo mohou spustit svůj výpočet paralelně napříč všemi výpočetními clustery tak, že se úloha provede na prvním dílčím výpočetním clusteru, který bude mít volné výpočetní zdroje a na ostatních výpočetních clusterech se úloha automaticky odebere z front úloh, čekajících na své zpracování.

Příloha č. 1_Specifikace prací

Realizace projektu

Realizace projektu proběhne min. v následujících krocích:

1) Vytvoření návrhu a konceptu řešení a jeho schválení,
2) PoC verze ClusterFIT prezentována ve VM na platformě CloudFIT, CPU i GPGPU výpočty,
3) Testovací verze ClusterFIT v rámci sady serverů a učebnových PC s GPU kartami, min. 1x řídící server, 1x frontend pro plánování úloh a spouštění interaktivních úloh, min. 4x výpočetní server s HW konfigurací typu A, min. 4x výpočetní server s HW konfigurací typu B, min. 2x GPGPU server s HW konfigurací typu C, kdy konfigurace A, B a C se můžou lišit velikostí RAM, počtem CPU a připojenými datovými úložišti a jsou provozovány na CPU architektuře x86-64 a GPU architektuře CUDA nad systémem OS openSUSE nebo s ním binárně kompatibilním. V rámci testovací verze se demonstruje spouštění a plánování interaktivních i neinteraktivní úloh. Jednotlivé úlohy budou mít následující charakter:
    a. Budou časově omezené,
    b. Budou časově neomezené,
    c. Budou prioritní,
    d. Budou obsahovat výpočty běžící na CPU i GPU min. pro architekturu CPU x86-64 a architekturu GPU CUDA,
    e. Budou obsahovat paralelní výpočty MPI,
4) Rozšíření testovací verze ClusterFIT z bodu (3) o systémové virtuální servery a monitoring služby,
5) Nasazení řešení ClusterFIT na vybrané testovací učebně a předání služby do provozu a správy ICT FIT včetně kompletní dokumentace, zaškolení obsluhy a zaučení ve způsobu rozšiřování služby přes další výpočetní systémy. V rámci této fáze se zrealizuje běh CPU a GPU výpočtů na učebnových PC jak v době mimo výuku, tak i v době výuky. Bude provedeno ladění funkčnosti, tak aby byl provoz umožněn bez negativního dopadu na probíhající výuku a s možností, že se učebnové PC může kdykoliv restartovat.

Příloha č. 1_Specifikace prací

Realizační výpočetní, systémové a komunikační zdroje pro ClusterFIT

Platforma se může sestávat zejména z následujících výpočetních, kapacitních a komunikačních zdrojů distribuovaných ve 3 lokalitách (oddělených budovách), propojených ethernetovou konektivitou s různou rychlostí:

- Výpočetní prostředky:
  - Počítače v učebnách
    - OS Windows 10, Linux Ubuntu,
    - CPU architektury x86-64 a ARM,
    - GPGPU typu CUDA,
  - Servery fyzické
    - OS Windows Server 2019, Linux SLES, Linux Debian, Linux Ubuntu,
    - CPU architektury x86-64, ARM a ppc64,
    - GPGPU typu CUDA,
  - Servery virtuální
    - OS Windows Server 2019, Linux SLES, Linux Debian, Linux Ubuntu,
    - CPU architektury x86-64,
    - GPGPU typu CUDA,
  - Uživatelské PC
    - OS Windows 10, Linux openSUSE, Linux Ubuntu, Linux Debian,
    - CPU architektury x86-64 a ARM,
    - GPGPU typu Intel, CUDA a RADEON,
  - Uživatelské notebooky
    - OS Windows 10, Linux openSUSE, Linux Ubuntu, Linux Debian,
    - CPU architektury x86-64 a ARM,
    - GPGPU typu Intel, CUDA a RADEON,

- Propojovací sítě typu ethernet s různou rychlostí sítí:
  - 1Gbps učebnová PC,
  - až 50Gbps fyzické servery,
  - až 25Gbps virtuální servery,
  - 1Gbps kancelářská PC,
  - do 1Gbps notebooky připojené bezdrátovým připojením WiFi; průměrně kolem 600Mbps, maximálně však 2Gbps,
  - Propoje mezi lokalitami od 2Gbps do 160Gbps,

- Datová disková úložiště:
  - Diskové pole, SSD s propojem SAN Fiber Channel 32Gbps,
  - Diskové pole, NLSAS s propojem SAN Fiber Channel 32Gbps,
  - Lokální NVMe, SSD, či magnetické HDD s propojením SATA

Pavel Suchánek

Digitálně podepsal Pavel Suchánek
Datum: 2021.11.23 09:07:41 +01'00'