1 ÚVOD

Pro realizaci požadované platformy nabízíme řešení postaveno a implementováno nad HPC řešením společnosti SchedMD (https://www.schedmd.com) – Slurm Workload Manager, https://slurm.schedmd.com používaný napříč celosvětovým spektrem clusterových aplikací používaných ve státní správě, laboratořích, či dokonce většině z TOP10 největších výpočetních systémů světa ze seznamu z TOP500 (https://www.top500.org/lists/top500/). Základní parametry nabízeného řešení v jeho základní formě, bez dodatečných konfiguračních úprav a modulů uvádíme v následujícím textu.

2 PŘEDSTAVENÍ TECHNOLOGIE

Slurm je open source, chybám odolný a vysoce škálovatelný systém správy clusteru a plánování úloh pro velké a malé clustery Linuxu. Slurm pro svůj provoz nevyžaduje žádné úpravy jádra systému a je relativně soběstačný. Jako správce zátěže clusteru má Slurm tři klíčové funkce. Za prvé, přiděluje uživatelům na určitou dobu výhradní a/nebo nevýhradní přístup ke zdrojům (výpočetním uzlům), aby mohli vykonávat práci. Za druhé, poskytuje rámec pro spouštění, provádění a monitorování práce (obvykle paralelní úloha) na sadě přidělených uzlů. Nakonec rozhoduje spory o zdroje řízením fronty práce. Volitelné pluginy lze použít pro účetnictví, pokročilé rezervace, plánování gangů (sdílení času pro paralelní úlohy), plánování zálohování, výběr zdrojů optimalizovaný podle topologie, limity zdrojů podle uživatele a sofistikované algoritmy pro prioritizaci více-faktorových úloh.

Slurm má k dispozici univerzální mechanismus zásuvných modulů pro snadnou podporu různých infrastruktur. To umožňuje širokou škálu konfigurací Slurm pomocí škálování přístupu přes tzv. stavební bloky. Mezi tyto pluginy patří např.:

- Accounting Storage: Primárně se používá k ukládání historických dat o zakázkách. Při použití SlurmDBD (Slurm Database Daemon) může také dodávat informace založené na limitech spolu s historickým stavem systémů.
- Account Gather Energy: Shromažďujte údaje o spotřebě energie na úlohu nebo uzly v systému. Tento plugin je integrován s pluginy Accounting Storage a Job Account Gather.
- Autentizace komunikace: Poskytuje ověřovací mechanismus mezi různými součástmi Slurm.
- Kontejnery : Podpora a implementace kontejnerů pracovních zátěží HPC.
- Pověření (Generování digitálního podpisu): Mechanismus používaný ke generování digitálního podpisu, který se používá k ověření, že krok úlohy je oprávněn provést uživatel pouze na konkrétních uzlech. To se liší od zásuvného modulu používaného pro ověřování, protože požadavek na krok úlohy je odeslán z příkazu uživatele, nikoli přímo z démona slurmctld, který generuje pověření kroku úlohy a její digitální podpis.
- Obecné zdroje: Poskytněte rozhraní pro ovládání generických zdrojů, včetně grafických procesorových jednotek (GPU).
- Odeslání úlohy : Vlastní plugin, který umožňuje specifickou kontrolu nad požadavky úlohy při odeslání a aktualizaci.
- Sdružování účetních informací: Shromažďujte data o využití zdrojů kroku úlohy.
- Protokolování dokončení úlohy: Protokolujte data o ukončení úlohy. Obvykle se jedná o podmnožinu dat uložených modulem Accounting Storage.
- Spouštěče: Řídí mechanismus používaný příkazem 'srun' ke spouštění úloh.
- MPI: Poskytuje mechanizmy pro různé implementace MPI. Můžete například nastavit proměnné prostředí specifické pro MPI.
- Preempt: Určuje, které úlohy mohou zabránit jiným úlohám, a mechanismus preempce, který se má použít.
- Priority: Přiřazuje úlohám priority při odeslání i průběžně (např. jak stárnou).
- Sledování procesů (pro signalizaci): Poskytuje mechanismus pro identifikaci procesů souvisejících s každou úlohou. Používá se pro účtování a signalizaci úloh.
- Plánovač: Plugin s možností konfigurace vlastního plánovače, který určuje, jak a kdy Slurm plánuje úlohy.
- Výběr uzlů: Plugin používaný k určení zdrojů použitých pro alokaci úlohy.
- Site factor (Priority) : Přiřazuje konkrétní komponentu multifaktorové priority úlohy k úlohám při odeslání i průběžně (např. jak stárnou).

4 KONFIGUROVATELNOST

Mezi sledované stavy uzlů patří: počet procesorů, velikost skutečné paměti, velikost dočasného místa na disku a stav (zapnutý, vypnutý, atd.). Mezi další informace o uzlu patří váhy (přednost při přidělování práce) a funkce (libovolné informace, jako je rychlost nebo typ procesoru). Uzly jsou seskupeny do oddílů, které mohou obsahovat překrývající se uzly, takže je lze nejlépe považovat za fronty úloh. Informace o oddílu zahrnují např. název, seznam přidružených uzlů, stav (zapnutý/vypnutý), maximální časový limit úlohy, maximální počet uzlů na úlohu, skupinový přístupový seznam, prioritu (důležité, pokud jsou uzly ve více oddílech) a sdílenou politiku přístupu k uzlu s volitelným překročením úrovně kvót pro plánování gangů. Bitové mapy je možné využít k reprezentaci uzlů a rozhodnutí o plánování lze provést provedením malého počtu porovnání a série rychlých manipulací s bitovou mapou.

    - 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