C2115
Praktický úvod do superpočítání

III. lekce

Petr Kulhánek

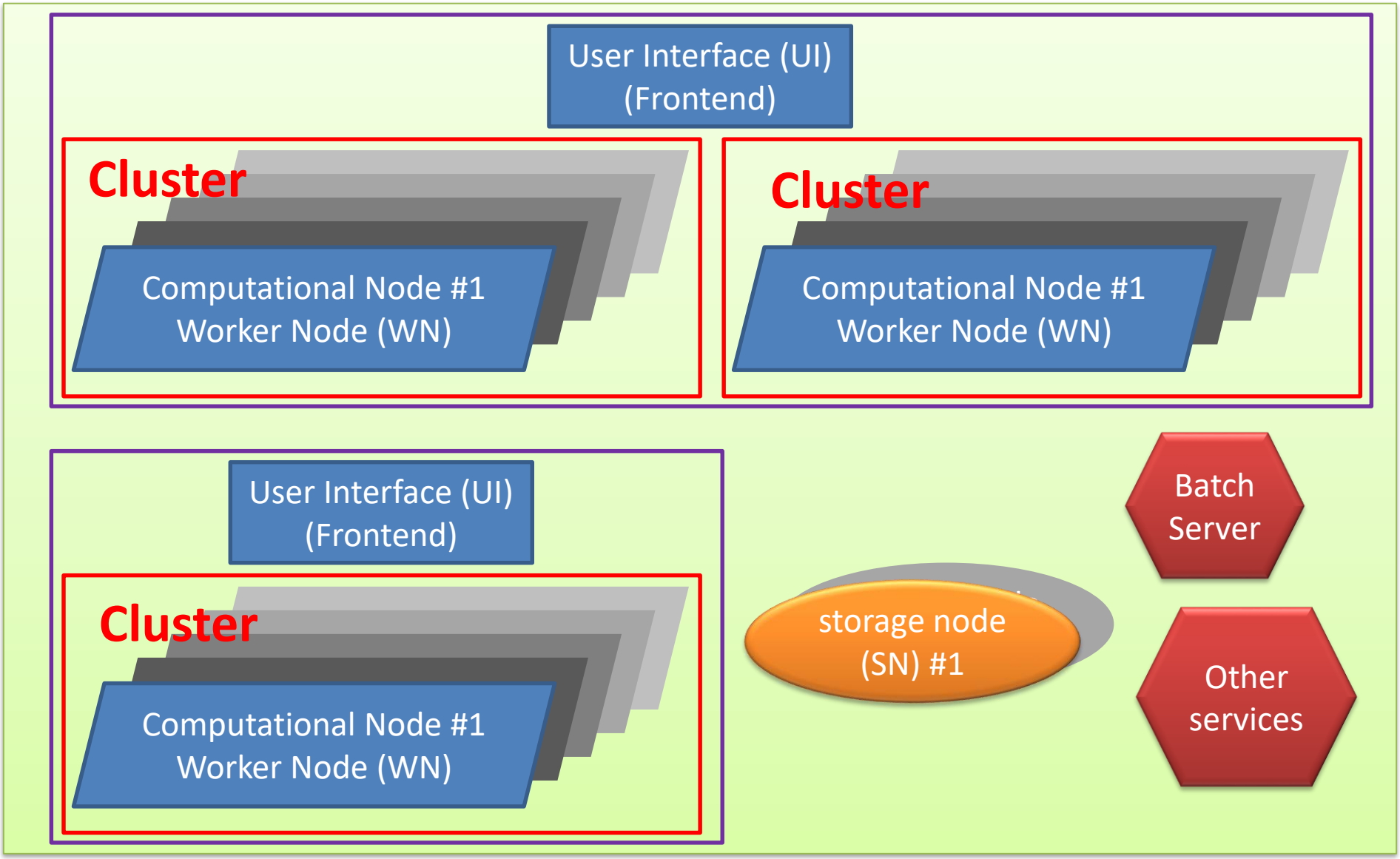
kulhanek@chemi.muni.cz

Národní centrum pro výzkum biomolekul, Přírodovědecká fakulta
Masarykova univerzita, Kamenice 5, CZ-62500 Brno

Obsah

- Architektura klastrů a (super)počítačů
  - Čelní uzly
  - Výpočetní uzly a elementy
  - Datové úložiště
  - Síťová infrastruktura

(Super)počítačová centra



Čelní uzel

Čelní uzel (**front-end node**, **user interface**) je počítač vyhrazený pro přímou interakci s uživatelem. Uživatel jej může použít pro přípravu vstupních dat úloh, **zadávání úloh do dávkového systému**, správu úloh a pro manipulaci s výsledky úloh (vizualizace).

Čelní uzel, pokud to není explicitně povoleno, by se **neměl používat pro spouštění CPU a paměťově náročných úloh**, případný pre-processing či post-processing dat úloh je nutné zadávat jako samostatné úlohy do dávkového systému.

U malých výpočetních klastrů je čelní uzel často zároveň i výpočetním uzlem.

Klastr či superpočítač většinou obsluhuje úlohy celé řady uživatelů, což vyžaduje přistupovat k čelnímu uzlu pouze vzdáleně. (Nehledě na to, že čelní uzel je typicky fyzicky přítomen v počítačovém sále, kde je poměrně velký hluk.).

Výpočetní uzel

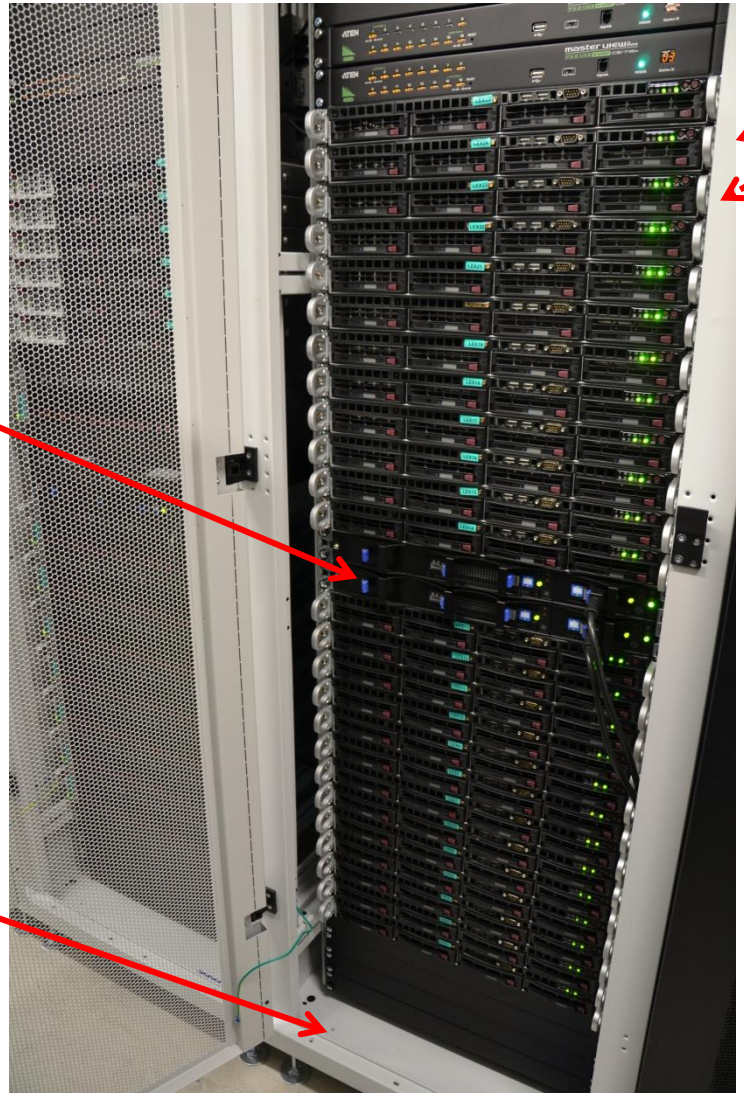
Výpočetní uzel (**worker node - WN, computational node**) je jednotka, která se chová jako samostatný počítač, který je vyhrazený pro řešení úloh uživatelů. Uzel může obsluhovat několik úloh současně. Počet úloh by však neměl překročit výpočetní zdroje (CPU, RAM, HDD), které tento uzel poskytuje. O efektivní využívání výpočetních zdrojů se stará OS (operační systém) ve spojení s **dávkovým systémem**.

Výpočetní element (**computational element** - **CE**), velmi často též nazýván jako **klastr**, je uskupení výpočetních uzlů nejčastěji se stejnou architekturou (homogenní klastr). Tyto uzly jsou většinou spojeny velmi rychlou lokální sítí (ethernet 1 Gbs, Infiniband, či proprietárním řešením). Kromě výše uvedeného může jedna úloha běžet na více výpočetních uzlech.

Úloha může běžet i na uzlech, které jsou v různých CE. Toto je však vhodné jen pro speciální typ úloh.

Výpočetní element / uzel

Klastr LEX
(výpočetní element)

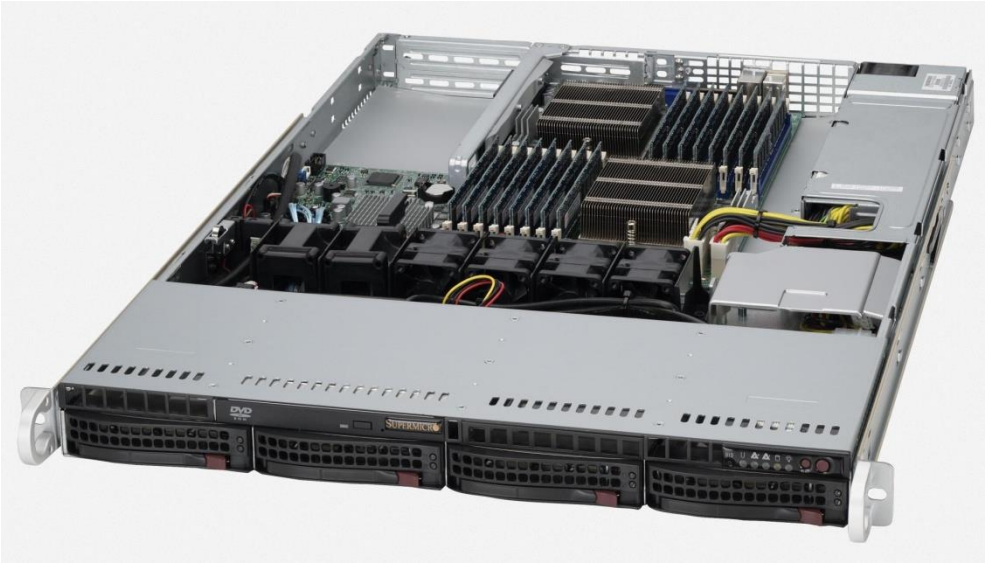


výpočetní uzly

Řadiče Infinibandu

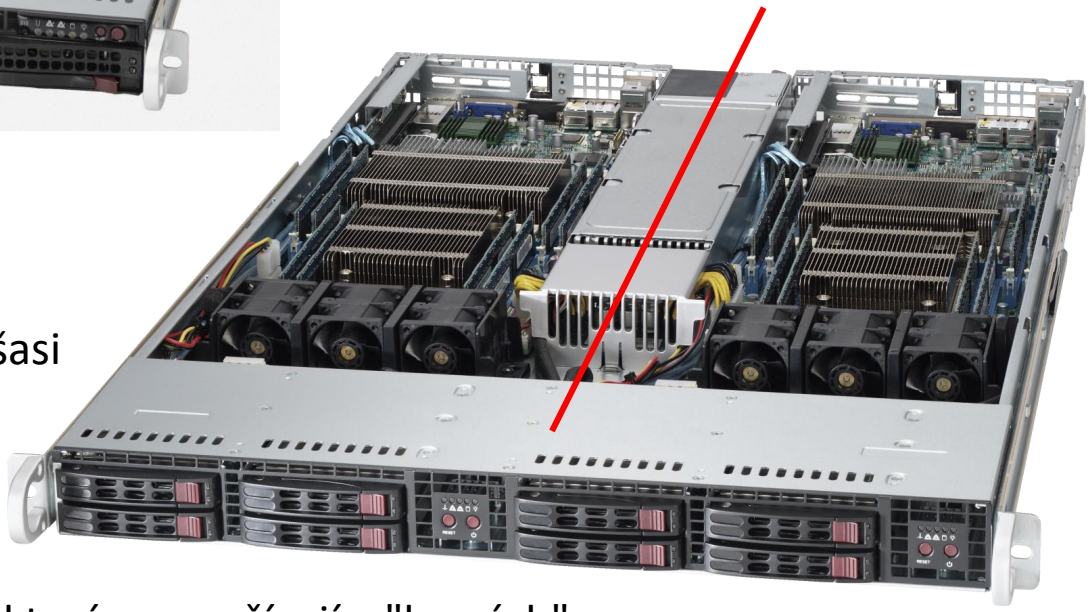
skříň (rack)

Výpočetní uzel



1U výška = 1,75 palce (44.45 mm)

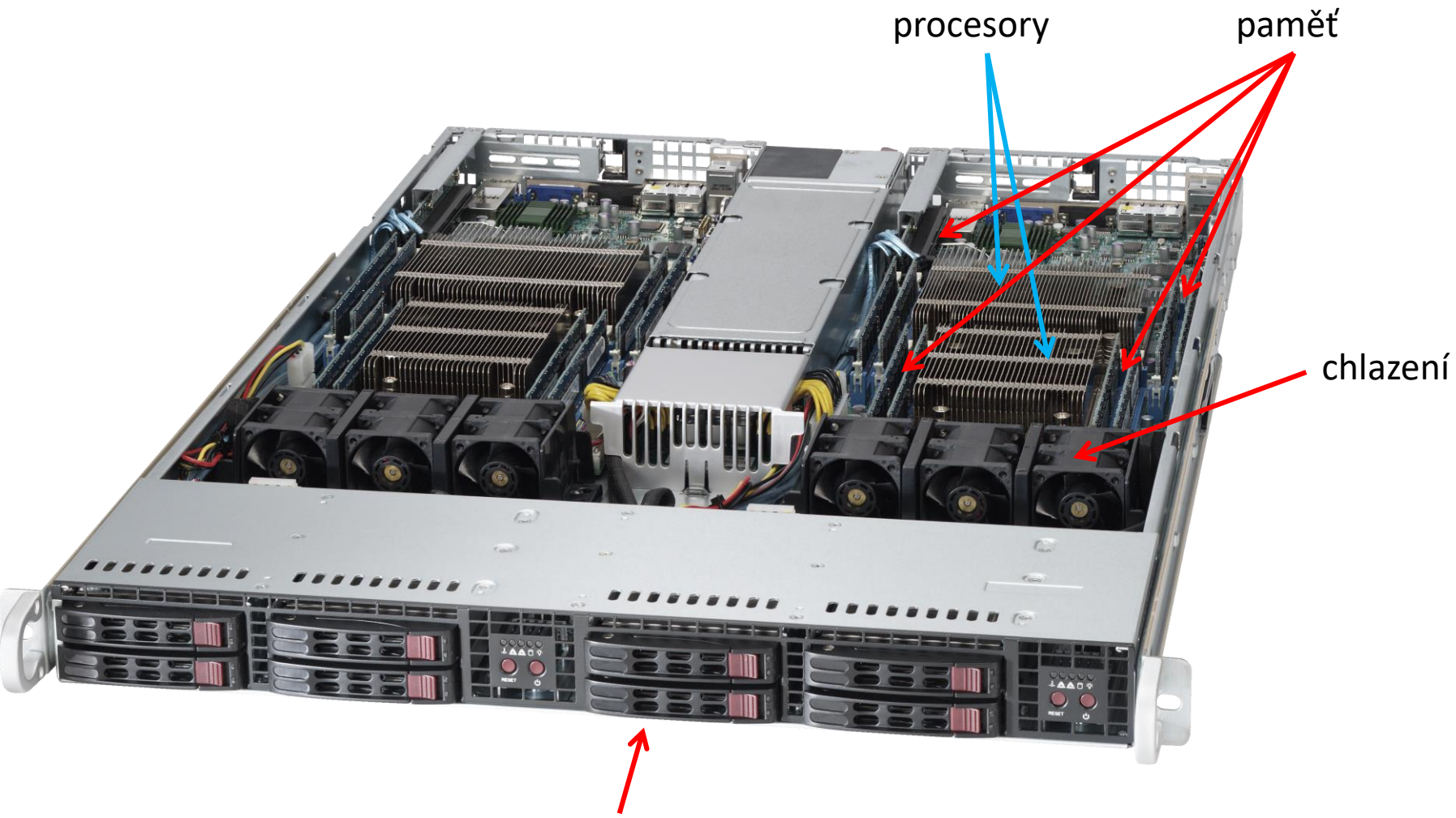
jeden počítač v jednom šasi



twin – dva počítače v jednom šasi

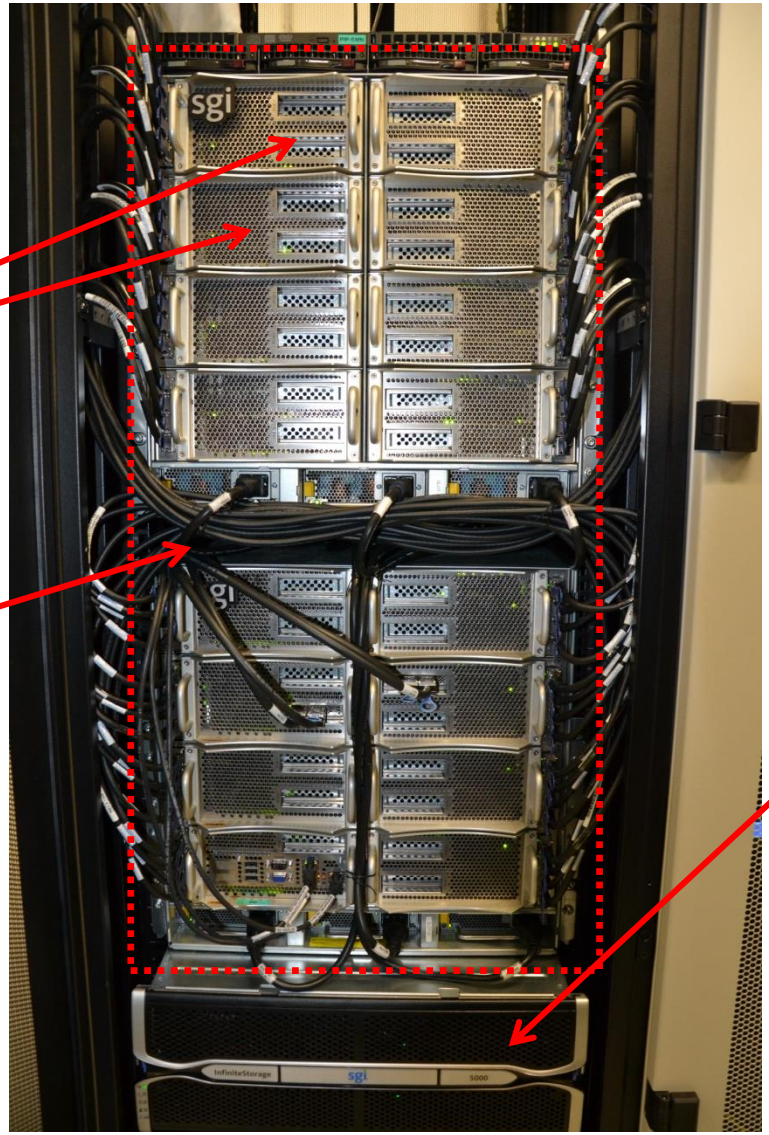
ukázky typických výpočetních uzlů, které se používají v "levných" klastrech – v superpočítačích se většinou používá proprietární řešení

Výpočetní uzel



Výpočetní uzel

SGI UV2000
pip



blades

sběrnice

jeden výpočetní uzel
192 CPU jader, 4 TB paměti

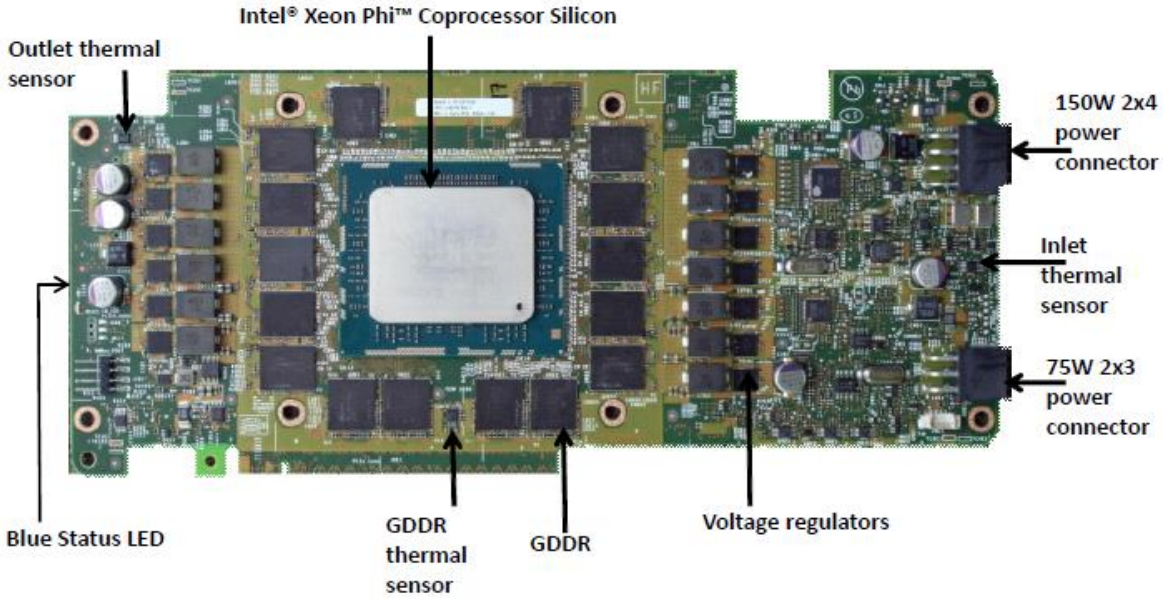
diskové pole – lokální datové úložiště

Výpočetní uzel - akcelerátory

NVidia Tesla K20 (GPGPU)

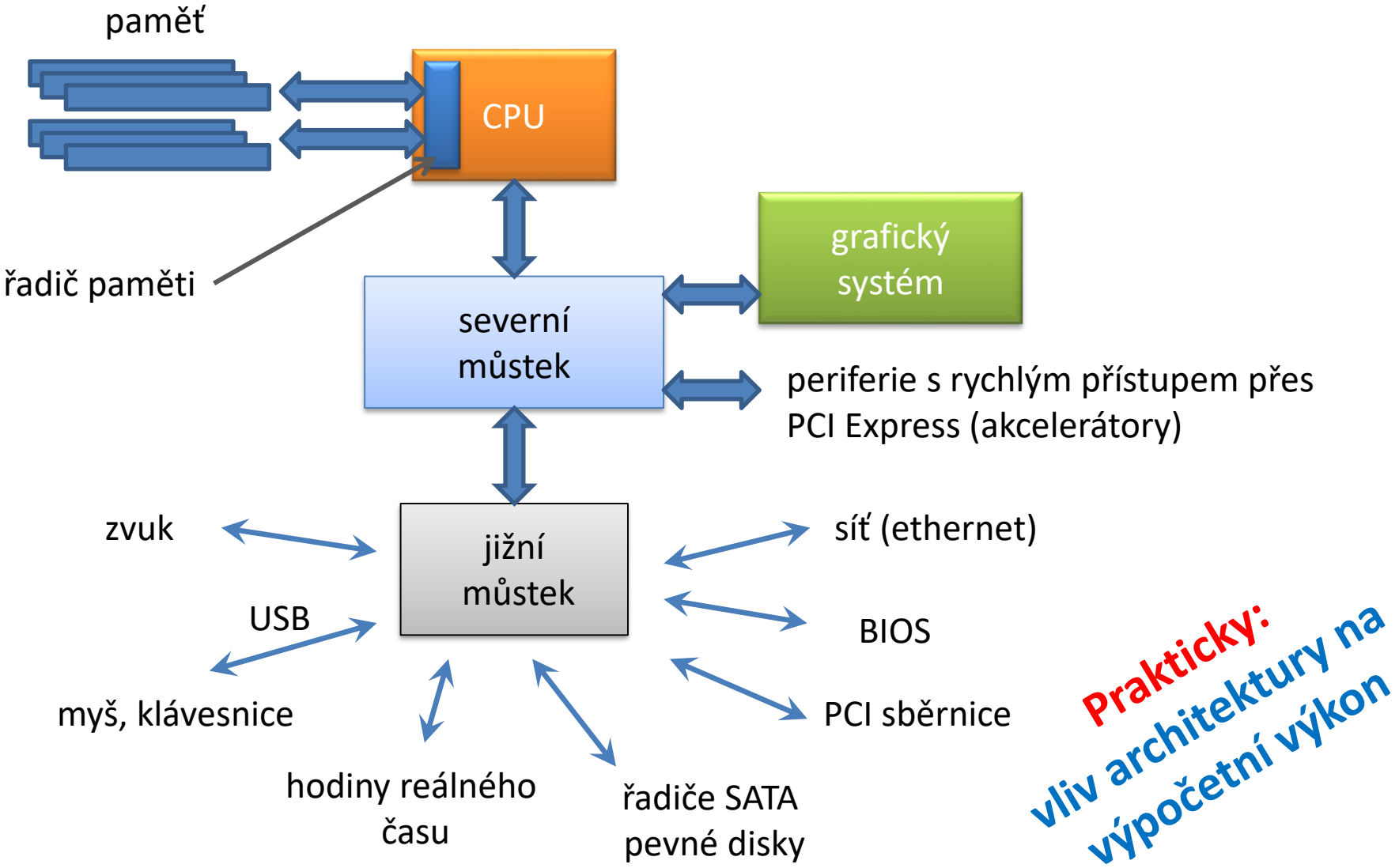
Prakticky:
MD simulace na GPU

Intel Xeon Phi (MIC)

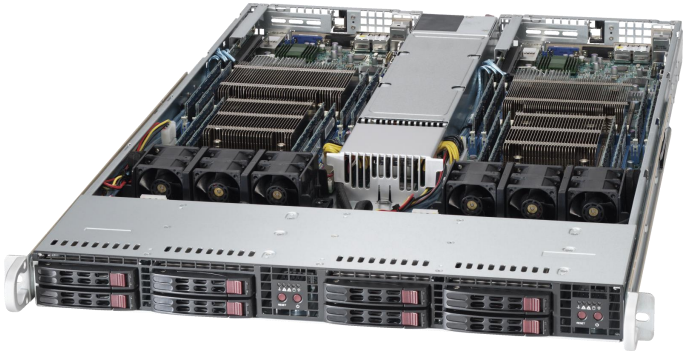


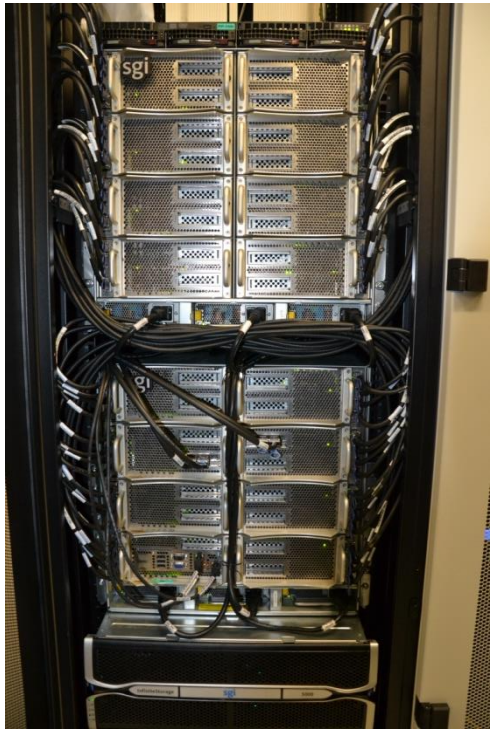
Výpočetní výkon akcelerátorů může přesáhnout výkon instalovaných CPU na výpočetním uzlu.

Typické schéma počítače



Víceprocesorové uzly





Výpočetní uzly v dnešní době obsahují **více fyzické procesorů** (minimálně dva), kdy každý obsahuje **více výpočetních CPU jader**. Operační paměť je pak většinou přístupná s různou rychlostí (**NUMA Non-Uniform Memory Architecture**).

Důvodem pro toto uspořádání je **navyšování výpočetního výkonu**, což však sebou přináší zvýšené nároky na přípravu a spouštění výpočetních úloh.

Prakticky: paralelizace úloh, spouštění paralelních úloh, limitace

Datové úložiště - dělení

Typy úložišť a jejich použití:

- **lokální datové úložiště** – dočasné data úloh
- **(vzdálená) datové úložiště (diskové pole)** – živá data úloh či řešených projektů
- **hierarchické datové úložiště** – ukončené projekty a zálohy

Prakticky:
datové úložiště MetaCentra

wikipedia.org

Lokální datové úložiště

- **Diskové pole** připojená lokálně k výpočetnímu uzlu.
- **HDD - pevný disk** (**Hard Disk Drive**) je zařízení, které se používá v počítačích k trvalému uchovávání většího množství dat pomocí magnetické indukce.
- **SSD - Solid-State Drive** je v informačních technologiích typ datového média, které na rozdíl od magnetických pevných disků neobsahuje pohyblivé mechanické části a má mnohem nižší spotřebu elektrické energie.

Lokální dočasné úložiště (scratch adresáře) jsou určené pro aktuálně běžící úlohy na výpočetním uzlu.

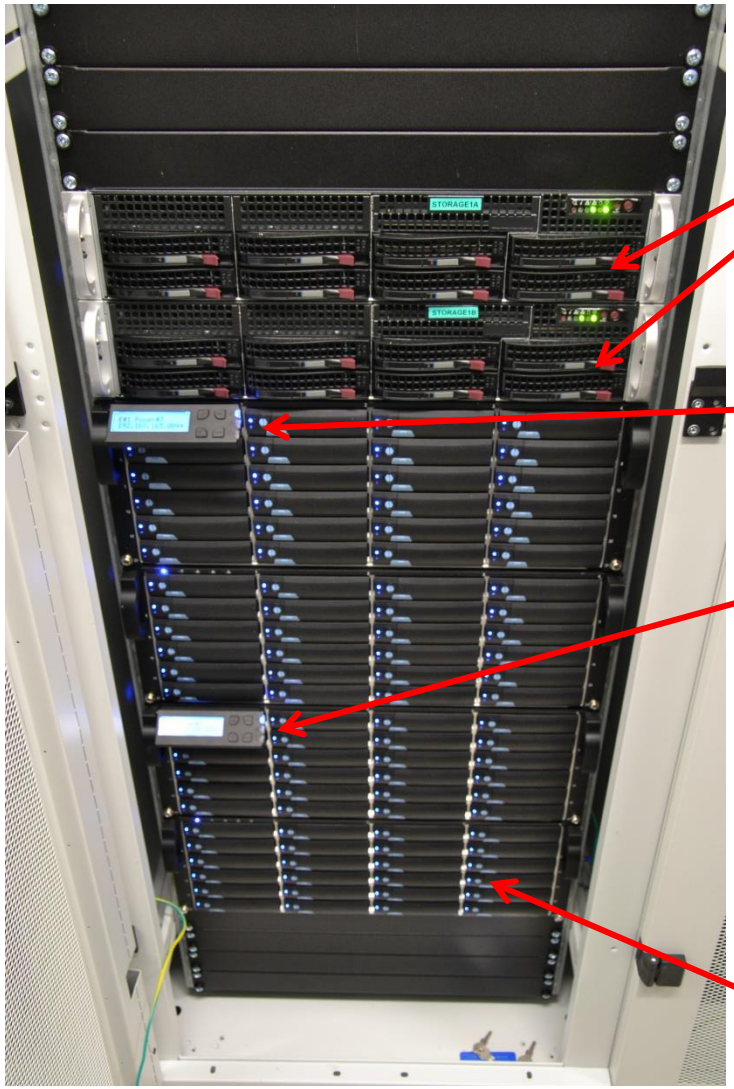
Tyto adresáře se NESMÍ* používat pro dlouhodobé ukládání dat.

*) samozřejmě můžete, ale pak se nedivte, že je jednoho krásného dne nenaleznete, protože administrátor, či jiný inteligentní nástroj, úložiště pročistil

wikipedia.org

Diskové pole

brno9-ceitec 269 TiB



souborové servery zpřístupňující data diskového pole vzdáleně přes NFS (**Network File System**) protokol

diskové pole RAID6

diskové pole RAID6

RAID 0

velké množství HDD

Disková pole jsou vhodná pro aktuálně řešené projekty.

Diskové pole – ochrana dat

Diskové pole obsahují velké množství HDD, což jsou mechanické komponenty, které jsou náchylné k selhání. Pro omezení poškození dat jsou data nejčastěji chráněna pomocí techniky RAID.

RAID (**Redundant Array of Inexpensive/Independent Disks**) je v informatice metoda zabezpečení dat proti selhání pevného disku. Zabezpečení je realizováno specifickým ukládáním dat na více nezávislých disků, kdy jsou uložená data zachována i při selhání některého z nich. Úroveň zabezpečení se liší podle zvoleného typu RAID, které je označováno čísly (nejčastěji RAID 0, RAID 1, RAID 5 či nověji RAID 6).

Při poškození části diskového pole běží pole v **degradovaném režimu**, kdy již další selhání by bylo neopravitelné. Proto jsou v diskových polích vyhrazeny tzv. spare disky, které se ihned použiji jako náhrada za poškozené. Po dobu **rebuildování diskového pole** (nový výpočet parity dat) může být přístupová rychlost k datům snížena.

wikipedia.org

Hierarchické datové úložiště

Hierarchical storage management (**HSM**) is a data storage technique, which automatically moves data between high-cost and low-cost storage media. While it would be ideal to have all data available on high-speed devices all the time, this is prohibitively expensive for many organizations. Instead, HSM systems store the bulk of the enterprise's data on slower devices, and then copy data to faster disk drives when needed. The HSM system monitors the way data is used and makes best guesses as to which data can safely be moved to slower devices and which data should stay on the fast devices.

brno10-ceitec-hsm
(páskový robot)

HSM úložiště jsou vhodná pro archivaci a zálohu dat.

wikipedia.org

Jednotky

Multiples of bytes V · T · E

Decimal			Binary				
Value	Metric		Value	IEC		JEDEC	
1000	kB	kilobyte	1024	KiB	kibibyte	KB	kilobyte
1000^2	MB	megabyte	1024^2	MiB	mebibyte	MB	megabyte
1000^3	GB	gigabyte	1024^3	GiB	gibibyte	GB	gigabyte
1000^4	TB	terabyte	1024^4	TiB	**tebibyte**	–	
1000^5	PB	petabyte	1024^5	PiB	pebibyte	–	
1000^6	EB	exabyte	1024^6	EiB	exbibyte	–	
1000^7	ZB	zettabyte	1024^7	ZiB	zebibyte	–	
1000^8	YB	yottabyte	1024^8	YiB	yobibyte	–	

Orders of magnitude of data

původní značení

wikipedia.org

Síťová infrastruktura

Ethernet (česky i eternet) je název souhrnu technologií pro lokální počítačové sítě (LAN), které používají kabely s kroucenou dvoulinkou, optické kabely pro komunikaci přenosovými rychlostmi od 10 Mbit/s po 100 Gbit/s.

InfiniBand (abbreviated IB), a computer-networking communications standard used in high-performance computing, features **very high throughput and very low latency**. It is used for data interconnect both among and within computers. InfiniBand is also utilized as either a direct, or switched interconnect between servers and storage systems, as well as an interconnect between storage systems.

Infiniband je vhodné použít pro datově náročné paralelní úlohy, které využívají více výpočetních uzlů.

wikipedia.org

Dávkový systém

Dávkové zpracování je vykonávání série programů (tzv. dávek) na počítači bez účasti uživatele. Dávky jsou připraveny předem, takže mohou být zpracovány předány bez účasti uživatele. Všechna vstupní data jsou předem připravena v souborech (skriptech) nebo zadána pomocí parametrů na příkazovém řádku. Dávkové zpracování je opakem interaktivního zpracování, kdy uživatel až teprve za běhu programu poskytuje požadované vstupy.

Výhody dávkového zpracování

- sdílení zdrojů počítače mezi mnoha uživateli a programy
- odložení zpracování dávek do doby, kdy je počítač méně vytížen
- odstranění prodlev způsobeným čekáním na vstup od uživatele
- maximalizace využití počítače zlepšuje využití investic (zejména u dražších počítačů)

Naše lokální klastry, MetaCentrum: PBSPro
IT4I: PBSPro

PBSPro je odvozen z OpenPBS.

Prakticky: PBSPro

wikipedia.org

Cvičení 1

1. Jaké jméno má vaše pracovní stanice (počítač) na klastru WOLF?
2. Jakou roli má tento počítač v rámci klastru WOLF?
3. Z dokumentace zjistěte jména čelních uzlů virtuální organizace MetaCentrum.
4. Ověřte, že se můžete přihlásit na jeden z čelních uzlů MetaCentra.
5. Kolik pevných disků může selhat ve skupině disků, které jsou chráněny pomocí RAID6?
6. Může sloužit RAID0 k ochraně dat?
7. Jak se označuje kombinace RAID6 a RAID0?
8. Jaký typ akcelerátoru je využit v superpočítači salomon (IT4I)?