

# C2115

# Praktický úvod do superpočítání

III. lekce

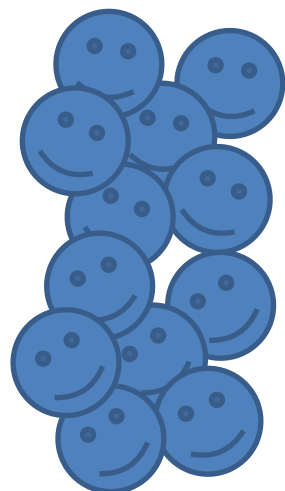
Petr Kulhánek

[kulhanek@chemi.muni.cz](mailto:kulhanek@chemi.muni.cz)

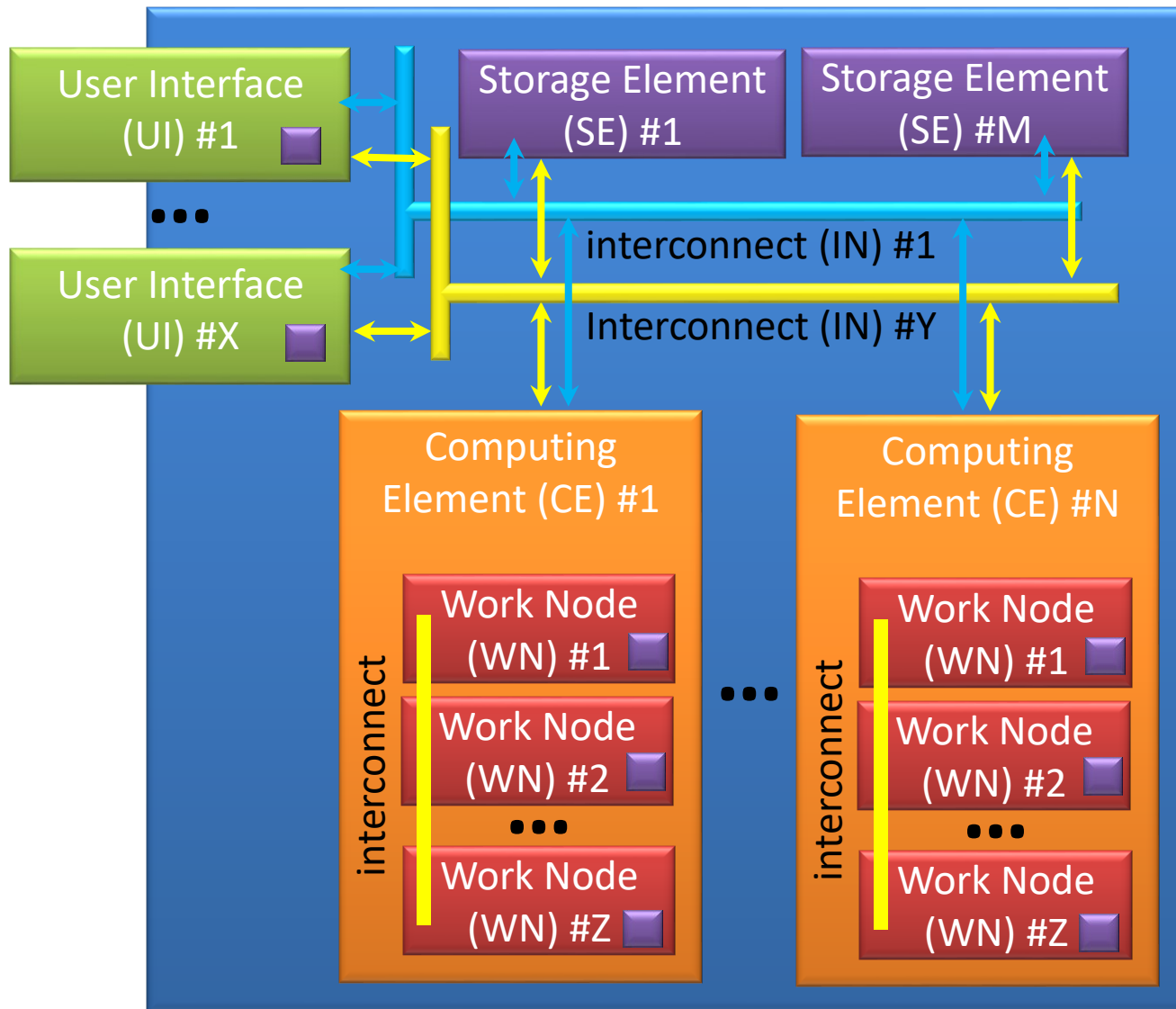
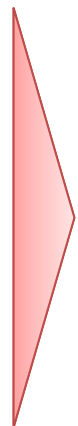
Národní centrum pro výzkum biomolekul, Přírodovědecká fakulta  
Masarykova univerzita, Kamenice 5, CZ-62500 Brno

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

# Klíčové části

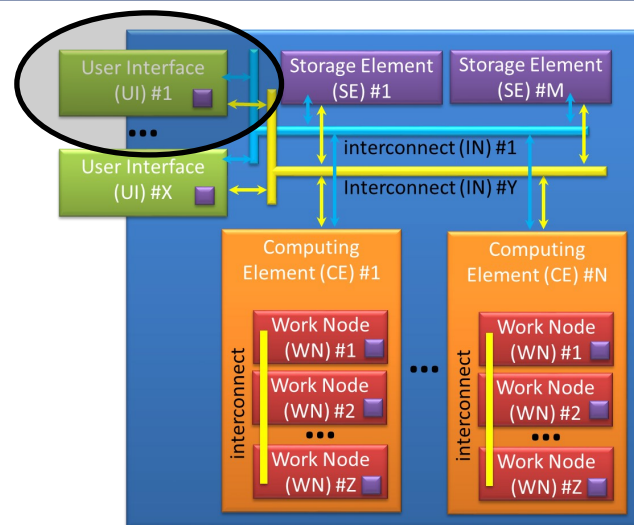


uživatelé



# Čelní uzel (UI)

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 spuštění 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.

Klaster 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.).

# Vzdálený přístup - čelní uzel (UI)

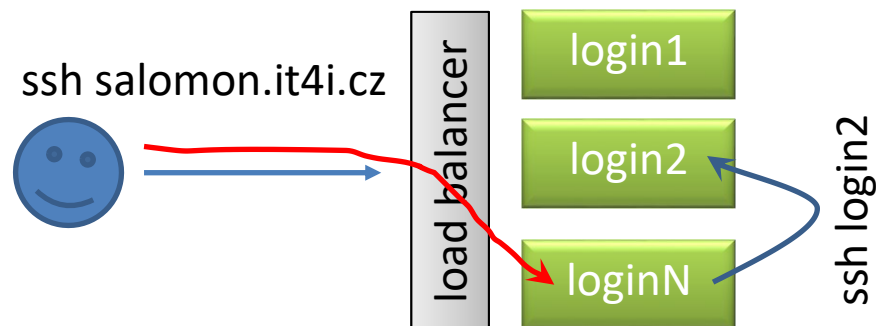
**Čelní uzel** (frontend) je počítač, který zprostředkovává interakci se superpočítačem.

- UI obvykle nabízí pouze příkazovou řádku (**CLI** - command line interface) pomocí zabezpečeného spojení protokolem ssh (secure shell).
- Pokud je dostupné grafické rozhraní (**GUI** - graphical interface) je vhodnější použít vzdálenou plochu (VNC - virtual network connection) než export X11 displeje.

## přímý přístup



## nepřímý přístup



**Na UI se NESPOUŠTÍ výpočetně náročné úlohy!!!!**

# Síťová infrastruktura

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

**InfiniBand** (zkráceně IB), je počítačový síťový komunikační standard používaný ve vysoce výkonných počítačích. IB se se vyznačuje **velmi vysokou propustností a velmi nízkou latencí**. Používá se pro datové propojení mezi počítači i uvnitř počítačů. InfiniBand se také používá buď pro přímé nebo přepínané propojení mezi servery a úložnými systémy, jakožto i jako propojení mezi úložnými systémy.

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

# Dávkový systém

**Dávkové zpracování** je vykonávání série programů (tzv. dávek) na počítači bez účasti uživatele. Dávky jsou připraveny předem, proto mohou být zpracovány bez účasti uživatele. Všechna vstupní data jsou předem připravena v souborech (skriptech) nebo zadána pomocí parametrů. 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

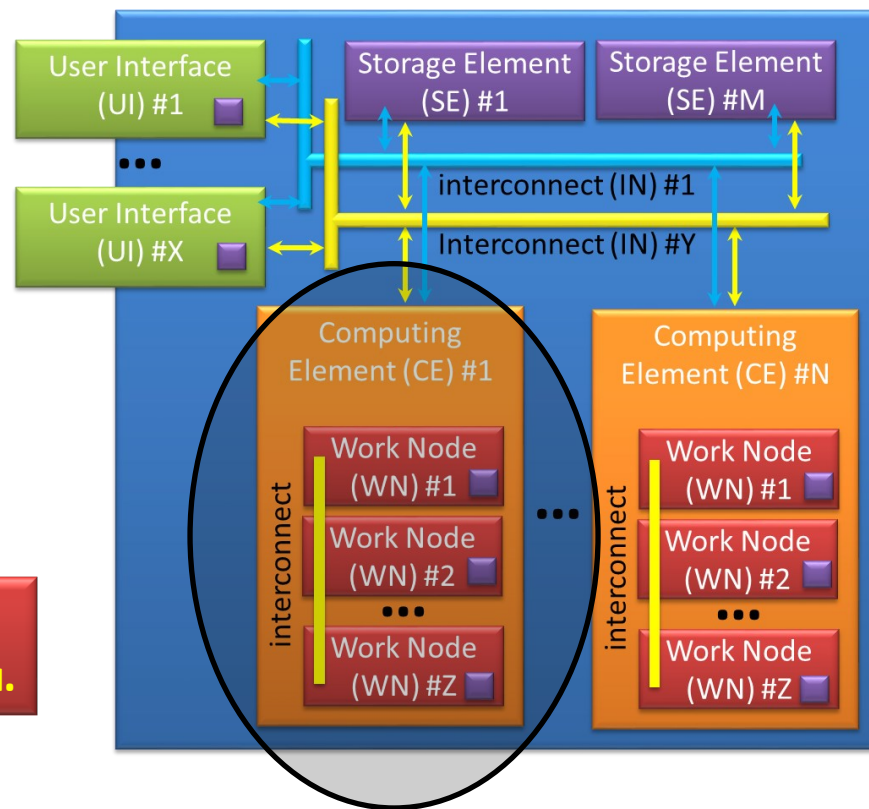
# Výpočetní elementy (CE) a výpočetní uzly (WN)



# Výpočetní element (CE)

**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).

Výpočetní uzel nebo počítačový uzel jsou synonymy pro pracovní uzel (**WN - worker node**).



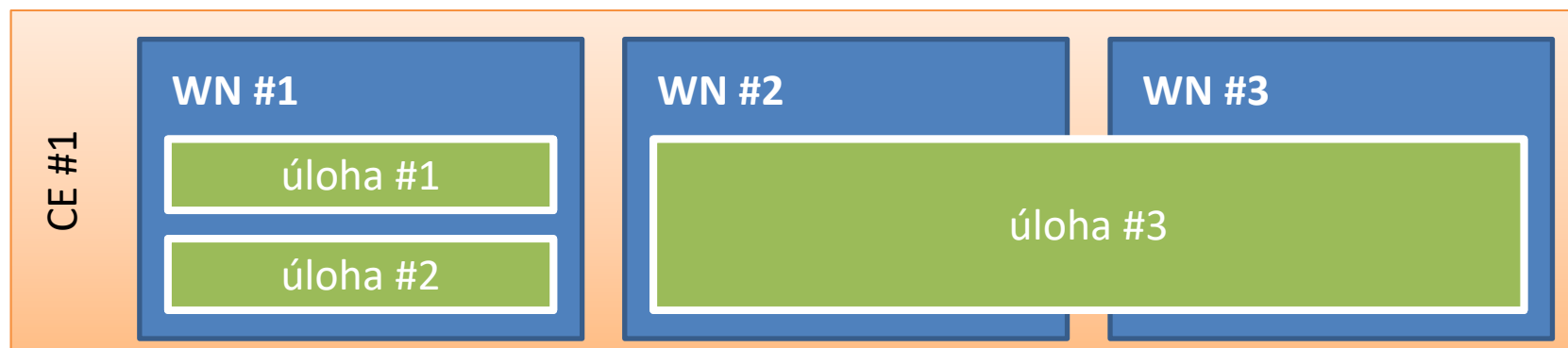
**Přímé používání CE and WN je zakázáno.  
Úlohy musí být zasílány do dávkového systému.**

# Výpočetní uzel (WN)

**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. Jedna úloha může využívat více výpočetních uzlů. 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**.

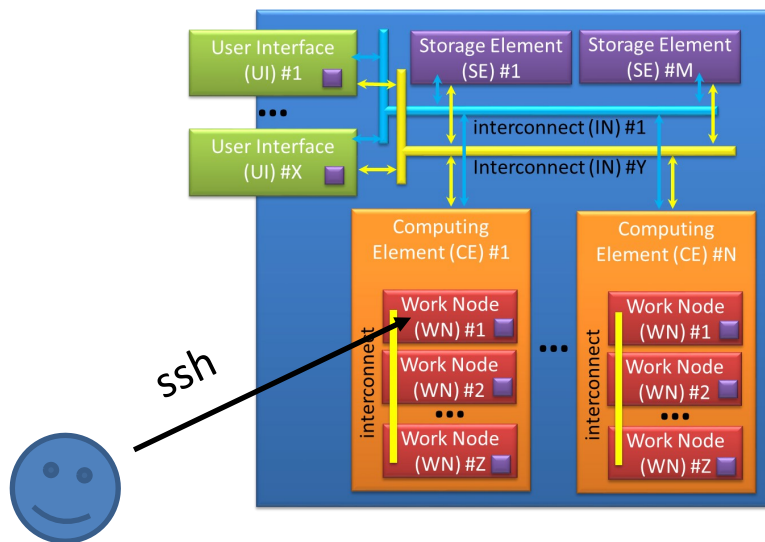
Úloha může běžet i na uzlech, které jsou v různých CE. Toto je však vhodné jen pro speciální typ úloh.

Možné uspořádání úloh na WN a CE:



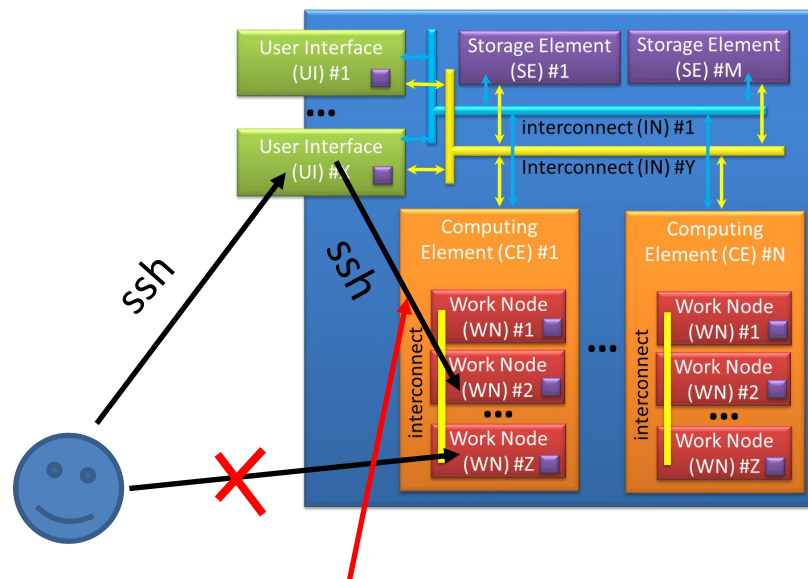
# Přístup na WN - pouze monitoring

## přímý přístup



MetaCentrum,  
všechny NCBR/CEITEC MU klastry

## nepřímý přístup



IT4I

Přístup je omezen pouze na  
uzly, kde uživateli běží úlohy.

**Tento přístup je možné použít pouze pro monitorování  
běžících úloh.**

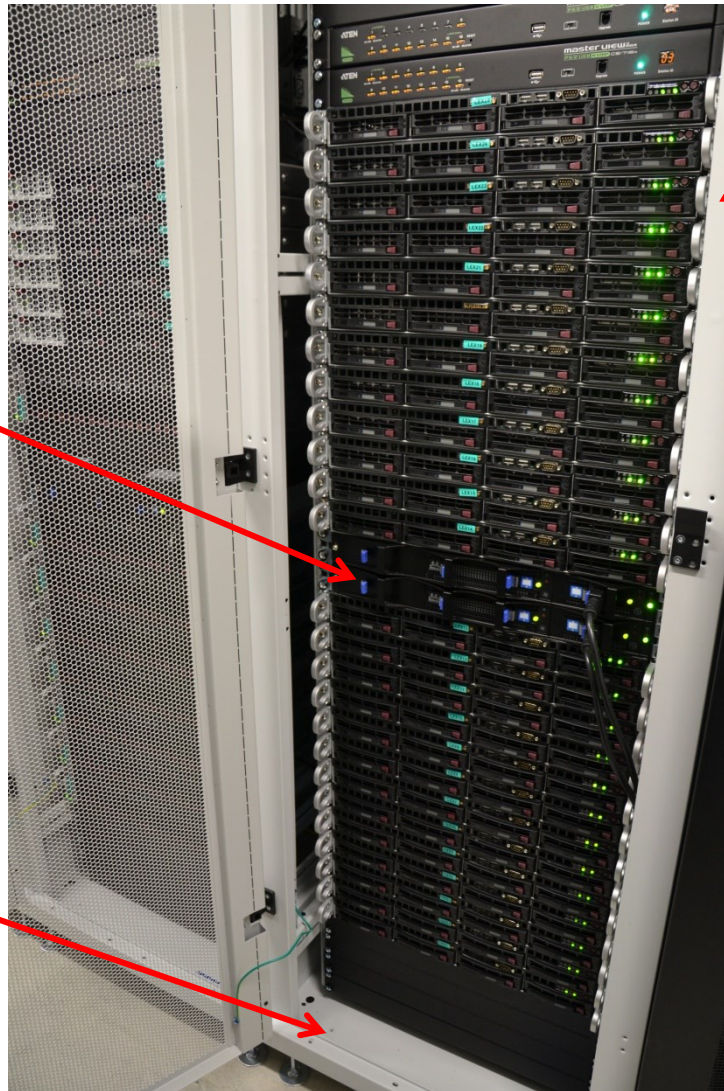
**NESMÍ se použít pro vlastní spuštění úloh**

# Výpočetní element / uzel

Klastr LEX  
(výpočetní element)

Řadiče Infinibandu

skříň (rack)

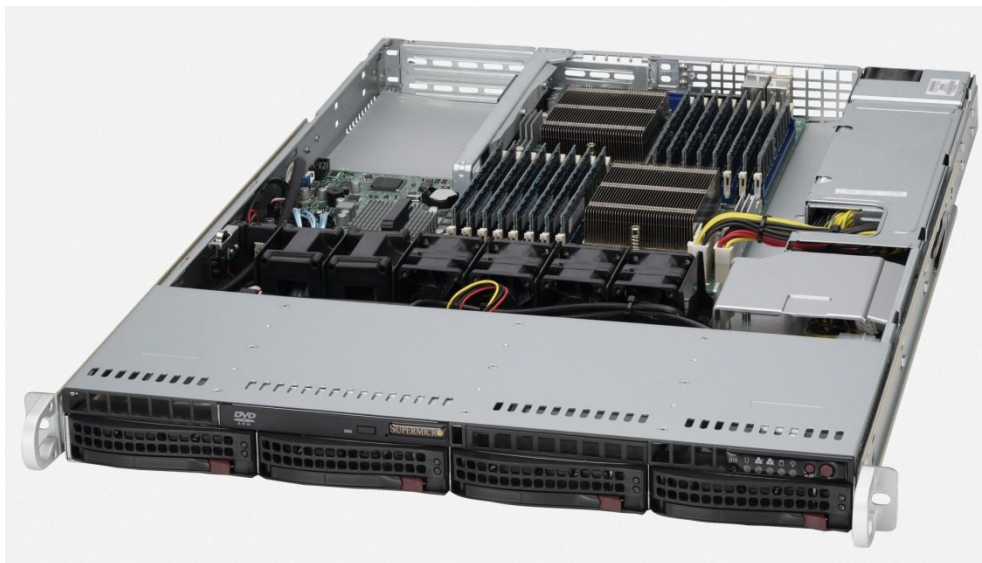


výpočetní uzly

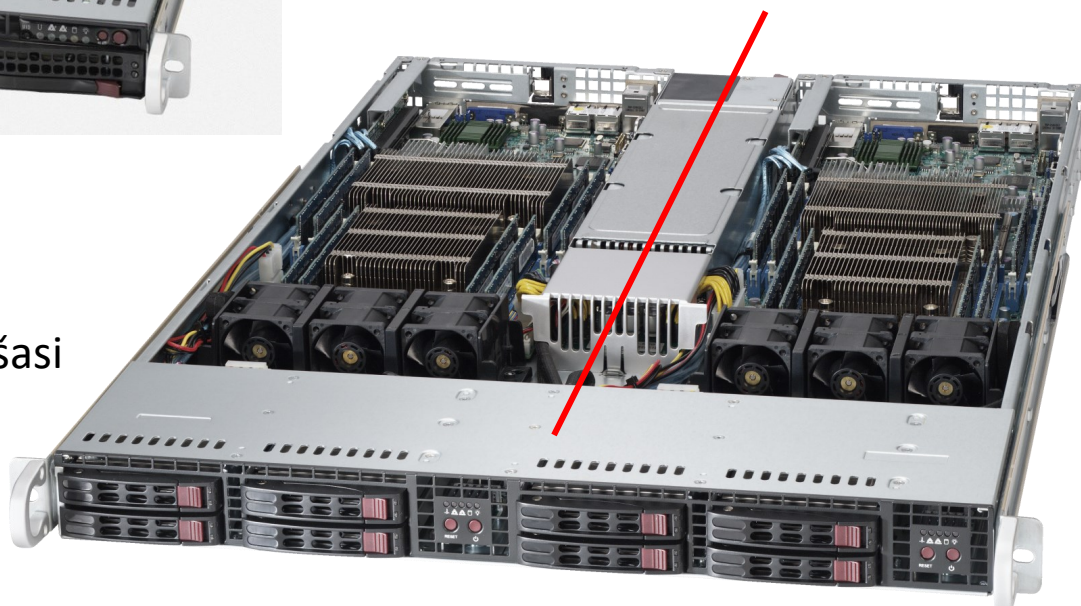
# Výpočetní uzel

1U výška = 1,75 palce (44.45 mm)

jeden počítač v jednom šasi

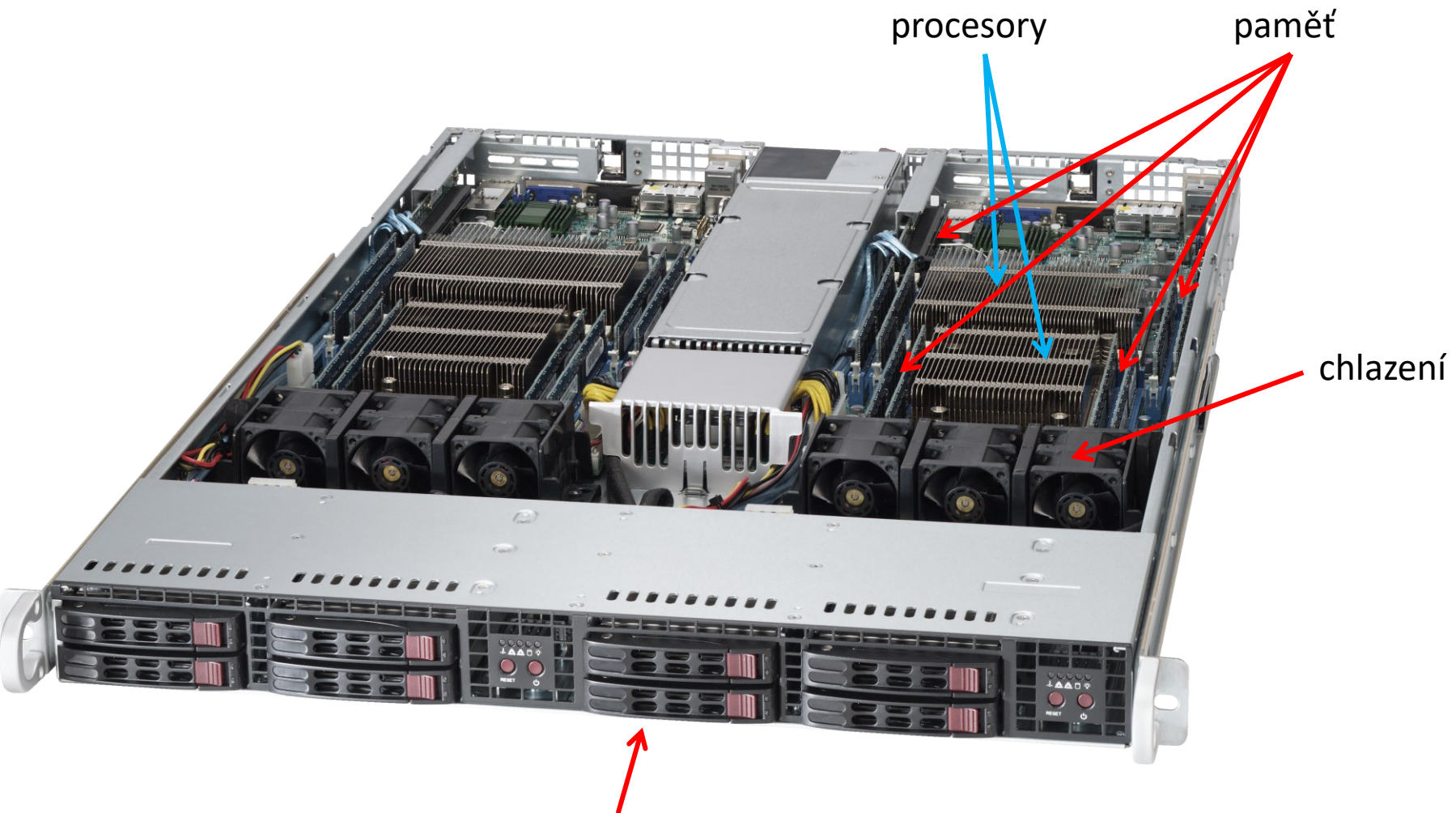


twin – dva počítače v jednom šasi



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

# Výpočetní uzel



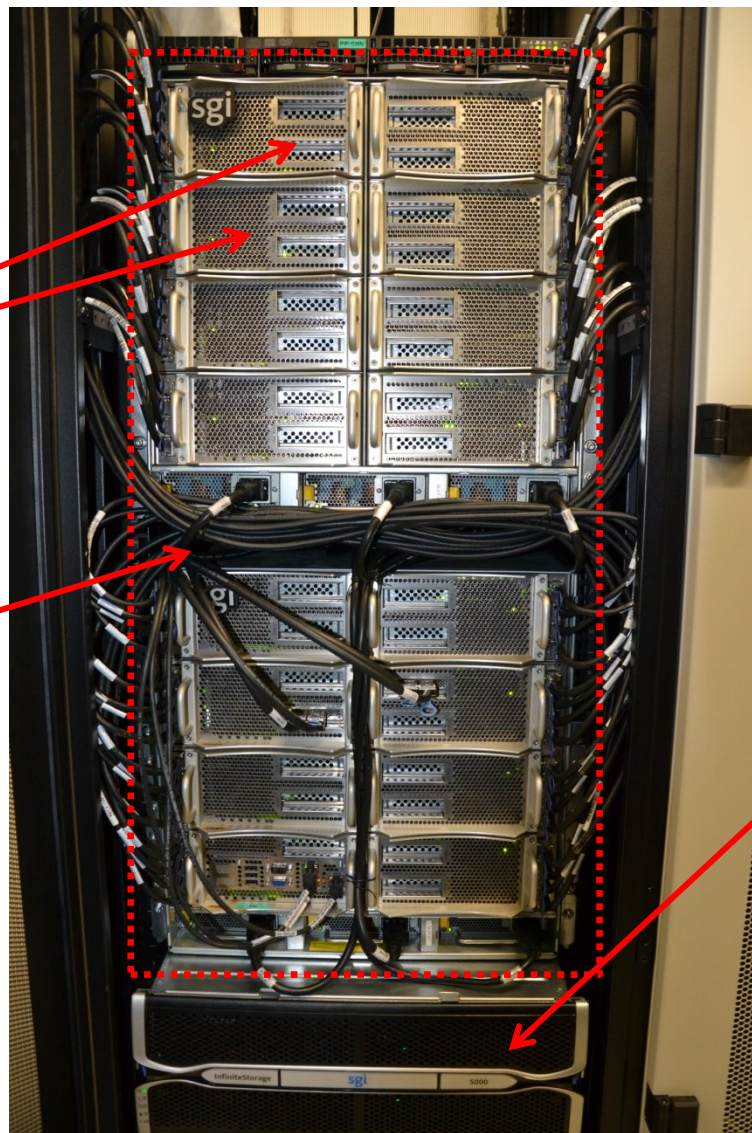
disky – lokální datové úložiště

# Výpočetní uzel

SGI UV2000  
pip

blades

sběrnice

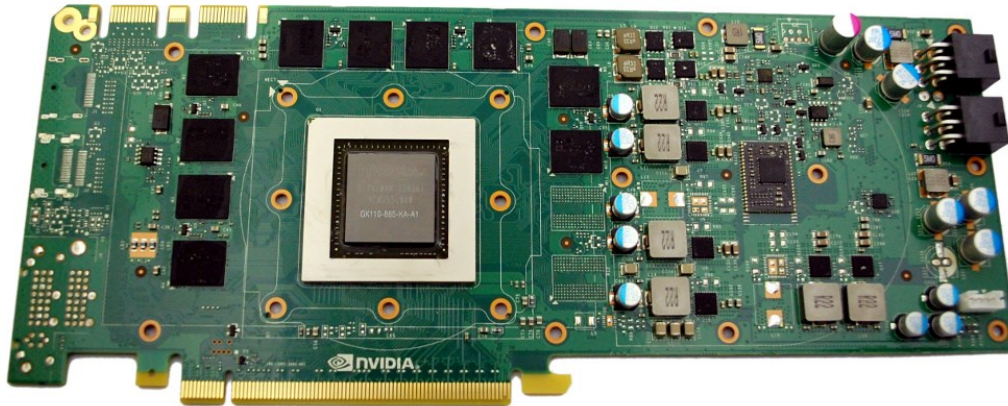


**jeden** výpočetní uzel  
192 CPU jader, 4 TB paměti

diskové pole – lokální  
datové úložiště

# Výpočetní uzel - akcelerátory

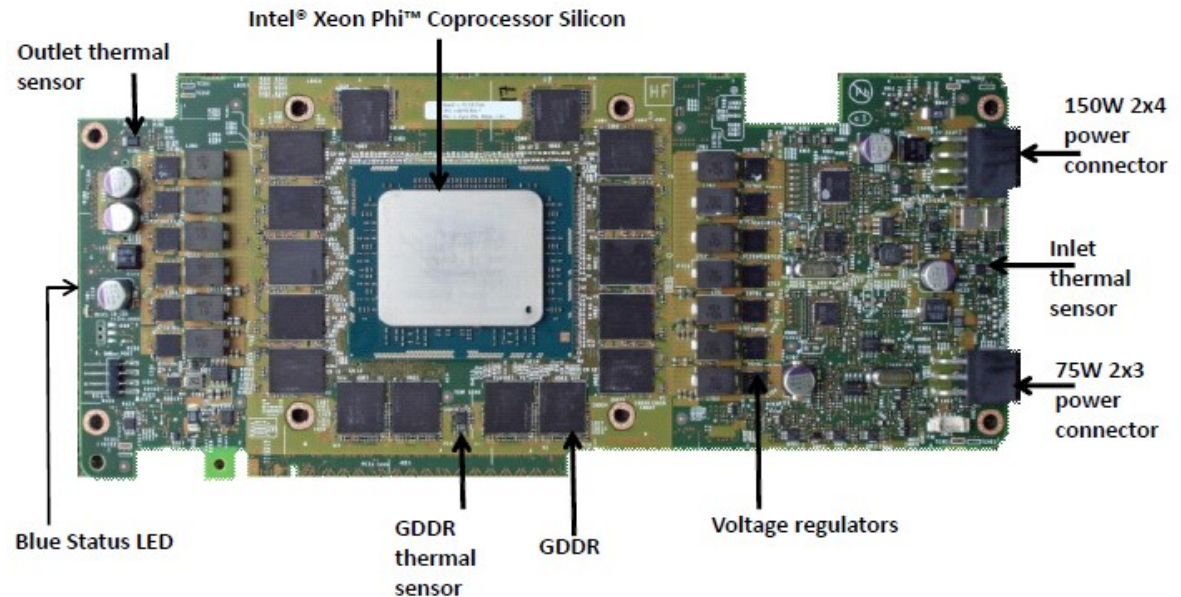
## NVidia Tesla K20 (GPGPU)



**Prakticky:**  
MD simulace na GPU

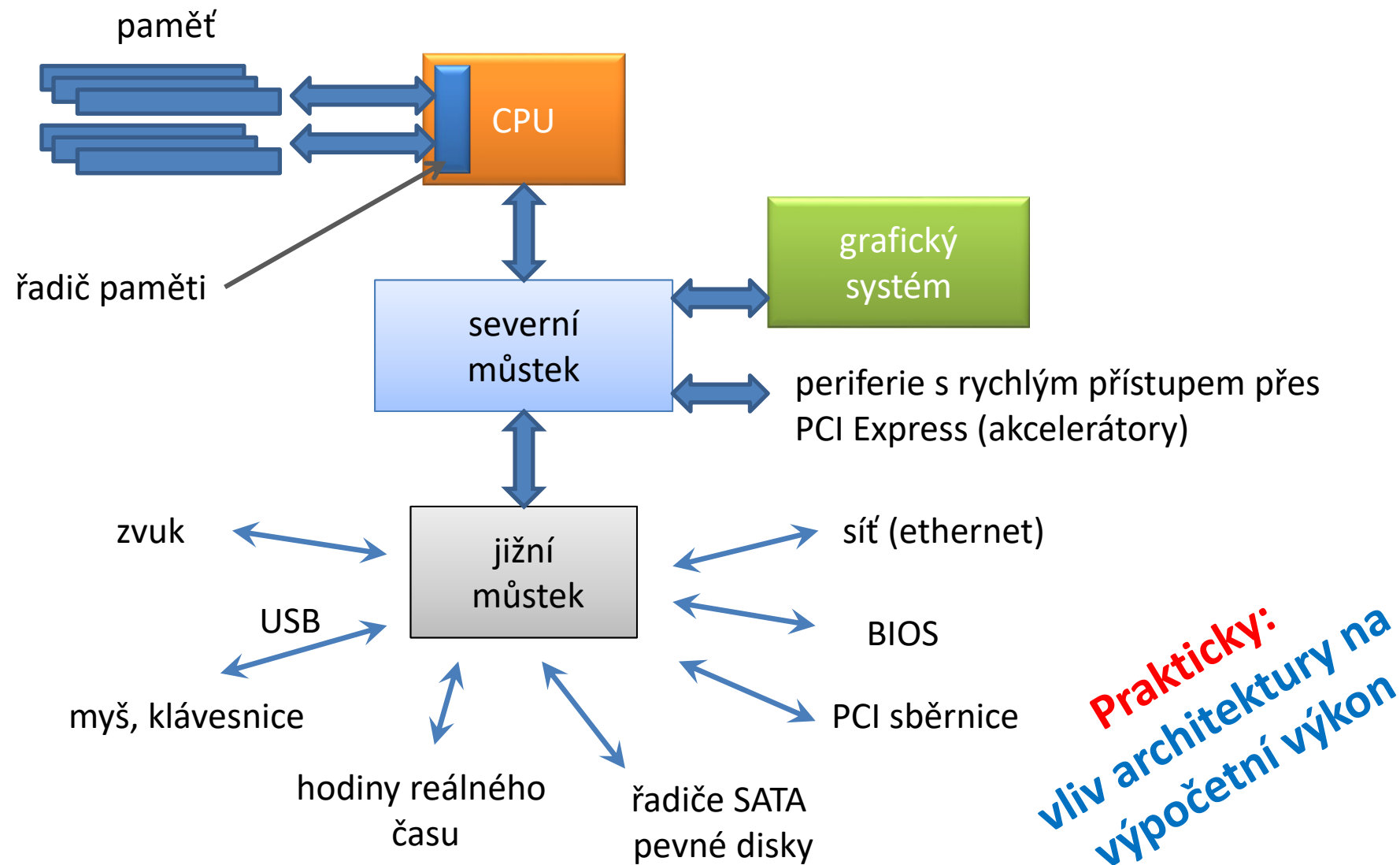
Výpočetní výkon akceleratorů může přesáhnout výkon instalovaných CPU na výpočetním uzlu.

## Intel Xeon Phi (MIC)

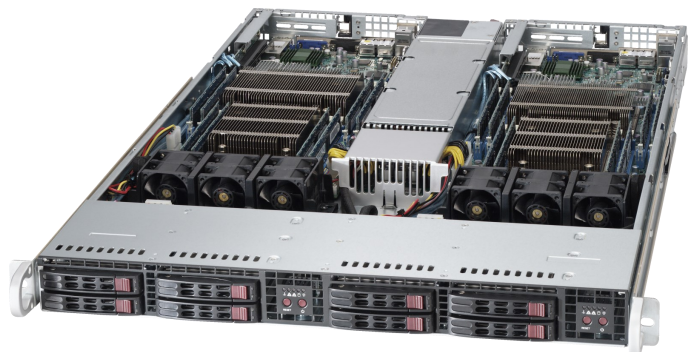




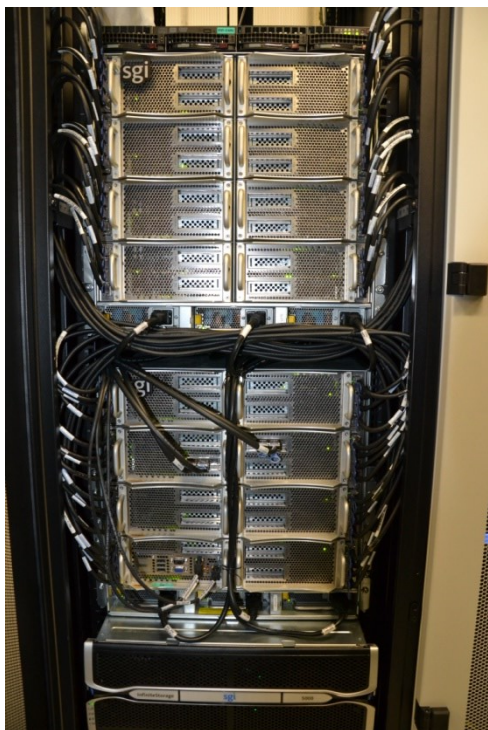
# Typické schéma počítače



# 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 (SE)

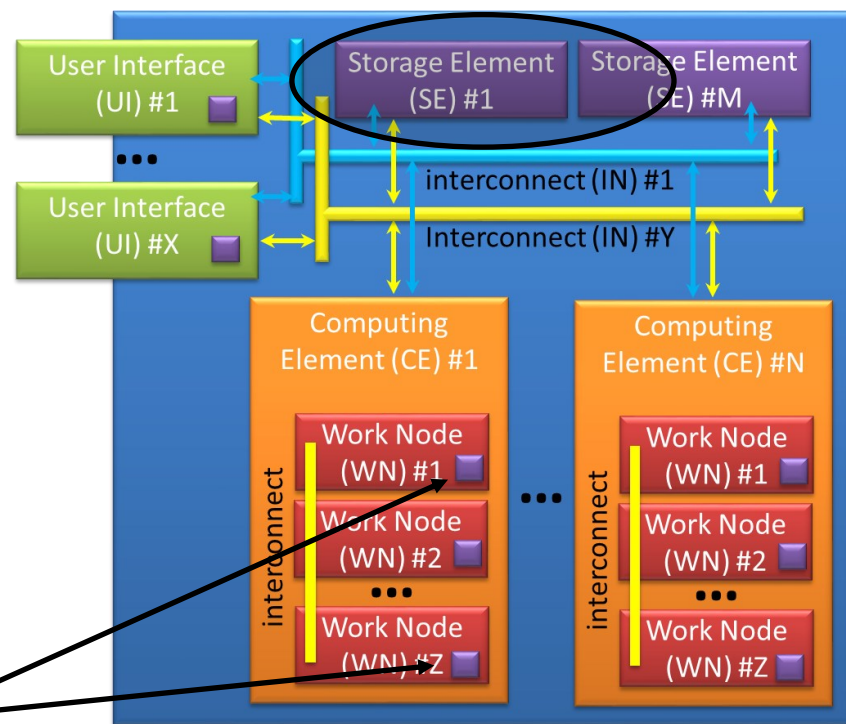
# Datové úložiště (SE) - dělení

Typy datových úložišť (SE - storage element) 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

lokální datové  
úložiště



# 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 uchování 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 určeny 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

# 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

velké množství HDD

RAID 0

**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 použijí jako náhrada za poškozené. Po dobu **rebuildování diskového pole** (nový výpočet parity dat) může být přístupová rychlost k datům snížena.

# Hierarchické datové úložiště

**Hierarchical storage management (HSM)** je technika ukládání dat, která automaticky přesouvá data mezi nákladnými a levnými úložnými médii. Udržování dat na vysokorychlostních zařízeních, je pro mnoho organizací neúnosně drahé, proto systémy HSM ukládají většinu podnikových dat na pomalejší zařízení a poté je v případě potřeby kopírují na rychlejší diskové jednotky. Systém HSM sleduje způsob, jakým se data používají, a odhaduje, která data lze bezpečně přesunout na pomalejší zařízení a která by měla zůstat na rychlých zařízeních.



**brno10-ceitec-hsm**  
(páskový robot)

**HSM úložiště jsou vhodná  
pro archivaci a zálohu dat.**



# Jednotky

Multiples of bytes				V · T · E	
Decimal		Binary			
Value	Metric	Value	IEC	JEDEC	
1000	kB kilobyte	1024	KiB kibibyte	KB kilobyte	
1000 <sup>2</sup>	MB megabyte	1024 <sup>2</sup>	MiB mebibyte	MB megabyte	
1000 <sup>3</sup>	GB gigabyte	1024 <sup>3</sup>	GiB gibibyte	GB gigabyte	
1000 <sup>4</sup>	TB terabyte	1024 <sup>4</sup>	TiB <b>tebibyte</b>	-	
1000 <sup>5</sup>	PB petabyte	1024 <sup>5</sup>	PiB pebibyte	-	
1000 <sup>6</sup>	EB exabyte	1024 <sup>6</sup>	EiB exbibyte	-	
1000 <sup>7</sup>	ZB zettabyte	1024 <sup>7</sup>	ZiB zebibyte	-	
1000 <sup>8</sup>	YB yottabyte	1024 <sup>8</sup>	YiB yobibyte	-	

Orders of magnitude of data

původní značení

# 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)?