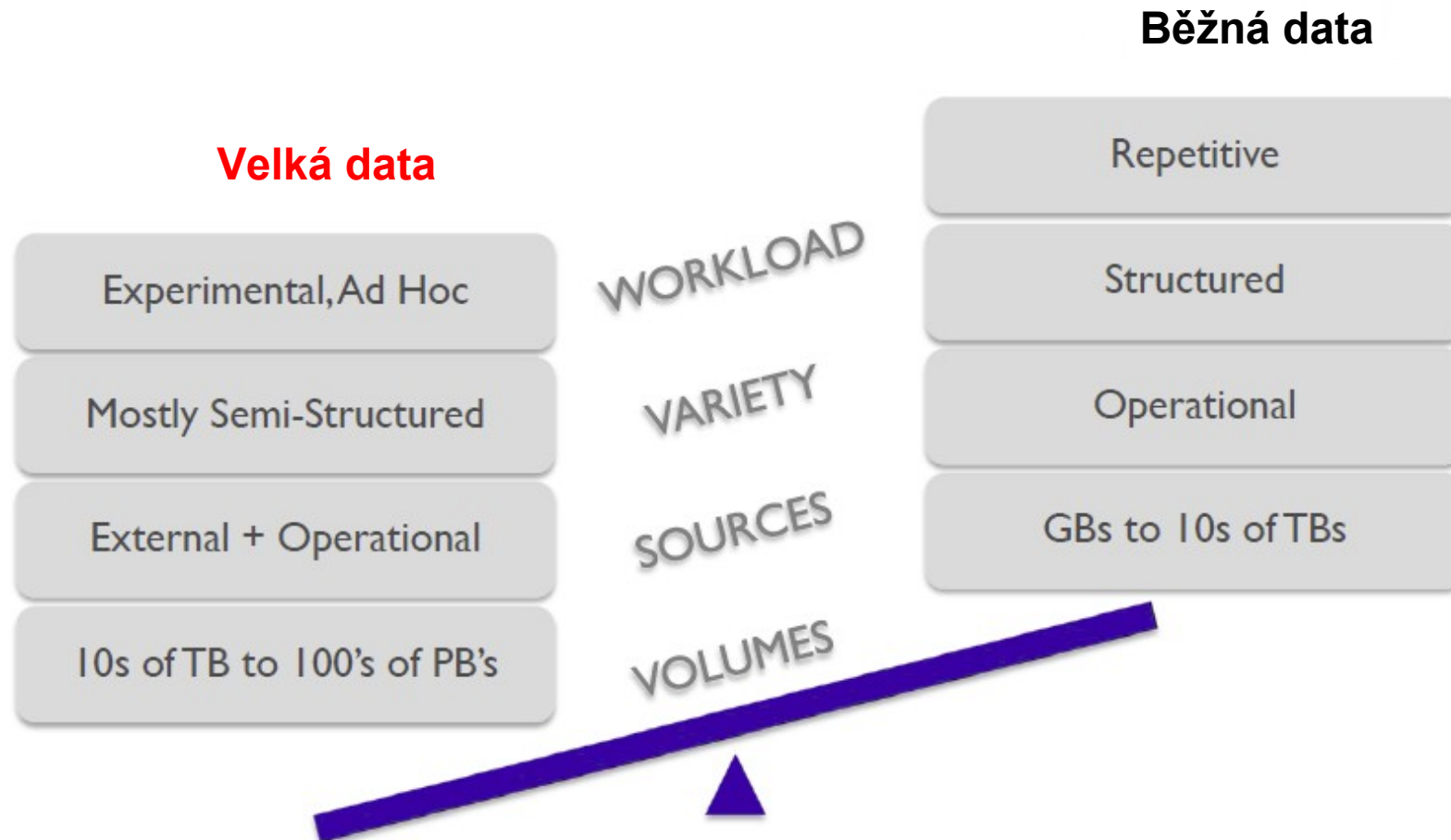


Virtuální paralelní infrastruktury pro velká data v metabolomice

**Fesl J., Doležalová M. , Cehák J.,
Moos M., Šimek P. a Janeček J.**

Velká data vs. běžná data



Velká data = obecně nezpracovatelná běžnými stolními počítači

Typické problémy při práci s velkými daty v metabolomice

- **Ukládání** a spolehlivé **skladování** dat
(nelze lokálně → datová úložiště, RAID svazky)
- Rychlé a spolehlivé **načítání**
(dnes zpravidla vysokorychlostní optické sítě)
- **Zpracování/Vytěžování** dat
(počítačové clustery, multiprocessorové karty CPU vs. GPU)
- **Statistická analýza** dat
(většinou pomocí konkrétního software -R, StatGraphics)

Infrastruktury pro práci s velkými daty

Rozpočet: stovky miliónů korun



**Velkokapacitní datová
úložiště**

**StovkyTB až
jednotky EB**

SEZNAM.CZ

Google

HPC výpočetní cluster



10-40 Gbit/s



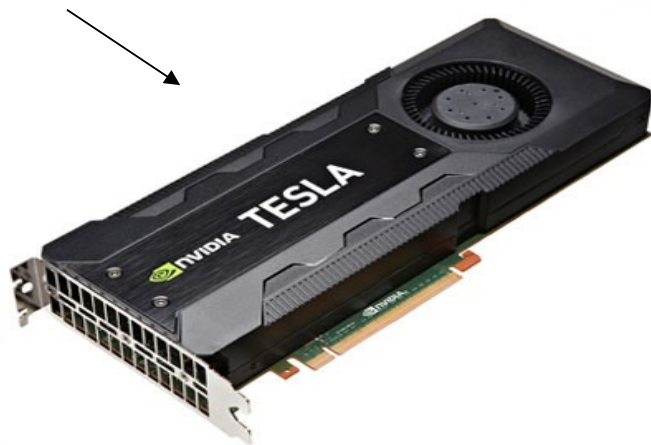
Rychlá počítačová síť

**Tisíce
CPU jader
TB paměti RAM**

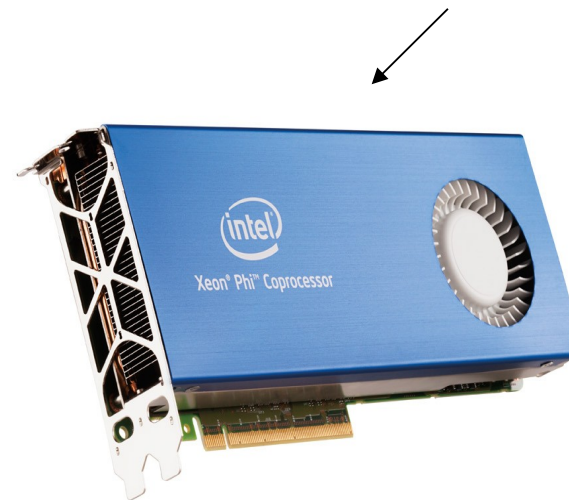


Doplňkové technologie pro zpracování dat

Vícejádrové GPU karty



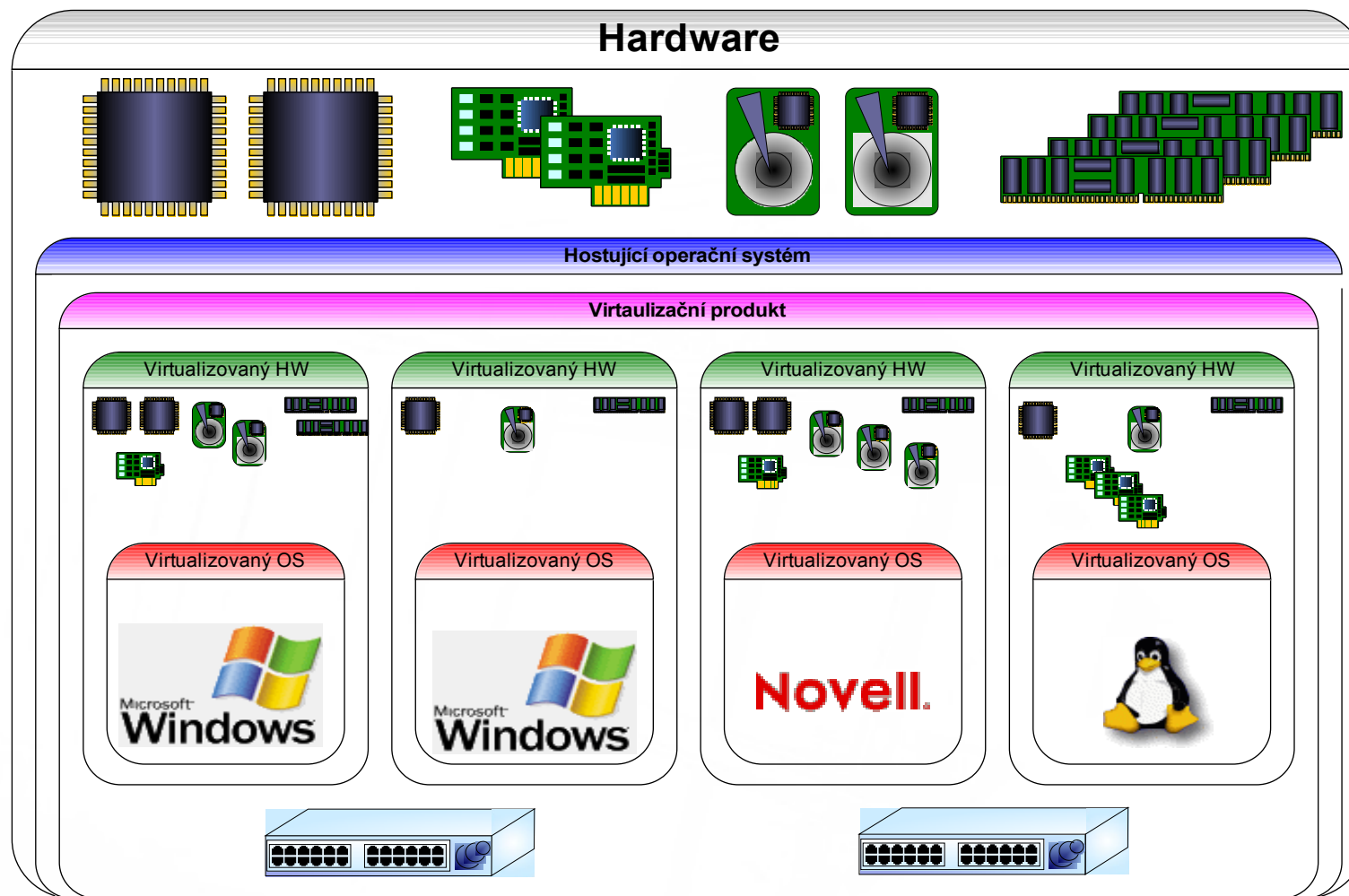
Vícejádrové CPU karty



+ Dokáží nahradit jednotky až desíky běžných počítačů

- Vyžadují speciální podporu konkrétními programy, specifické způsoby programování, ne každý software lze upravit pro jejich využití

Virtualizace – motivace a schéma

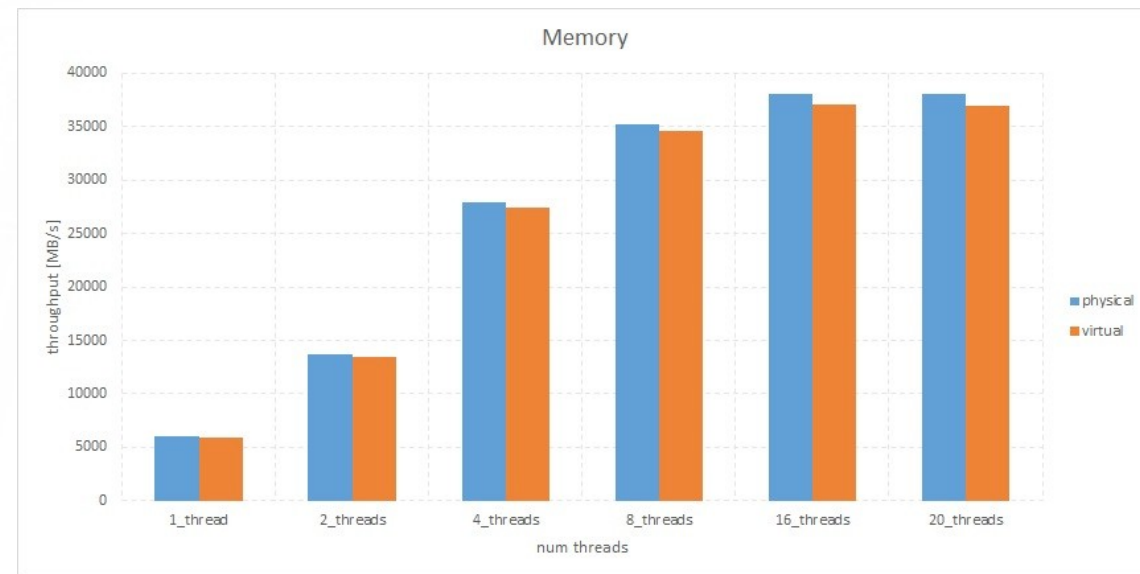
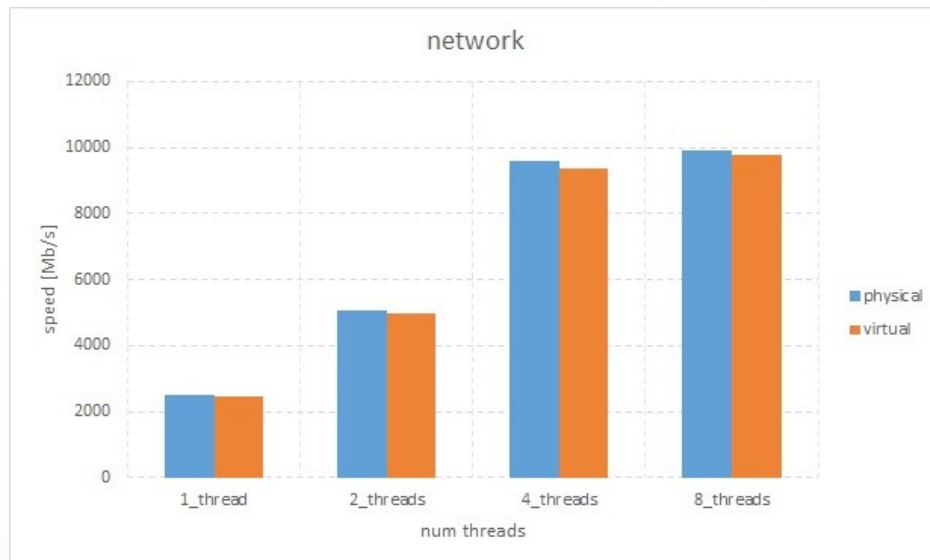
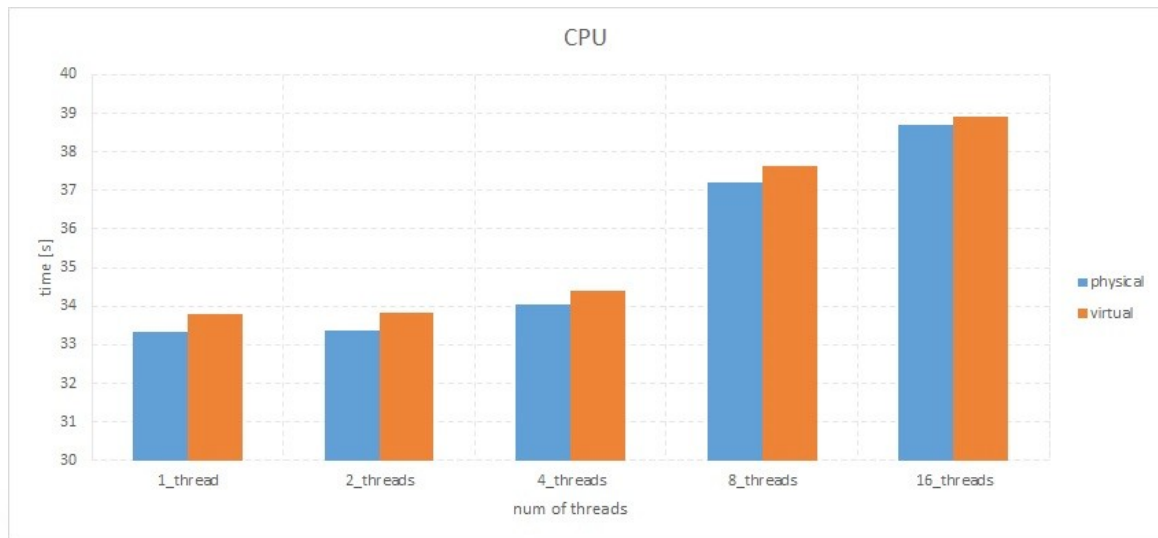


Na jediném fyzickém počítači může současně běžet **několik počítačů** virtuálních a to i ve **stejnou dobu**.

Co všechno lze/nelze virtualizovat ?

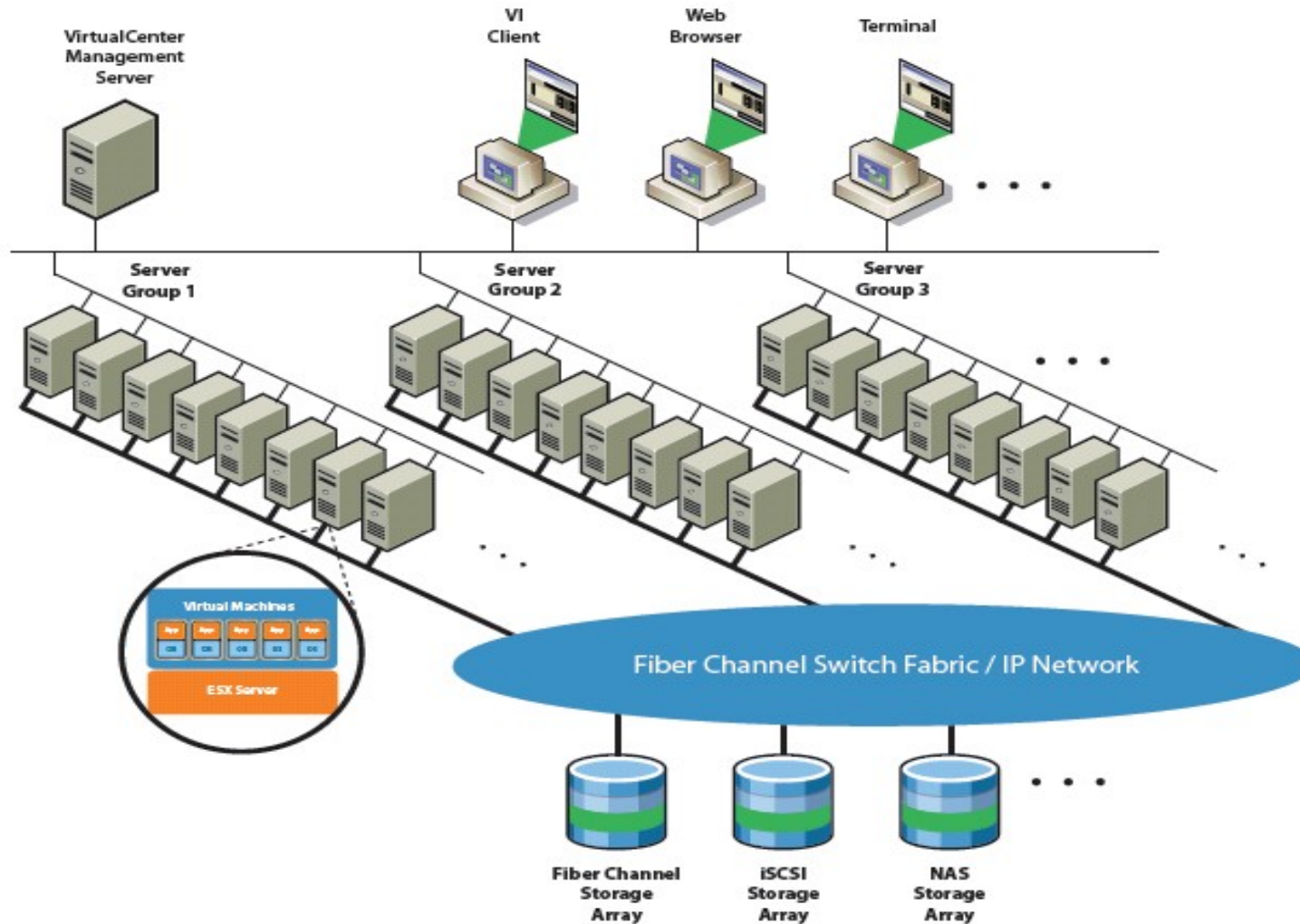
- Teoreticky **lze** libovolný **operační systém** (Linux, Windows, Android), což umožňuje **spuštění libovolného programu**
- Virtuální počítače **lze** propojovat **virtuálními počítačovými sítěmi** (VPS), které mohou mít i složitější topologii
- Spojením virtuálních počítačů přes VPS, lze vytvářet plnohodnotné **virtualizační clustery**
- Virtuální počítače **nemohou** (v současné chvíli) využívat prostředky pro masivní paralelizaci pomocí **GPU karet**, nicméně **multiprocessorové karty Xeon Phi** využít **lze**

Vyplatí se virtualizovat ?



Pro hardwarově asistovanou virtualizace je její režie pouze 1.5-4% v závislosti na dané operaci.

Distribuovaná virtualizační infrastruktura (DVI) obecné schéma



DVI pro zpracování velkých dat v univerzitních podmínkách

4 výpočetní uzly
1 Řídicí uzel

Dokáže nahradit 30 běžných počítačů.



10 Gbit/s optický switch



20 TB datové úložiště
RAID 5 HW controller



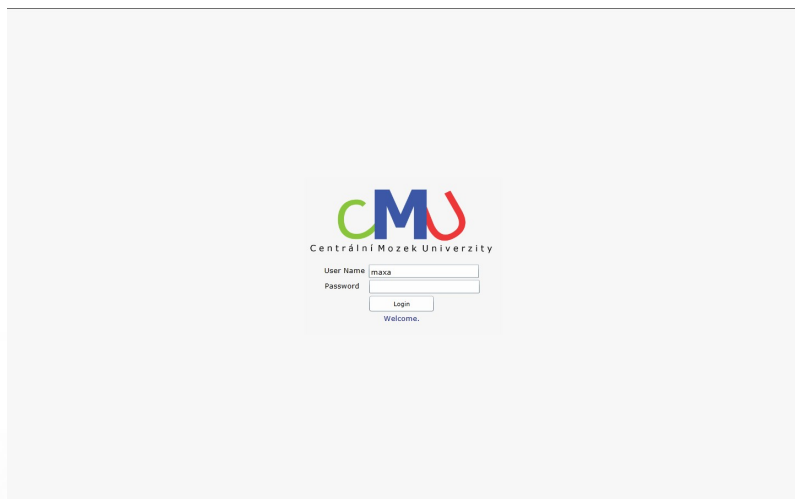
156 CPU jader + 360 XEON Phi jader
416 GBs RAM

Rozpočet: 620 tisíc korun



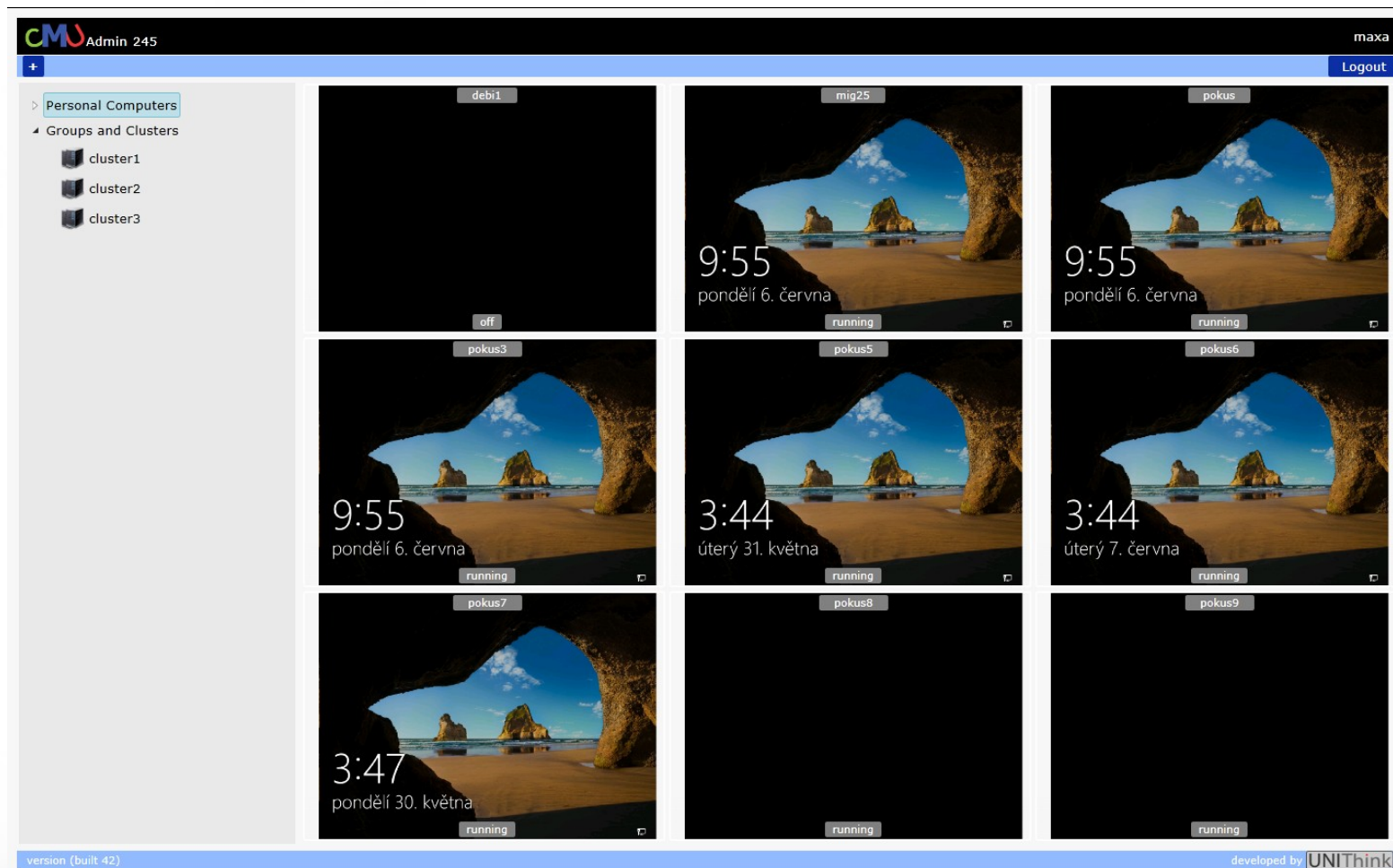
Centrální Mozek Univerzity

System CMU – ukázka ovlání

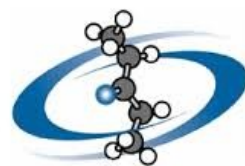
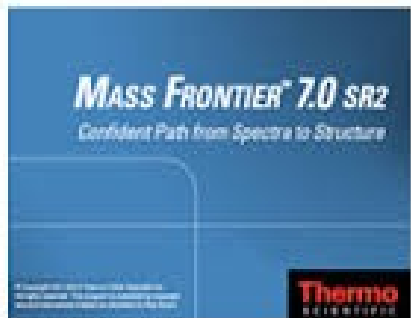


Přihlášení do systému

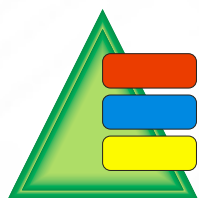
Uživatelská aplikace



BioInformatický software provozovaný na CMU



ACD/ChemSketch



Metabolite Mapper



Prostor prostor pro Vaše dotazy



Přírodovědecká
fakulta
Faculty
of Science



BIOLOGICKÉ
CENTRUM
AV ČR, v. v. i.



BIOLOGY
CENTRE
CAS

LABORATORY OF ANALYTICAL BIOCHEMISTRY & METABOLOMICS



Centrální Mozek Univerzity