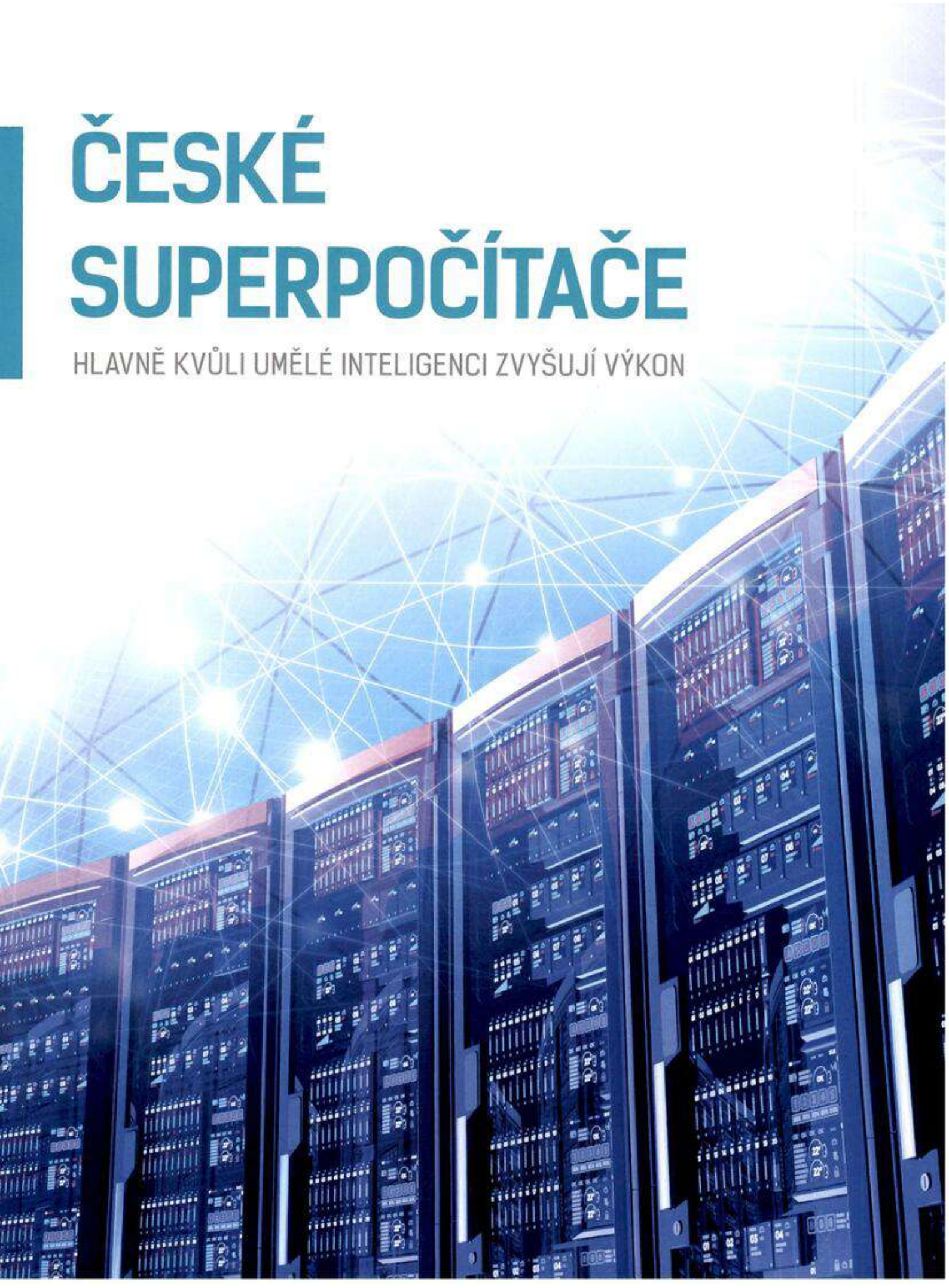
ČESKÉ SUPERPOČÍTAČE

HLAVNĚ KVŮLI UMĚLÉ INTELIGENCI ZVYŠUJÍ VÝKON



Superpočítače jsou i v Česku důležitým středobodem pro náročné výpočty, které se týkají různých segmentů života kolem nás. Pracuje se na nich nejen v akademické, ale i v soukromé sféře, což je podníceno masivním rozmachem velkých jazykových i obrazových modelů, které vyžadují extrémně vysoký výkon při trénování umělé inteligence (AI).

POKROK NEZASTAVÍŠ

Ať už hrajete hry na počítači nebo na moderní herní konzoli, možná si neuvědomujete, že to, co máte za „pár tisíc" připojené v zásuvce, bylo ještě kolem roku 2005 něco, co stálo stovky milionů dolarů, mělo spotřebu 1,5 MW a pro všechny skříně, ve kterých byl takový stroj rozložen, bylo potřeba přes 232 m^2. Rychlost technologického vývoje jde rychle kupředu a stále se řídí Moorovým zákonem. Ten říká, že počet tranzistorů, které mohou být umístěny na integrovaný obvod, se při zachování stejné ceny zhruba každých 18 měsíců zdvojnásobí. Definice superpočítače proto není určená samotným výkonem, ale jeho konstrukcí, která zahrnuje velké množství moderních komponent dané doby do společného výpočtu.

Vzhledem k tomu, jak rychle takový superpočítač zastarává, je třeba jeho výkon v dané době co nejvíce využít – ať pro akademické a vědecké výpočty, tak i jeho pronajímáním soukromým subjektům. Dnes už si sice můžete v cloudu pronajmout výkon, jaký chcete, třeba i jen na pár minut, avšak pokud to myslíte s náročnými výpočty vážně a potřebujete jich opravdu hodně, je lepší koupit vlastní superpočítač. A v Česku jich máme několik.

NEJDŘÍVE VELKÉ FIRMY

Pokud se podíváme do velmi vzdálené historie, „superpočítače" tehdy v Česku znamenaly sálové počítače, které byly určené pro potřeby velkých podniků. V té době byly i pro základní „excelové" výpočty nutné obří stroje typu amerického IBM 360/30, který měla například Škoda Auto (tehdy AZNP) v Mladé Boleslavi od roku 1968. Podobný model se podílel na prvním přistání člověka na Měsíci, kdy bylo nutné spočítat na tehdejší dobu opravdu hodně dat. Ve Škodovce si počítač IBM odpracoval bez větší chyby 20 let administrativních prací a pak byl sešrotován.

Škoda Auto od té doby několikrát modernizovala. V nejnovějším případě má nový superpočítač HPE SGI 8600 s původním výkonem 2 PFLOPS (petaflops, tedy 10^{15} operací v pohyblivé řádové čárce za sekundu), ale upgradovatelným na 15 PFLOPS. A i když je spotřeba jen kolem 400 kW, potřebuje plochu hokejového hřiště na všechny skříně a dodatečné komponenty.

Kolik superpočítačů bude potřebovat Grok Elona Muska?

SUPERPOCITAC RYCHLE ZASTARÁVÁ A JE POTŘEBA JEHO VÝKON V DANÉ DOBĚ CO NEJVÍCE VYUŽÍT – AŤ PRO AKADEMICKÉ A VĚDECKÉ VÝPOČTY, TAK I JEHO PRONAJÍMÁNÍM SOUKROMÝM SUBJEKTŮM.

A na co Škoda Auto potřebuje superpočítač? Vysoký výpočetní výkon je v tomto případě používán hlavně na simulace všeho druhu. Od samotné fáze výzkumu prvních prototypů motorů i konstrukcí, až po testování finálních návrhů. Týká se to fyzikálních vlastností motoru při provozu, vlastností konstrukčních prvků v simulovaných scénářích reálného světa a stejně tak simulace crash testů nebo optimalizace výroby jednotlivých dílů pro stavbu automobilu na výrobní lince.

Vždy je totiž levnější co nejvíce problémů otestovat, a tím pádem odchytit v co nejranějších fázích vývoje – pokud by se pouze testovalo na vyrobené „hmotě", bylo by to zdlouhavé a velmi nákladné. Díky simulacím mohou inženýři udělat velké množství testů, a tím pádem i velké množství prototypů a vylepšení za krátký čas. Protože vždy je co zlepšovat – ať už z pohledu kvality, tak i ceny, respektive výrobních nákladů.

Velké množství superpočítačů se objevilo v Českém hydrometeorologickém ústavu (CHMI), ve kterém je používají pro analýzu a zpracování vlastních modelů počasí pro krátkodobou a dlouhodobou předpověď – ano, to je známý model Aladin-Climate/CZ. Už od roku 2021 disponuje CHMI superpočítačem NEC SX-Aurora TSUBASA s výkonem až 940 TFLOPS, 24 TB operační paměti DDR4 a 18 TB rychlé paměti HBM2. Síťové úložiště má 2 PB.

OSTRAVSKÉ EPICENTRUM

V případě akademického výzkumu, který je ale úzce navázán i na soukromou sféru, je v oblasti superpočítačů v Česku aktuálně největší středobod v Ostravě, konkrétně v rámci národního výzkumného inovačního centra IT4Innovations, které je pod záštitou VŠB – Technické univerzity Ostrava. Už v roce 2013 se zde objevil první superpočítač s označením Anselm, který fungoval až do roku 2021. Poskytoval výpočetní výkon asi 94 TFLOPS, což sice není mnoho, ale někde se začít muselo.

V roce 2015 došlo na instalaci dalšího superpočítače – Salomon s původní cenovkou 270 milionů korun, který už přinesl na tehdejší dobu pořádný skok ve výkonu. Díky teoretickému výkonu až 2 PFLOPS se v době spuštění umístil dokonce v žebříčku nejvýkonnějších superpočítačů na světě, a to na 40. pozici. Technologie ale postupuje neúprosně, takže v červnu letošního dubna došlo k jeho

definitivnímu vypnutí, protože poměr mezi výkonem a spotřebou (náklady) byl už neefektivní, a navíc je problém s náhradními díly. Za devět let provozu se postaral o spuštění 8,7 milionů samostatných úloh z 1 085 různých výzkumných projektů jak z Česka, tak i ze zahraničí díky akademické spolupráci. Vědci, studenti i firmy využívali výkon pro oblast inženýrství, ale také biologie nebo fyzikální a materiální vědy.

Ostravská univerzita má však už nové modely superpočítačů ve svém portfoliu. V roce 2019 spustila Barboru s výkonem 849 TFLOPS a od roku 2021 ještě mnohem výkonnější Karolinu od HPE, která už má kombinaci jak výkonných serverových procesorů AMD Epyc 7763 s 64 jádry, tak hlavně výpočetních grafických karet Nvidia A100, takže celkový výkon je maximálně 15,7 PFLOPS. Díky tomu je v celosvětovém žebříčku na 135., respektive 54. místě, pokud jde jen o modely s výpočetními grafickými kartami. Z pohledu efektivity (spotřeba vs. výkon) je v celosvětovém žebříčku na 36. místě, byť v době spuštění byla na 8. místě.

Akcelerovanou část ostravského superpočítače Karolina tvoří 72 serverů s celkovým teoretickým výpočetním výkonem až 360 PFlop/s pro výpočty umělé inteligence

A TY DALŠÍ

K dispozici jsou i specializované superpočítače, které jsou vhodné jen pro určitý typ úloh. V datacentru nechybí Nvidia DGX-2, platformy od Fujitsu (A64FX s HBM2), akcelerační platformy s programovatelnými kartami Intel Stratix 10 (FPGA) nebo řešení od AMD v podobě Xilinx Alveo se staršími výpočetními kartami MI100. V poslední době se trh také přiklání k řešení na bázi procesorů s architekturou Arm. Inovační centrum má k dispozici platformu Arm Ampere Altra s Nvidia A30 a DPU čipy Bluefield-2 série E.

Protože i IBM si vyrábí vlastní procesory, nesmí chybět ani zde, konkrétně IBM Power10. Mezi další speciality pak patří Nvidia Grace CPU Superchip, Intel Sapphire Rapids s pamětí HBM, serverové procesory AMD Epyc s velkou cache paměti (Milan-X), systémy s kartami Nvidia A40 a ještě další varianty.

To vše mají studenti a vědci k dispozici k tomu, aby si nová řešení nejen vyzkoušeli a naučili se s nimi pracovat, ale také je využili pro konkrétní věci, kde bude daný hardware nejlépe využitý. Z těchto systémů se totiž staví velké superpočítače, které pak musí někdo obsluhovat a spravovat.

V automobilce Škoda mají zkušenosti se superpočítači již desítky let

EVROPSKÁ SPOLUPRÁCE

Čeští vědci mají přístup nejen na české superpočítače, ale také například i na evropský LUMI (HPE Cray EX235A) ve finském Kajaani, který nabízí výkon 580 PFLOPS a v celosvětovém žebříčku je na 5. místě. Základem jsou téměř 3 miliony jader procesory AMD Epyc a výpočetní grafické karty AMD Instinct MI250X, každá s 128 GB pamětmi HBM2e. Celý superpočítač je složen z 2 978, přičemž každý z nich obsahuje jeden zmíněný 64jádrový procesor a čtyři výpočetní karty. Vše je spojené rychlou síťovou infrastrukturou s propustností 800 Gb/s. Jen pro vizualizaci se používá 64 karet Nvidia A40.

Celý systém zabírá plochu dvou tenisových kurtů a hmotnost veškerého železa je 150 tun. Tento superpočítač je součástí EuroHPC a Česká republika je v rámci IT4Innovations partnerem, takže má možnost ho ve sdíleném módu využívat pro vlastní potřeby.

NEJLEPŠÍ UMĚLÁ INTELIGENCE

O nový vítr do plachet technologickému pokroku v oblasti superpočítačů se zasloužil rozmach nového druhu umělé inteligence – velkých jazykových modelů (LLM), ale stejně tak s tím spojených systémů pro umělé generativní tvorby obrázků, fotografií a v nejnovějším případě i videí, hudby či hlasu.

K dosažení revolučního stupně inteligence, kterou přinesl model GPT od OpenAI, bylo zapotřebí nejen nových algoritmů, ale především

Čip Cell z PlayStation 3 si našel svého času cestu až do superpočítačů Roadrunner

extrémně velké množství dat a extrémně vysoký výpočetní výkon. Právě trénink nového modelu umělé inteligence a jeho náročnost je důvod, proč se Nvidia tak raketově stala nejhodnotnější společností na světě. Její výpočetní grafické karty a rozšířená softwarová podpora pro masivní zpracování dat totiž umožňují provádět trénování i těch největších modelů, které jsou aktuálně na světě k dispozici.

Jen pro představu – nový jazykový model Llama 3.1 od Mety (dříve Facebook), který svými schopnostmi pomalu dosahuje na GPT-4, byl trénován na superpočítači s 16 tisíci kartami Nvidia H100, a to téměř tři měsíce v kuse s tím, že se plánuje nový model další čtvrtletí. Jedna tato grafická karta stojí přes půl milionu korun, přičemž xAI od Elona Muska, který vyvíjí vlastní umělou inteligenci s názvem Grok, bude mít na konci roku 2024 nový model natrénovaný na 100 tisících kartách.

Jedno trénování takto velkého modelu tak vyjde na stovky milionů dolarů (zajímavostí Llamy je, že je open source, a tedy zcela zdarma) a do budoucna se počítá s tím, že velikosti budou rychle narůstat, protože souboj velkých hráčů nezná konce – každý si chce získat silnou pozici na trhu. Lze tedy očekávat, že se velmi brzy dostaneme do fáze, kdy náklady vystoupí i na miliardy dolarů. A když si vezmeme analogii z filmového průmyslu, vlastně to není tak moc. Stovky milionů dolarů stojí jeden větší film.

SUPERPOČÍTAČ POD STOLEM

A v čem se vlastně liší vaše herní grafická karta od té výpočetní? Jedna klíčová věc je velikost grafické paměti, která je u herních modelů poměrně malá, a to i u GeForce RTX 4090 s 24 GB. Výpočetní karty v superpočítačích mají už standardně 80 GB a pro náročné úlohy je stejně nutné je spojovat, aby bylo dosaženo několik stovek gigabajtů extrémně rychlé grafické paměti, ve kterém se úloha nachází.

OSTRAVSKÁ UNIVERZITA V ROCE 2019 SPUSTILA BARBORU S VÝKONEM 849 TFLOPS A OD ROKU 2021 JEŠTĚ MNOHEM VÝKONNĚJŠÍ KAROLINU OD HPE.

Extrémně důležitá je přesnost výpočtů, respektive výkon při dané přesnosti. Nvidia, ale i AMD mají chytře oddělené, že grafické karty pro hraní her mají maximální výkon při nižší přesnosti, která stačí pro běžné věci a také pro hry. Pro vědecké účely je ale nutná mnohem vyšší přesnost (FP64), kde už jsou čipy herních grafik ořezané na naprosto tragická čísla.

Na druhou stranu, v případě zmíněné umělé inteligence už se ve velkém pracuje právě s mnohem nižší přesností, protože to pro řadu věcí (včetně umělých neuronových sítí) zcela dostačuje a benefit i stonásobného výkonu oproti vyšší přesnosti je prostě obrovský. Doma si tak sice vlastní velkou umělou inteligenci nenatrénujete, ale snadno si ji pustíte a na moderní grafické kartě vám poběží velmi rychle, byť pouze zmenšený model kvůli menší paměti (například Llama 3.1 8B).

GLOBÁLNÍ SUPERPOČÍTAČ

Už poměrně dlouho funguje i model distribuovaných výpočtů v rámci celosvětových superpočítačů zaměřených na různé oblasti vědeckého výzkumu. Mezi nejznámější projekty patří BOINC (Berkeley Open Infrastructure for Network Computing) a v době koronaviru byl extrémně populární Folding@Home nebo hledání mimozemských signálů SETI@home.

Tyto platformy fungují tak, že jednotlivým počítačům, které mají spuštěný daný program, rozešlou miniaturní oddělené úlohy, váš procesor nebo grafická karta úlohu spočítá a pošle výsledky zpátky. Řešení se tak postupně skládá díky obrovskému množství dobrovolníků jak z řad jednotlivců, tak celých superpočítačů.

Výhodou je, že si můžete vybrat, jaké výzkumy a projekty chcete výkonem podpořit, přičemž máte pochopitelně pod kontrolou, kolik výkonu a v jaký čas chcete poskytovat – například, když nic na počítači neděláte, náročná úloha se spustí, ale když pracujete nebo hrajete, nic se nepočítá. Projekty jsou z oblasti výzkumu matematiky, vesmíru, biologie, ale i fyziky pro Velký hadronový urychlovač v Cernu (LHC).

Takže i váš počítač může být superpočítačem už dnes.

Karel Javůrek