188GB HBM3 Nvidia H100 NVL vznikla na míru pro Chat-GPT

Zdroj: Nvidia

Nvidia se pochlubila novou verzí akcelerátoru Hopper, která vznikla na míru pro Chat-GPT. Zajímavá je teoreticky až 2× vyšší energetickou efektivitou a nezvyklou kapacitou paměti, 2× 94 GB HBM3…

Nvidia loni uvedla akcelerátory Hopper ve dvou základních podobách. Jednak jako SXM5 moduly a jednak jako PCIe karty. Jako obvykle dosahuje SXM5 varianta vyššího výkonu i TDP, neboť je na modulu podstatně více prostoru pro velký radiátor chladiče. Nová varianta označená jako NVL má však opět podobu PCIe karty, respektive karet. Nvidia H100 NVL jsou totiž dvě PCIe karty spojené můstkem NVLINK. Nejde však o původní H100 PCIe. Nvidia k tomuto řešení sáhla patrně pro vyšší výpočetní denzitu takového řešení. Zkrátka méně prostoru zabraného chladičem ~ prostor pro více GPU. Jak vyplývá z parametrů, energetická efektivita má být 2× vyšší (2× více výkonu při stejném TDP díky dvojici karet s TDP poloviční oproti SXM5 provedení). Lze však očekávat, že realita bude trochu níže, protože reálné a papírové takty se budou patrně nezanedbatelně lišit.

	Nvidia A100	Nvidia H100		
GPU	GA100	GH100		2× GH100 „NVL“
architektura	Ampere	Hopper		
formát	SXM4	SXM5	PCIe	2× PCIe
CU/SM	108	132	114	2× 132?
FP32 jader	6912	15872 16896	14592	2× 16896?
FP64 jader	3456	8448	7296	2× 8448?
INT32 jader	6912	8448	7296	2× 8448?
Tensor Cores	432	528	456	2× 528?
takt	1410 MHz	1980 MHz	1750 MHz	1980 MHz?
		↓↓↓ T(FL)OPS ↓↓↓		
FP16	78	120 134	102	2× 134
BF16	39	120 134	102	2× 134
FP32	19,5	60 67	51	2× 67
FP64	9,7	30 34	26	2× 34
INT4	?	?	?	?
INT8	?	?	?	?
INT16	?	?	?	?
INT32	19,5	30 34	26	2× 34
FP8 tensor	✗	1979/3958*	1513/3026*	2× 1979/3958*
FP16 tensor	312/624*	989/1979*	757/1513*	2× 989/1979*
BF16 tensor	312/624*	989/1979*	757/1513*	2× 989/1979*
FP32 tensor	19,5	60? 67?	51?	2× 67?
TF32 tensor	156/312*	495/989*	378/757*	2× 495/989*
FP64 tensor	19,5	67	51	2× 67
INT8 tensor	624/1248*	1979/3958*	1513/3026*	2× 1979/3958*
INT4 tensor	1248/2496*	?	?	?

	↑↑↑ T(FL)OPS ↑↑↑				
TMU	432		528	456	2× 528
LLC	40 MB		50 MB		2× 50 MB?
sběrnice	5120bit		5120bit		6144bit
paměť	40 GB	80 GB	80 GB		2× 94 GB
HBM	2,43 GHz	3,2 GHz	HBM3 5,23 GHz	HBM2E 3,2 GHz	HBM3 5,1 GHz
pam. prop.	1555 GB/s	2048 GB/s	3350 GB/s	2048 GB/s	2× 3,9 TB/s
TDP	400 W		700-800 W	350 W	700 W
transistorů	54,2 mld.		80 mld.		2× 80 mld.
plocha GPU	826 mm²		814 mm²		2× 814 mm²
proces	7 nm		4nm		
datum	5. 2020	11. 2020	2022?		H2 2023

Protože jsou jazykové modely Chat-GPT velké, osadila Nvidia poprvé GPU GH100 všemi šesti HBM moduly (tzn. plně využila 6144 bit sběrnici) a použila navíc HBM3 (oproti HBM2E na původní PCIe variantě). To by teoreticky odpovídalo 6× 2× 16 GB = 192 GB paměti, jenže reálně specifikace uvádí 188 GB. Jak toho lze dosáhnout, když žádné 15,66GB moduly HBM neexistují? Nvidia se nejspíš dohodla s některým výrobcem pamětí, který jí dodává moduly s deaktivovanými vadnými buňkami, což by mohlo znamenat mírně nižší kapacitu, ale i podstatně výhodnější cenu.

Přestože Nvidia hovoří o vydání H100 NVL, reálně mají být tyto akcelerátory dostupné někdy ve druhém pololetí letošního roku.

nahlásit chybu

Nvidia nechala redaktory ukázat GeForce RTX 4090 [fotografie]

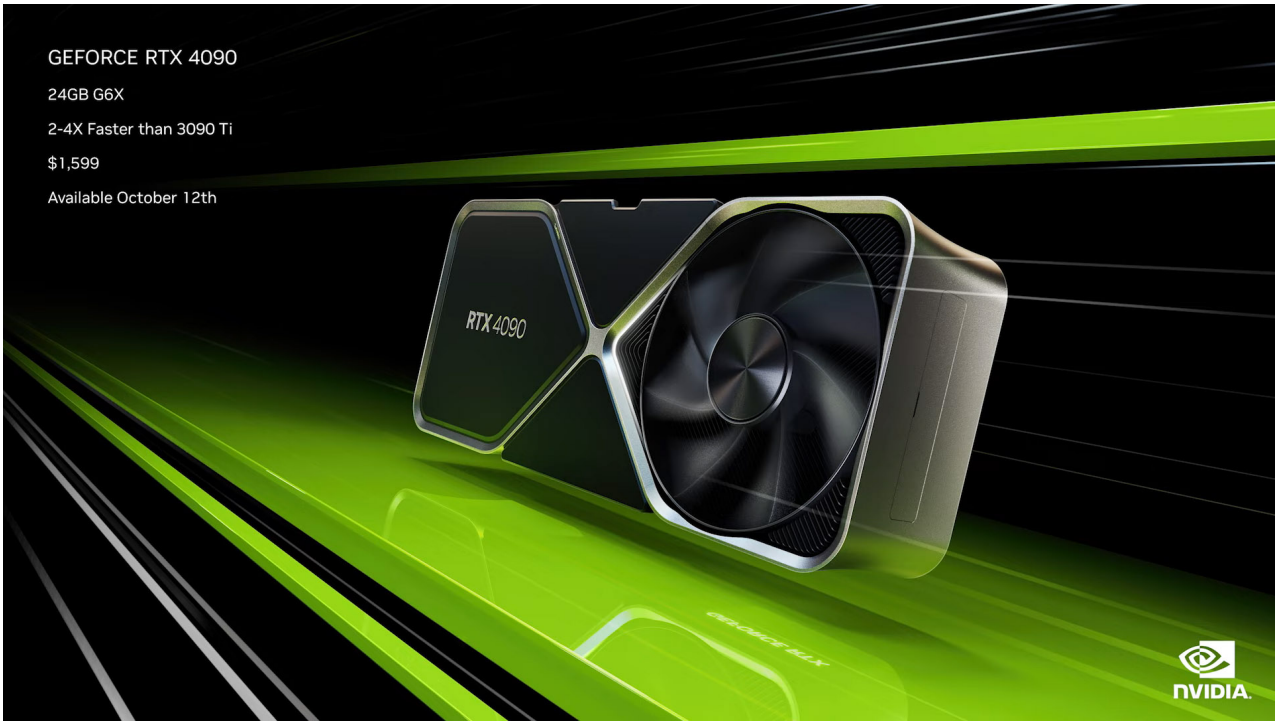
GeForce RTX 4090 sice bude vydána až příští týden (zároveň se zahájením prodejů Intel Arc A7x0), ale Nvidia nechala redakce zveřejnit fotografie karty již nyní, v podstatě zároveň s recenzemi Arc…

GeForce RTX 4080 v GeekBench: 10 % nad RTX 3090 Ti, 34 % nad RTX 3080

16GB (tedy ta nezrušená) verze GeForce RTX 4080 byla otestována v GeekBench 5. Její výkon dosahuje v průměru 10 % nad GeForce RTX 3090 Ti a 34 % nad GeForce RTX 3080...

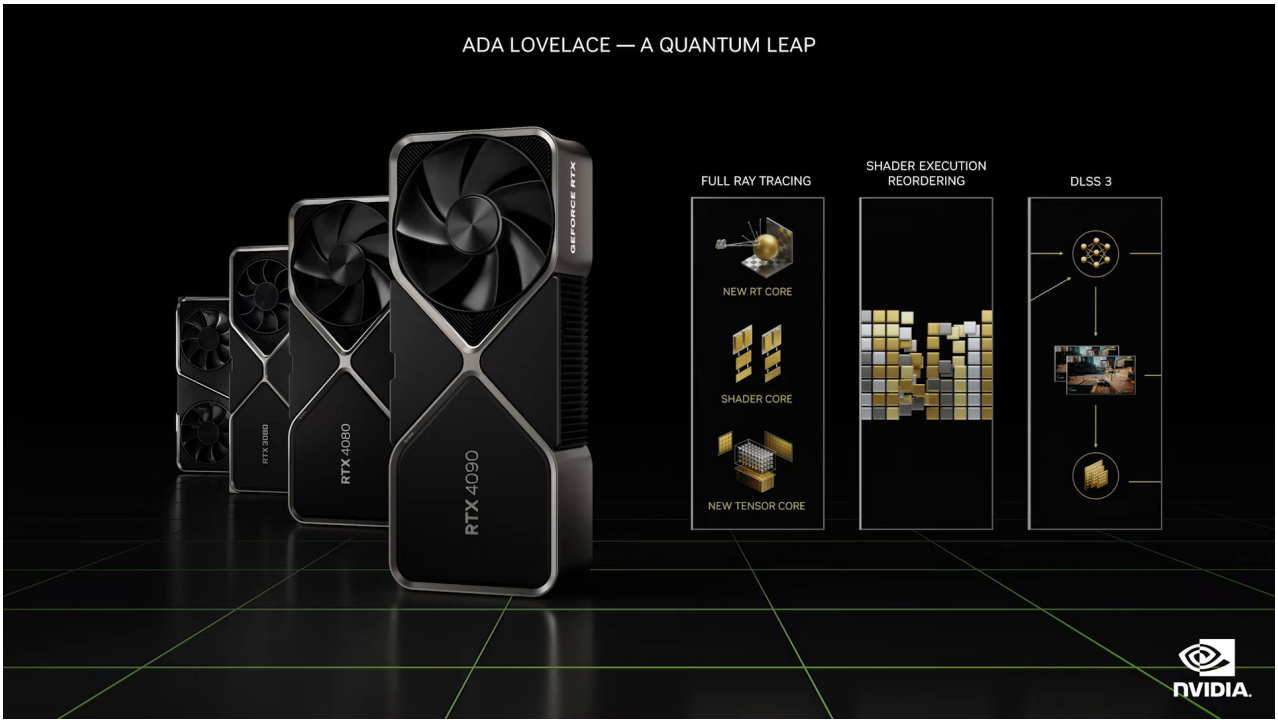
Uživatelé: GeForce RTX 4080 by měla stát nanejvýš $800

Z uživatelské ankety, které se zúčastnilo téměř 12 tisíc respondentů, vyplynulo, že 97,8 % uživatelů považuje GeForce RTX 4080 za předraženou, pouze 0,9 % se s její cenou ztotožňuje…



GeForce RTX 4090 v GeekBench: 60-63 % nad GeForce RTX 3090 Ti

Skóre GeForce RTX 4090, nové generace grafického high-endu od Nvidie, se objevilo v online databázi benchmarku GeekBench…



GeForce RTX 4080 16GB se původně měla jmenovat GeForce RTX 4070

Nakonec vše nasvědčuje tomu, že původní plán Nvidie byl podobný jako u generace Ampere. Na největším jádru řady 102 postavit GeForce RTX x090 a o stupeň menší řadu 103 použít pro GeForce RTX x070...

Jiří "no-X" Souček

více článků, blogů a informací o autorovi

Diskuse ke článku 188GB HBM3 Nvidia H100 NVL vznikla na míru pro Chat-GPT

před 1 dnem | Wladows | Poběží na tom i Skynet?
před 2 dny | no-X | To jsou pěkné teorie, které by dávaly smysl, ale...
před 2 dny | melkor | A pak, že Sci-Fi je brak. 87 let od vydání a...
před 2 dny | TyNyT | a tohle sedí taky: Přes tento střízlivý úsudek...
před 2 dny | TyNyT | Na konci protokolu shrnula odborná komise...

před 2 dny | jk2 | Dle mých zkušeností je nutné výsledky z Chat-GPT...
před 2 dny | danieel | Ta pamet bude mit spise chybejici kanal. Kazdy...
před 2 dny | RedMaX | Ptali se Chat-GPT a říkal, že je současný český...
Zobrazit diskusi
Pro psaní komentářů se, prosím, přihlaste nebo registrujte.