AI med minne som aldrig bleknar

EXPERTARTIKEL

Så kan NVDIMM dramatiskt höja prestanda på AI-beräkningar i datacenter

Av Arthur Sainio, Smart Modular Technologies

Arthur Sainio är produktmarknadschef på iSMART Modular Technologies där han arbetar med att föra fram nya arkitekturer som persistent MRAM och NVDIMM för användning inom industriell IoT (IIoT), telekom, flyg och rymd och försvar. Det gör han även i egenskap av vice ordförande i NVDIMM Special Interest Group på SNIA/SSSI. Han har tidigare jobbat på Hitachi Semiconductor och har en MBA från San Francisco State University och en MS från Arizona State University.

Det finns två stora flaskhalsar i dagens datacenter: kommunikationen med sekundärminnet och storleken på primärminnet.

Notera att dessa två problem uppträder i vad som normalt uppfattas som två väsensskilda delar av datorn. Klassiskt används primärminne (arbetsminne eller RAM) för att lagra kod och data tillfälligt under programkörning medan hårddiskar och annat sekundärminne eller datalager används för långsiktig lagring.

När ett datorprogram behöver hämta data från sekundärminnet – vilket kan bli ganska ofta när datavolymerna är större än primärminnet – tar det tid, vilket straffar tillämpningens prestanda.

PROBLEMET KAN ADRESSERAS via persistent minne (beständigt, icke-flyktigt), en teknik som inneburit en vändpunkt i lagringshierarkierna i traditionella datacenter. Det öppnar möjligheter för en ny unifierad hyperkonvergerad arkitektur som dramatiskt kan öka prestanda på sekundärminne.

Den pågående dataexplosionen har ackompanjerats av omfattande ökning av användandet av artificiell intelligens (AI) och maskininlärning (ML). Problemet är att traditionella system inte är konstruerade för att hantera de gigantiska datavolymer som används inom detta område.

Innan AI och ML kan bli vardagsmat behöver vi hitta ett sätt att reducera den tid som krävs för den dataintensiva så kallade ETL (extract–transform–load, omvandling av rådata till lagrad data) som används i AI- och ML-algoritmer för att transformera rådata till observationer och användbara insikter.

DESSUTOM BEHÖVER TIDSÅTGÅNGEN reduceras för det som kallas checkpointing – att kopiera data från primär- till sekundminne för att gardera mot dataförluster.

ETL för AI och ML kräver GPU-accelererade beräkningar och mycket IO. Beräknings- och IO-prestanda bestäms i hög utsträckning av bandbredd och latens (tidsåtgång för datatransport). För att implementera den beräkningstunga dataanalys som AI och ML använder, krävs system med hög bandbredd och låg latens.

Utgifterna för AI- och ML-system kommer att uppgå till 110 miljarder dollar år 2024. Det är mer än dubbelt så mycket som de 50 miljarder dollar som spås spenderas under 2020. Allt enligt prognoser från IDC (International Data Corporation).

Det krävs en exponentiell ökning i beräkningskraft för att hålla jämna steg med den utvecklingen. Å ena sidan utvecklas nya parallella arkitekturer för att svara på det trycket. Å andra sidan saknas en vital komponent i konventionella minneslösningar: NVM (nonvolatile memory, icke-flyktigt minne). Även om arkitekturerna förbättras, kommer strömavbrott att kunna kosta datacenter miljontals dollar. Det finns därför ett omedelbart behov av icke-flyktigt minne.

CHECKPOINTING INNEBÄR i detta sammanhang att ML-nät som är under träning då och då lagras i ett permanentminne för att säkra att resultat av mellanliggande beräkningar så långt inlärningen fortskridit, inte går förlorade. Detta är en extra stor utmaning för AI och ML eftersom checkpointing kostar beräkningskraft och energi utan att direkt föra lösningen i sig närmare målet.

▶▶

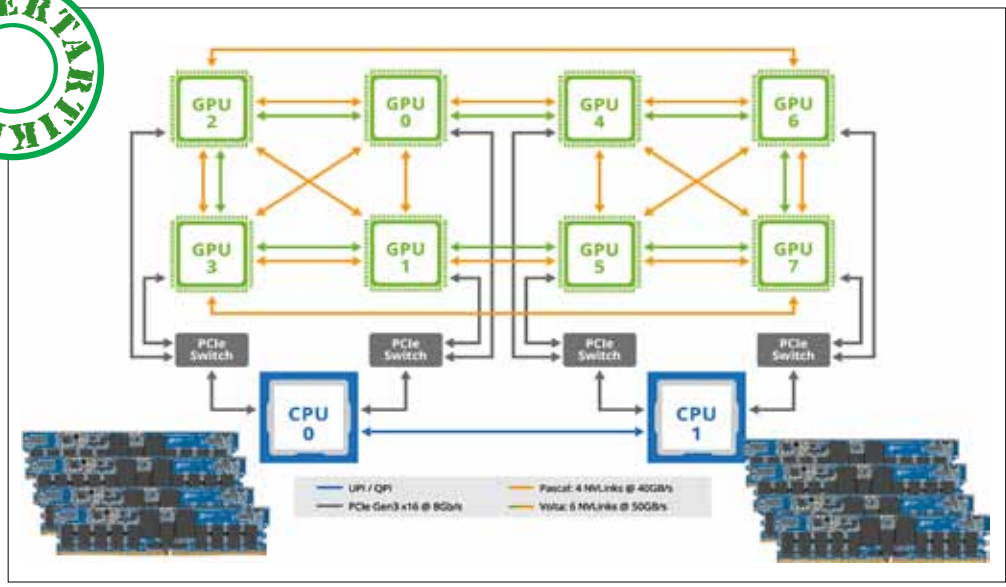
Penguin Relion 4118GTS – AI / ML Performance at Scale	
Processor/ Chipset	2x Intel® Xeon® Scalable Processor family, TDP up to 205W *Support for Cascade Lake
Data and Storage Layer	24x DIMM slots, 6 DPC, DDR4 - DRAM - NVDIMM - SCM 10 x 2.5" hot-swappable HDD/SSD bays - 4 x U.2 (Secure) NVMe devices only - 6 x 2.5" (Secure) SATA/SAS devices
PCIe Accelerated GPU and Networking	8X NVIDIA V100 SXM2 w/ NVLINK 4x 100G Low Latency High Speed Network 2x 25GbE Ethernet
Workloads and Verticals	

En server optimerad för AI och ML med hjälp av persistenta minnen.

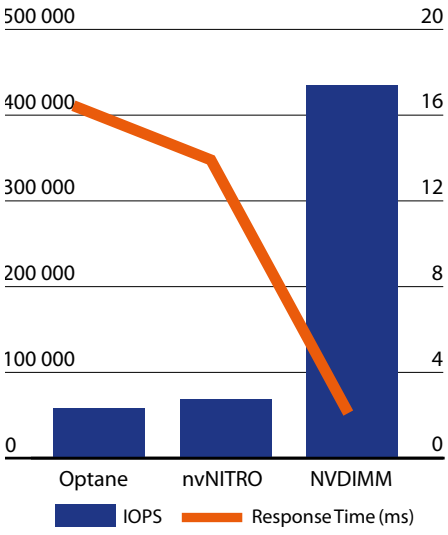
Bearbetningen i andra noder pausas medan informationen skrivs in i centralminnet. Operationen är skrivintensiv vilket förvärrar problemet i vissa lägen, eftersom hårddiskar och andra konventionella datalager arbetar ineffektivt när man skriver till dem.

Som svar på att checkpointing mot centrala minnen tar allt för lång tid i AI- och ML-tillämpningar, har man börjat placera icke-flyktigt minne nära cpu:n. Detta görs för att minimera påverkan från checkpointing, som trots allt är nödvändig. Lösningen ger en bättre balans mellan data och beräkningar, vilket gör att systemet kan möta produktionskraven.

Persistenta minnen av typen NVDIMM (Non-Volatile Dual In-line Memory Module) kan användas för att öka prestanda på tillämpningar där skrivningen är känslig för latens. Lagringsmodellen är persistent och prestanda blir i nivå med DRAM. Att NV-DIMM sänker latens och höjer prestanda skapar en unik möjlighet för datacenter att möta prestandakraven inom AI och ML utan att i större omfattning behöva bryta med existerande tekniklösningar.



Varje cpu får fyra NVDIMM-moduler på 32 GB som utgör snabba byte-adresserbara persistenta minnesceller.



En prestandajämförelse enligt ezFIO mellan en 2,5" Intel Optane NVMe-SSD, en MRAM NVMe U.2-SSD och en NVDIMM.

NÄR NVDIMM ANSLUTS till en server kan Bios:et se till att de uppfattas som en del av sekundärminnet. Därmed kan tillämpningen använda NVDIMM för att snabba upp checkpointing. Det traditionella alternativet vore att överföra checkpoint-data via IO-stacken över NVMe (NVM Express) och sedan spara till en SSD. Lösningen belastas med den latens som adderas i IO-stacken och NAND-minnet.

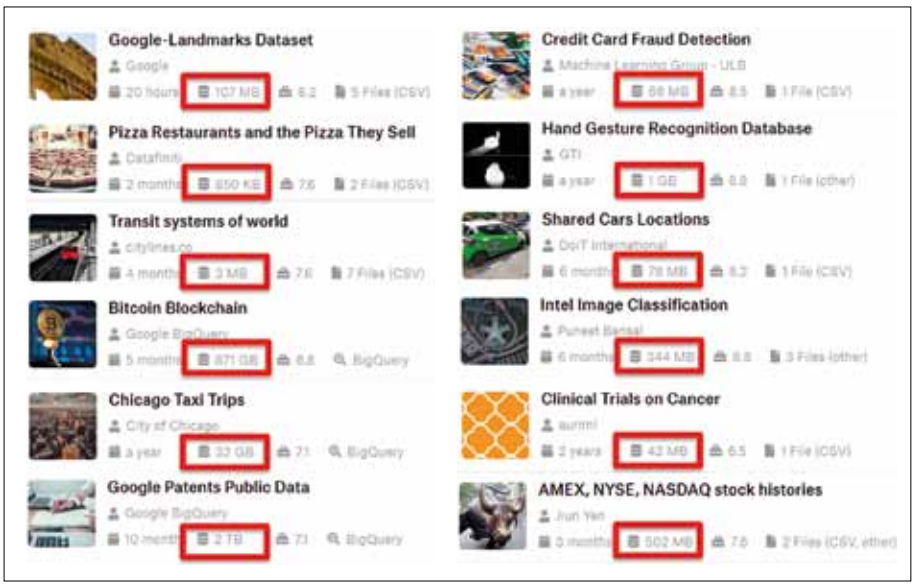
NVDIMM är en ideal lösning för högprestandaservrar för AI och ML. Genom att utnyttja den persistenta minnesregionen i primärminnet kan dataintensiv ETL och checkpointing arbeta med bandbredd och latens på DRAM-nivå – 25,6 GB/s respektive under 100 ns.

NVDIMM används för att snabba upp checkpointing men kan också användas inom ML för att öka prestanda och skydda de data som samlats in av algoritmer.

GPU-konfigurerade lagringsservrar kör algoritmer som är delsteg i simuleringar och ML. NVDIMM kan här användas för att skydda GPU-servrarna från att förlora simuleringsdata. Datavolymerna för typiska algoritmer mäts i allt mellan kilobyte och terabyte. Om data förloras måste hela processen börja om från början. Med fyra NVDIMM-konfigurerade servrar kan datavolymer upp till en terabyte utnyttja persistent minne snarare än traditionell lagring och därmed dramatiskt förbättra prestanda utan risk för förlorad data.

Data för simuleringar och AI/ML har snarlik karaktär och slussas vanligen in i servern via Infiniband eller Ethernet varefter det mellanlagras i SSD för att eliminera risken för dataförluster. Därefter flyttar GPU:n delar av data till DRAM för beräkningar.

LÅT OSS TA SOM EXEMPEL en ML-beräkning som syftar till att avgöra om data från en viss bild beskriver en katt eller en hund. När be-



I dessa exempel ligger datavolymerna för maskininlärning på allt mellan 850 kilobyte och 2 terabyte.

räkningen är klar skickas svaret tillbaka upp i nätverket och om det då sker en systemkrasch under processen går alla beräkningar förlorade.

Genom att byta till NVDIMM kan denna process effektiviseras rejält. Det finns inget behov av att mellanlagra inkommande data i SSD. Data kan flyttas raka vägen till DRAM och GPU:n kan omedelbart starta sina beräkningar. Svaret på frågan om bilden visar en hund eller en katt kan komma magnituder snabbare. Och det finns ingen risk att förlora data eller beräkningar eftersom NVDIMM är persistent.

NVDIMM är inte bara väl lämpat för AI och ML, utan också för finansiella tillämpningar, så kallad FinTech. Sådana tillämpningar kräver hög prestanda (låg latens och höga transaktionshastigheter) eftersom tid är pengar, helt enkelt.

AVSLUTADE TRANSAKTIONER måste loggas synkront innan nästa transaktion kan startas. Synkroniseringen är kritisk för revision men utgör en flaskhals för många system och sänker transaktionshastigheten. Genom att använda NVDIMM kan denna process för att logga data till SATA eller NVMe SSD elimineras. Istället för att skicka loggdata via IO till SSD-flash kan det läggas direkt i det höghastighets-DRAM som gjorts persistent via NVDIMM. De gör det möjligt att starta nästa transaktion med fullt förtroende för att den tidigare transaktionen loggats till en säker plats utan risk för att data går förlorade.

NVDIMM har funnits i mer än ett decennium men fördelarna med att använda denna typ av beständiga minnen för AI och ML är något som fortfarande testas inom olika användningsområden, från bank och detaljhandel till vård, yrkestjänster och verkstads- och processindustri.

Ekosystemet för NVDIMM, inklusive operativsystem, hårdvaruaktivering och JEDEC-standardisering uppstod en gång som resultat av att många företag arbetade tillsammans med att börja använda persistenta minnen. Nu korsar NVDIMM vägarna med AI och ML och visar sig vara ett perfekt sätt att öka deras systemprestanda. ■