BM-311 Bilgisayar Mimarisi

Hazırlayan: M.Ali Akcayol
Gazi Üniversitesi
Bilgisayar Mühendisliği Bölümü

Konular

- Giriş
- CUDA
- GPU ve CPU
- GPU mimarisi
- Intel Gen8 GPU
- Yardımcı işlemci olarak GPU

Giriş

- Graphical Processor Unit (GPU), **üç boyutlu grafiklerin** ve **video'ların** işlenmesini **hızlandırmak** için tasarlanmıştır.
- GPU günümüzde hemen hemen tüm iş istasyonlarında, laptop'larda, tablet'lerde ve akıllı telefonlarda vardır.
- Yüzlerce hatta **binlerce paralel işlemci core** bir entegre devrede yer alabilmektedir.
- **Büyük GPU sistemler** genellikle **PCIe** bus üzerinden iletişim yapan ayrı bir **yardımcı işlemci kartında yer alır.**
- **Küçük GPU sistemler** akıllı telefon veya tablet'lerde bulunur ve **birkaç tane işlemci core vardır.**
- GPU; biyoinformatik, moleküler analiz, petrol ve gaz arama, finansal uygulamalar, sinyal ve ses işleme, istatistiksel modelleme, bilgisayarla görme, medikal görüntüleme alanlarında kullanılmaktadır (General-Purpose GPU, GPGPU).

Konular

- Giriş
- CUDA
- GPU ve CPU
- GPU mimarisi
- Intel Gen8 GPU
- Yardımcı işlemci olarak GPU

CUDA

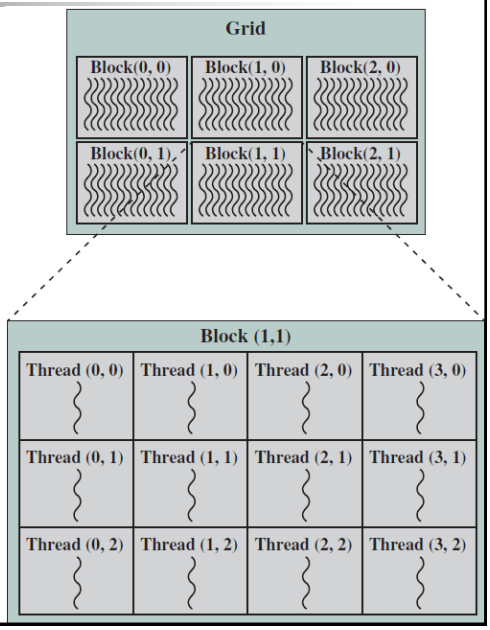
- Compute Unified Device Architecture (CUDA), NVIDIA tarafından geliştirilen ve GPU kullanan **paralel hesaplama platformu** ve **programlama modelidir.**
- CUDA; C, C/C++ tabanlı bir dildir.
- Bir CUDA programı genel olarak **3 kısma bölünebilir:**
  - **Host üzerinde çalışan kod (CPU)**
  - **Cihaz üzerinde çalışan kod (GPU)**
  - **Host ve cihaz arasında veri transferi yapan kod**
- **Host** üzerindeki **kod seri çalışır**, paralelleştirilemez.
- **GPU** üzerindeki **kod paralel çalışır**, kernel olarak adlandırılır.
- **Kernel'da branch deyimi bulunmaz** veya çok az olabilir.
- **Branch** deyimleri thread'lerin **seri çalışmasını gerektirir.**

CUDA

- **Programcı,** kernel fonksiyon çağırıldığında **kaç tane thread çalışacağını belirler.**
- GPU **işlemci core'ları** (CUDA cores) **maksimum verimde** kullanmak için **binlerce thread tanımlanır.**
- **Programcı** thread'lerin **nasıl gruplandırılacağını** da **belirler.**
- Thread'ler bloklar halinde gruplanır.
- **Her kernel için blok sayısı** grid **olarak adlandırılır.**
- Bir blok sadece **bir GPU** streaming multiprocessor'e (SM) atanır.

CUDA

- **İki boyutlu thread blokları ve iki boyutlu grid yapısı.**
- **Bir blok** sadece **bir GPU** streaming multiprocessor'e **(SM) atanır.**
- **Bir blok** SM'ler arasında **bölünmez.**
- **Tek blok olursa**, bir SM tüm işi yaparken diğerleri boş bekler.
- **Blok sayısı,** GPU üzerindeki SM sayısından az olmamalıdır.

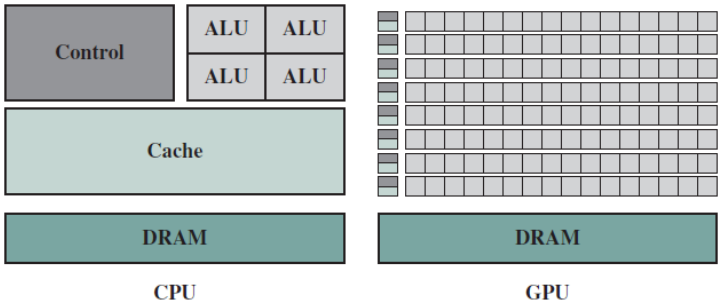


Konular

- Giriş
- CUDA
- GPU ve CPU
- GPU mimarisi
- Intel Gen8 GPU
- Yardımcı işlemci olarak GPU

GPU ve CPU

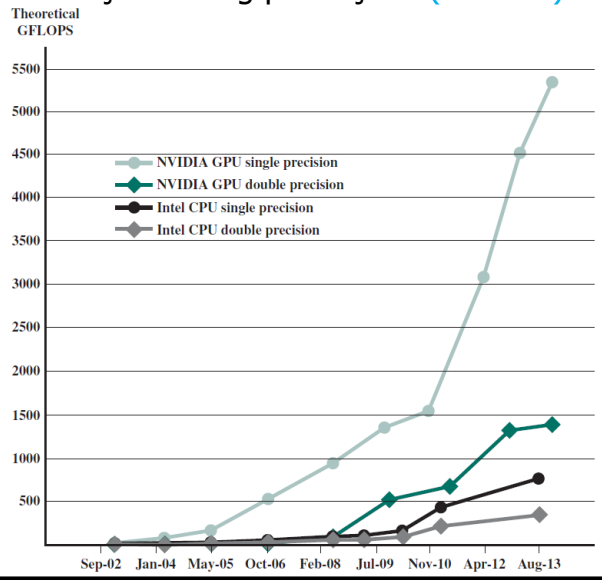
- **GPU** ve **CPU** farklı uygulamalar için tasarlandığından **mimarileri çok farklıdır.**
- GPU ve CPU **önbellek**, **kontrol birimi**, **işlem birimi** farklıdır.



- **CPU**'nun önemli bir kısmını **önbellek** ve **kontrol birimi** kaplar.
- **GPU**, matematiksel işlemler için **SIMD** mimarisi kullanır.
- **GPU karmaşık kontrol birimine** (out of order exec, branch prediction, data hazards, vb.) **ihtiyaç duymaz.**

GPU ve CPU

- GPU ve CPU'lar için floating point işlem (GFLOPS) sayıları.



Konular

- Giriş
- CUDA
- GPU ve CPU
- GPU mimarisi
- Intel Gen8 GPU
- Yardımcı işlemci olarak GPU

GPU mimarisi

- **1980-1990**'lı yıllarda **GPU** mimarisi sabit ve **programlanamaz yapıya sahipti ve özel amaçlı** işlem birimlerinden oluşmaktaydı.
- Daha sonraki yıllarda **paralel SIMD işlemcilerle GPU mimarisi geliştirilmiştir.**
- 2006 yılında **NVIDIA** tarafından **GPGPU** dili **CUDA geliştirilmiştir.**
- NVIDIA tarafından geliştirilen **GeForce 8800 GTX** ilk **GPGPU donanımıdır.**
- Genel amaçlı **uygulamaları paralel çalıştırmak için hiyerarşik önbellek** ve **paylaşılmış hafıza eklenmiştir.**
- Programlanabilir **GPU işlemci core'ları** eşit sayıda **SM'ye bölünmüştür.**

GPU mimarisi

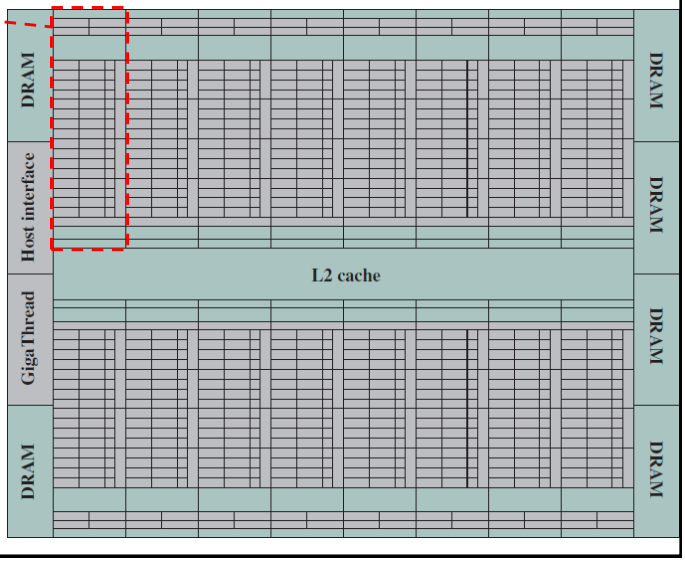
- **NVIDIA tarafından** Tesla, Fermi, Kepler ve Maxwell gibi **çok sayıda versiyon geliştirilmiştir.**
- SM mimarileri her versiyonda giderek geliştirilmiştir.
- NVIDIA Fermi mimarisi temel mimariyi yansıtır.
- **Fermi mimarisi 16 SM'den oluşur.**
- Fermi mimarisinde **her SM 32 CUDA core'a sahiptir.**
- **Fermi GPU**, 16 SM*32 CUDA core = **512 CUDA core'a** sahiptir.
- **Fermi GPU**, her SM'deki CUDA core sayısı az olduğu için **GPU donanımı ile CUDA yazılımı arasında eşleştirmeyi kolay yapar.**

GPU mimarisi

- Fermi **L2 önbellek** 16 SM'nin arasında yer alır.

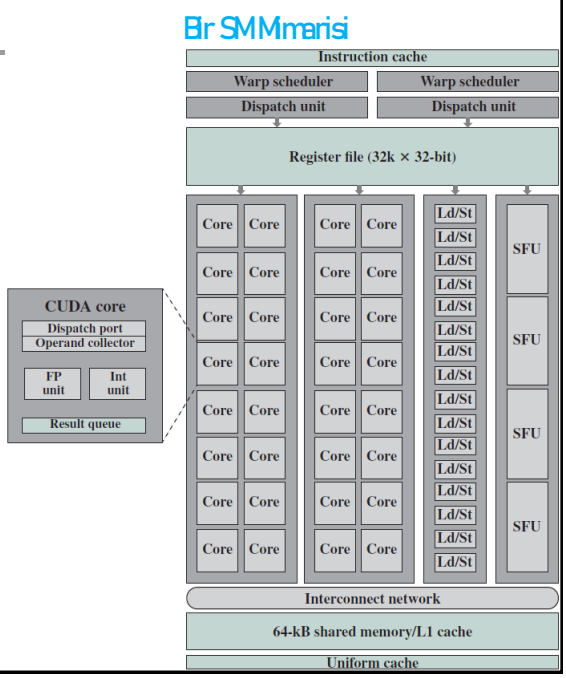
1 adet SM
Her SM'de;
- **32 core**
- **16 Ld/St** birimi
- **4** özel **işlem birimi**
- **6 tane 64-bit (384 bit) DRAM** I/O arayüzü
- Host interface ile **GPU ve CPU arasında PCIe bağlantı** sağlanır.
- **GigaThread scheduler,** thread bloklarını SM'lere dağıtır.

Fermi Mimarisi



GPU mimarisi

- Her SM'de **32 adet GPU core** (32 CUDA core)
- **2 adet Warp Scheduler/ Dispatch Unit**
- **16 adet Load/Store** birimi
- **4 adet özel işlem birimi** (SFU-special function units)
- **32k * 32 bit register**
- **Paylaşılmış L2** ve **L1 cache**



GPU mimarisi

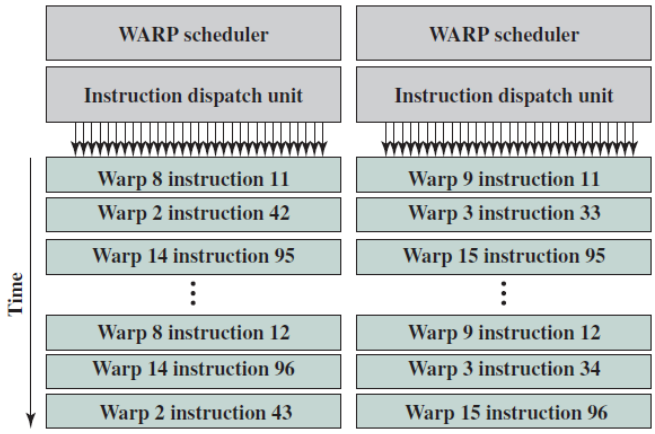
Dual warp scheduler

- GigaThread scheduler **thread bloklarını SM'lara dağıtır.**
- Dual warp scheduler thread bloklarını **32'şer thread halinde gruplandırır.**
- **Her thread** kendi **instruction address counter** ve **register kümesine** sahiptir.
- SM içindeki **her thread** bağımsız **branch** ve **execute** işlemi **yapabilir**.
- GPU performansı, **CUDA core'larının** maksimum **dolu olmasına bağlıdır** (SM donanımlarının yüksek verimli olması).
- Her bir **warp scheduler** ve **dispatch birimine, 16 CUDA core** (toplam 32 CUDA core) **atanır.**

GPU mimarisi

Dual warp scheduler

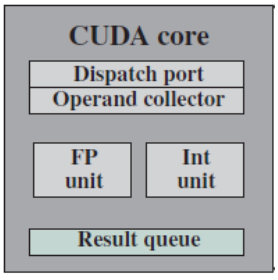
- Dual warp scheduler ve instruction dispatch birimleri



GPU mimarisi

CUDA cores

- NVIDIA **Fermi mimarisinde** GPU içerisinde **her SM'ye 32 adet CUDA core atanmıştır.**
- **Her CUDA core** ayrı **pipeline** ve **datapath'e sahiptir** (bir integer unit pipeline ve bir floating-point unit pipeline).
- **Int unit**, 32-bit, 64-bit **integer** ve **logic/bitwise işlemleri yapar.**
- **FP unit**, single-precision (bir CUDA core gerekir) veya double-precision (iki CUDA core gerekir daha uzun sürer) işlem yapar.



GPU mimarisi

Special function units

- Her SM **4 adet SFU**'ya sahiptir.
- SFU; **trigonometrik** ve **square root** gibi işlemleri bir clock cycle'da yapar.

Load/Store units

- **16 load/store birimi** ile SM kaynak ve hedef adresleri hesaplar.
- **Adresler,** cache veya DRAM üzerine **thread'lerin yazma** ve **okuma** yapması için kullanılır.

GPU mimarisi

Registers, shared memory ve L1 cache

- Her SM, register kümesi, shared memory/L1 cache sahiptir.
- **Fermi mimarisinde**, **her SM'de 32k*32-bit register** vardır, **her thread için 64*32-bit register atanabilir** (CUDA 2.x).
- **Register erişimi birkaç ns'dir** (ardından, L1, L2 ve memory erişimi yapılır).
- Bir thread'e atanan **register'daki verinin yaşam süresi, thread'in yaşam süresi kadardır.**
- Bir SM'deki **shared memory'deki verinin yaşam süresi, thread bloğunun yaşam süresi kadardır.**

GPU mimarisi

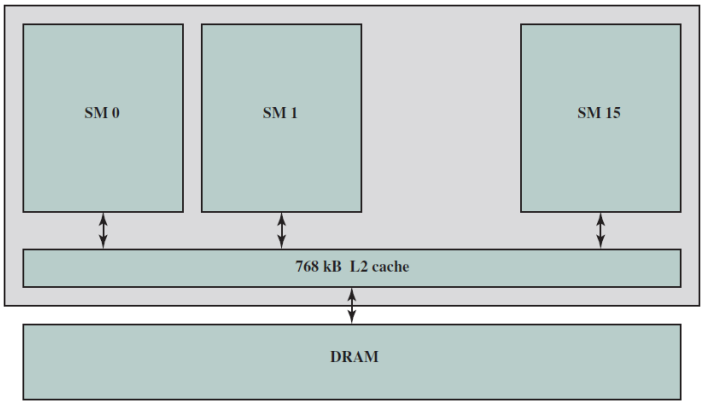
Registers, shared memory ve L1 cache

- Bir SM içindeki **GPU core'ları** adreslenebilir **on-chip shared memory'ye sahiptir** (multicore CPU'larda yoktur.)
- On-chip memory, GPGPU'ların off-chip memory ihtiyacını azaltır.
- **Shared memory boyutu** küçüktür, ancak global memory'ye göre **100-150 kat daha hızlıdır.**
- SM memory mimarisinde, **her SM için 64-kB L1 instruction cache** vardır.
- Toplam **128kB** (32k*32-bit) **register file** vardır.
- Bir bölümü **shared memory**, kalan kısım **data cache** olarak kullanılan **64kB L1 data cache** vardır.

GPU mimarisi

Registers, shared memory ve L1 cache

- Fermi memory mimarisinde, tüm SM'ler için shared **768 kB L2 unified cache** vardır.
- **DRAM** shared memory olarak **tüm SM'ler tarafından kullanılır.**

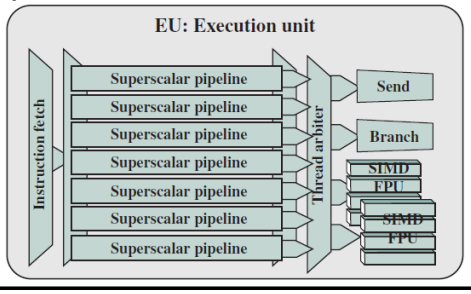


Konular

- Giriş
- CUDA
- GPU ve CPU
- GPU mimarisi
- Intel Gen8 GPU
- Yardımcı işlemci olarak GPU

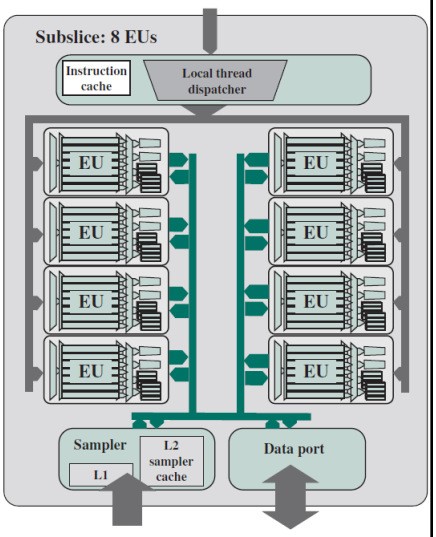
Intel Gen8 GPU

- **Gen8** mimarisinde temel yapısal blok **Execution Unit (EU)**'tir.
- EU, **7 thread** ile **simultaneous multithreading (SMT)** ve **superscalar pipeline** mimarisine sahiptir.
- **Her thread 128 genel amaçlı register'a** sahiptir.
- **Her EU'da**, SIMD floating-point ve integer hesaplamaları yapar.
- Branch unit ile branch; Send unit ile hafıza işlemleri yapılır.
- Her thread **4kB GPR** file'a, her EU **28 kB GPR** file'a sahiptir.
- **Thread arbiter,** her komutu 4 functional unit'ten birisine atar.



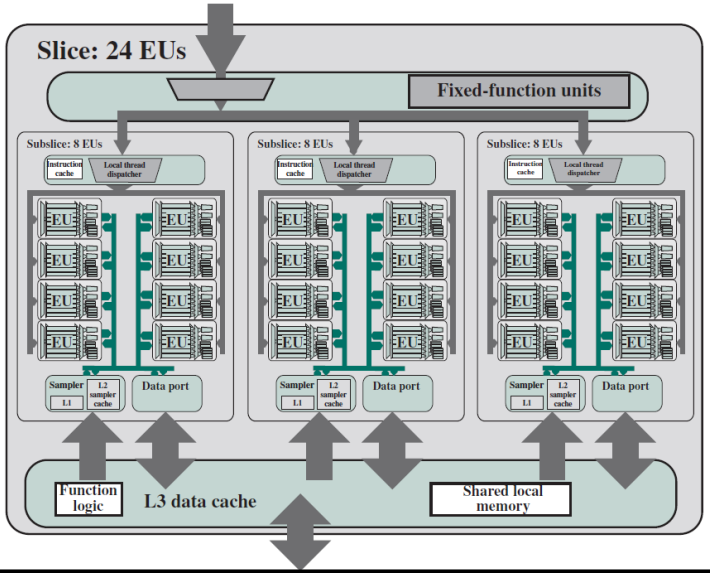
Intel Gen8 GPU

- EU'lar subslice şeklinde organize edilir **(her subslice'ta 8 EU)**.
- Subslice, **thread dispatcher** ve **instruction cache'e sahiptir.**
- **Her subslice** 56 eş zamanlı thread çalıştırabilir **(7 pipeline*8 EU)**.
- **Sampler**, görüntü yüzeyinde **örnekleme** yapmak için kullanılır.
- **Sampler**, farklı **filtreleme modlarına sahiptir** (point, bilinear, trilinear, anisotropic).
- **Dataport**, yazma/okuma yapar.



Intel Gen8 GPU

- **Subslice'larla slice oluşturulur** (Intel Gen8'de **3 subslice 24 EU)**.



Intel Gen8 GPU

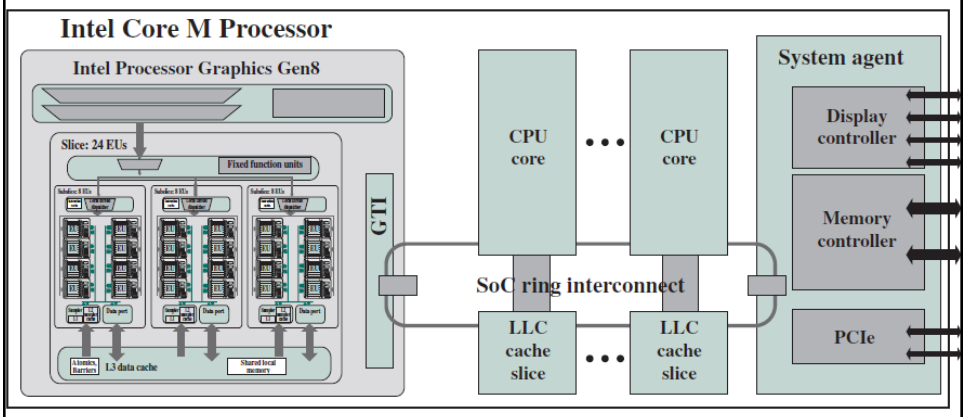
- Intel **Gen8 slice'ta**, **thread dispatch** ve **grafik veri işleme optimizasyonu** için **function logic** bulunur.
- **L3 cache** ve küçük boyutlu **shared local memory** vardır.
- Shared local memory ile EU'lar arasında geçici değişkenler paylaştırılır.
- Performansı artırmak için **cache banking tekniği kullanılır.**
- **Cache banking** ile cache eşit boyutlu **n adet parçaya** (bank) **bölünür.**
- Tüm cache parçalarına **eş zamanlı erişim** yapılır.
- **Aynı cache bank'i adresleyen erişimler ardışık yapılır.**

Intel Gen8 GPU

- SoC (System-on-Chip) mimarisinde, bir veya **birden fazla slice bir chip içerisinde** yer alır.
- Slice'lar ek bileşenlerle (3D rendering, media pipeline) birleştirilir.
- Intel **Gen8 mimarisi**, **GTI** (Graphics Technology Interface) **aracılığıyla diğer SoC bileşenleriyle birleştirilir.**
- **Intel Core M işlemci** (Intel HD Graphics 5300 Gen8) SoC mimarisine sahiptir.
- **Intel Core M**, GPU kısmına ek olarak **çok sayıda CPU core'a**, **LLC (Last Level Cache)'e**, **DRAM controller**, **display controller** ve **PCIe device**'lara sahiptir.
- **Gen8 işlemci**, CPU'lar, LLC cache, system agent (DRAM controller, display controller ve PCIe device'lar) arasında **ring şeklinde bir bağlantıya sahiptir.**

Intel Gen8 GPU

- Intel Core M işlemci



Konular

- Giriş
- CUDA
- GPU ve CPU
- GPU mimarisi
- Intel Gen8 GPU
- Yardımcı işlemci olarak GPU

Yardımcı işlemci olarak GPU

- **Bir GPU,** yüzlerce hatta **binlerce SIMD** mimarisine sahip **işlemci** core'una sahiptir.
- **Yüksek oranda paralel çalışabilen kod parçaları**, GPGPU sistemlerde **çok hızlı çalışırlar.**
- Binlerce **lightweight thread** ile büyük **veri kümelerinde eşzamanlı çalışan programlar**, GPGPU sistemlerde **çok hızlı çalışırlar.**
- **Lightweight thread'lerde** hemen hemen hiç **branch komutu yoktur.**
- Iterasyonları arasında **veri bağımlılığı olmayan çok büyük iterasyonlara sahip döngüler** (matris işlemleri) GPGPU sistemlerde **çok hızlı çalışır.**

Yardımcı işlemci olarak GPU

- Yüksek oranda paralel çalışabilecek bir seri kod parçası **derleyici tarafından paralel hale getirilmelidir.**
- **CUDA, OpenCL** vb. ile **seri kod** parçası **paralel çalışacak şekilde düzenlenebilir.**
- Compiler directive diller (**OpenACC, hiCUDA**, vb.) paralelleştirme için kullanılabilir.
- Compiler directive diller, paralelleştirilebilir kısımlar için **ipuçları olacak açıklamalar yerleştirir.**
- **Yeni** versiyon **CUDA derleyiciler**, OpenACC dil desteğine sahiptir.