

۷۰۰۶۷۷۷
به نام خدا



برنامه نویسی حرفه ای

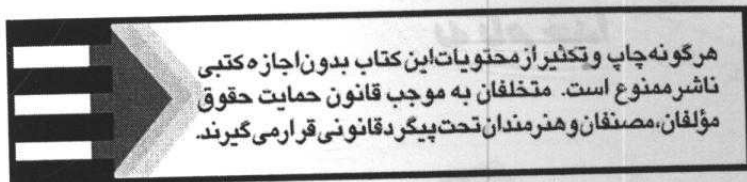
CUDA C

با مقدمه دکتر باران چمن
مرکز محاسبات و سیستم های پیشرفته دانشگاه هوستون

مترجمان:

دکتر صبا جودکی (استادیار - عضو هیات علمی دانشگاه آزاد اسلامی واحد خرم آباد)

مهندس ندا محمودی (کارشناس ارشد مهندسی کامپیوتر - نرم افزار)



هرگونه چاپ و تکثیر از محتویات این کتاب بدون اجازه کتبی
ناشر ممنوع است. متخلفان به موجب قانون حمایت حقوق
مؤلفان، مصنفان و هنرمندان تحت پیگرد قانونی قرار می گیرند.



عنوان کتاب: برنامه نویسی حرفه ای CUDA C

عنوان کتاب اصلی: Professional CUDA C programming

مترجمان: دکتر صبا جودکی

تهران: ندا محمودی

ناشر: موسسه فرهنگ سری دیباگران تهران

صفحه آرای: نازنین صیر

طراح جلد: داریوش فرسای

نوبت چاپ: اول

تاریخ نشر: ۱۳۹۷

چاپ و صحافی: درج عقیق

تیراژ: ۱۰۰۰ جلد

قیمت: ۱۴۲۰۰۰۰ ریال

شابک: ۹۷۸-۶۲۲-۲۱۸-۰۳۷-۹

نشانی واحد فروش: تهران، میدان انقلاب،

خ کارگر جنوبی، روبروی پاساژ مهستان،

پلاک ۱۲۵۱

تلفن: ۲۲۰۸۵۱۱۱-۶۶۴۱۰۰۴۶

فروشگاههای اینترنتی دیباگران تهران :

WWW.MFTBOOK.IR

www.dibagarantehran.com

www.mftdibagaran.ir

نشانی تلگرام: @mftbook

سرشناسه: چنگ، جان، cheng.jon

عنوان و نام پدیدآور: برنامه نویسی حرفه ای CUDA C /

تالیف: جان چنگ، ماکس گروسمن، تای مکرچر، با مقدمه باربارا

چپمن، مترجمان: صبا جودکی، ندا محمودی

مشخصات نشر: تهران: دیباگران تهران: ۱۳۹۷

مشخصات ظاهری: ۷۳۶ ص: مصور.

شابک: ۹۷۸-۶۲۲-۲۱۸-۰۳۷-۹

وضعیت فهرست نویسی: فیا

داشت: عنوان اصلی: professional cuda c programming

موضوع: برنامه نویسی موازی (computer science) parallel programming

موضوع: نواحی، بخش گرافیک-برنامه نویسی

موضوع: graphics processing units-programming

موضوع: نرم افزار، توسعه (application software-development)

موضوع: کودا (computer architecture) cuda

موضوع: نواحدهای پردازش، فیک: graphics processing unit

شناسه افزوده: گروسمن، ماکس- grossman, max

شناسه افزوده: مکرچر، تای- mc, cher, ty

شناسه افزوده: چپمن، باربارا، مقدمه نویسنده: chapman, barbara

شناسه افزوده: جودکی، صبا، ۱۳۶۰- مترجم

شناسه افزوده: محمودی، ندا، ۱۳۶۸- مترجم

رده بندی کنگره: ۴۱۳۹۷: ج۹ ۷۶/۶۴۲ QA

رده بندی دیویی: ۰۰۵/۲۷۵

شماره کتابشناسی ملی: ۵۴۴۲۰۹۶

اپلیکیشن دیباگران تهران را از سایت های اینترنتی دیباگران دریافت نمایید.

۱۹	مقدمه ناشر
۲۰	درباره نویسندگان
۲۳	سخن مترجم
۲۴	پیشگفتار
۲۶	برف
۲۷	مقدمه

فصل اول / محاسبات موازی ناهمگن با کودا

۳۶	محاسبات درازی
۳۸	برنامه نویسی تریبی و واری
۴۰	موازات
۴۲	معماری کامپیوتر
۴۶	محاسبات ناهمگن
۴۷	معماری ناهمگن
۵۰	الگوی محاسبات ناهمگن
۵۳	کودا: پلتفرمی برای محاسبات ناهمگن
۵۷	HELLO WORLD از GPU
۶۱	آیا برنامه نویسی CUDA دشوار است؟
۶۴	خلاصه
۶۵	تمرین های فصل ۱

فصل دوم / مدل برنامه نویسی کودا

۶۷	معرفی مدل برنامه نویسی کودا
----	-----------------------------

۶۹		ساختار برنامه نویسی کودا
۷۱		مدیریت حافظه
۷۸		سازماندهی نخ ها
۸۶		فراخوانی یک کرنل کودا
۸۸		نوشتن کرنل شما
۹۰		بررسی کرنل شما
۹۱		مدیریت طلاها
۹۲		کامپایل اجرا
۹۶		زمانبندی کرنل شما
۹۷		زمانبندی با تایمر C++
۱۰۳		زمانبندی با nvprof
۱۰۵		سازماندهی نخ های موازی
۱۰۶		اندیس گذاری ماتریس ها با بلاک ها و نخ ها
۱۱۰		مجموع ماتریس ها با گرید دو بعدی و بلاک دو بعدی
۱۱۷		مجموع ماتریس ها با گرید یک بعدی و بلاک های یک بعدی
۱۱۹		مجموع ماتریس ها با گرید دو بعدی و بلاک های یک بعدی
۱۲۲		مدیریت دستگاه ها
۱۲۳		استفاده از API زمان اجرا برای پرسوجوی اطلاعات GPU
۱۲۸		تعیین بهترین GPU
۱۲۹		استفاده از nvidia-smi برای پرسوجوی اطلاعات GPU
۱۳۰		تنظیمات دستگاه در زمان اجرا
۱۳۱		خلاصه
۱۳۲		تمرین های فصل ۲

۱۳۳	معرفی مدل اجرایی کودا
۱۳۴	بررسی معماری GPU
۱۳۹	معماری Fermi
۱۴۲	معماری Kepler
۱۴۷	بهینه سازی مبتنی بر پروفایل
۱۵۰	تأیید ماهیت اجرای چند تار
۱۵۰	تارها و بلاک های نخ
۱۵۲	انشعاب چند تار
۱۶۲	تقریب بندی منابع
۱۶۵	نادیده گرفتن ناخ
۱۷۰	مشغولیت
۱۷۵	همگام سازی
۱۷۷	مقیاس پذیری
۱۷۸	نمایش موازات
۱۸۰	بررسی چند تارهای فعال با nvprof
۱۸۲	بررسی عملیات حافظه با nvprof
۱۸۳	ارائه موازات بیشتر
۱۸۸	اجتناب از انشعاب
۱۸۸	مساله کاهش توازی
۱۹۱	انشعاب در کاهش توازی
۱۹۸	بهبود انشعاب در کاهش توازی
۲۰۲	کاهش با جفت های درهم تنیده
۲۰۵	باز کردن حلقه ها

۲۰۷	کاهش بار کردن
۲۱۰	کاهش بار چند تارهای بار شده
۲۱۴	کاهش بار کردن کامل
۲۱۶	کاهش بار توابع نمونه
۲۲۰	تواری پویا
۲۲۱	اجرای تودرتو
۲۲۳	برنامه Hello World تودرتو در GPU
۲۳۰	کاهش بار
۲۳۸	خلاصه
۲۳۹	تمرین های فصل ۳

فصل ۴: بارم / حافظه سراسری

۲۴۲	معرفی مدل حافظه کودا
۲۴۳	مزایای سلسله مراتب حافظه
۲۴۴	مدل حافظه کودا
۲۴۶	نبات ها
۲۴۸	حافظه محلی
۲۴۸	حافظه اشتراکی
۲۵۰	حافظه ثابت
۲۵۱	حافظه بافت
۲۵۱	حافظه سراسری
۲۵۲	کش های GPU
۲۵۴	حافظه سراسری ایستا
۲۵۷	مدیریت حافظه
۲۵۷	تخصیص و آزادسازی حافظه

- ۲۵۸ انتقال حافظه
- ۲۶۱ حافظه متصل
- ۲۶۳ حافظه Zero-Copy
- ۲۷۰ آدرس دهی مجازی متحد
- ۲۷۲ حافظه متحد
- ۲۷۳ الگوهای دستیابی به حافظه
- ۲۷۴ دسترسی متصل و تراز شده
- ۲۷۶ خواندن از حافظه سراسری
- ۲۷۸ بازیابی های کش شده
- ۲۸۰ بازیابی های کش نشده
- ۲۸۲ مثالی از خواندن تراز نشده
- ۲۸۷ کش فقط خواندنی
- ۲۸۸ نوشتن در حافظه سراسری
- ۲۸۹ مثالی از نوشتن تراز نشده
- ۲۹۱ آرایه ای از ساختارها در مقابل ساختاری از آرایه ها
- ۲۹۳ مثال: مثال ریاضی با طرح بندی داده AOS
- ۲۹۶ مثال: مثال ریاضی با طرح بندی داده SOA
- ۲۹۸ تنظیم کارایی
- ۲۹۸ روش های باز کردن
- ۳۰۰ ارائه موازات بیشتر
- ۳۰۲ کرنل در چه پهنای باندی میتواند اجرا شود؟
- ۳۰۲ پهنای باند حافظه
- ۳۰۳ مسأله ترانهاده ماتریس
- ۳۰۵ تنظیم یک حد کارایی بالا و پایین برای کرنل های ترانهاده

۳۱۰	ترانهاده ساده: خواندن سطرها در مقابل خواندن ستون ها
۳۱۵	ترانهاده قطری: خواندن سطرها در مقابل خواندن ستون ها
۳۲۰	ارائه موازات بیشتر با بلاک های سبک
۳۲۲	جمع ماتریس با حافظه متحد
۳۲۹	خلاصه
۳۳۰	تمرین های فصل ۴

فصل پنجم / حافظه اشتراکی و حافظه ثابت

۳۳۴	معرفی حافظه اشتراکی کوانتا
۳۳۵	حافظه اشتراکی
۳۳۷	تخصیص حافظه اشتراکی
۳۳۸	بانک های حافظه اشتراکی و حالت دسترسی
۳۳۸	بانک های حافظه
۳۳۹	ناسازگاری بانک
۳۴۱	حالت دسترسی
۳۴۵	لایه بندی حافظه
۳۴۶	پیکربندی حالت دسترسی
۳۴۷	پیکربندی میزان حافظه اشتراکی
۳۴۹	همگام سازی
۳۵۰	مدل حافظه ضعیف
۳۵۰	مانع صریح
۳۵۱	حفاظ حافظه
۳۵۲	توصیف کننده فرار
۳۵۳	بررسی طرح داده حافظه اشتراکی
۳۵۴	حافظه اشتراکی مربعی

۳۵۵		دستیابی سطری در مقابل ستونی
۳۵۹		نوشتن سطری و خواندن ستونی
۳۶۰		حافظه اشتراکی پویا
۳۶۲		لایه بندی ایستا حافظه اشتراکی اعلان شده
۳۶۳		لایه بندی پویا حافظه اشتراکی اعلان شده
۳۶۵		مقایسه عملکرد کرنل های حافظه اشتراکی مربعی
۳۶۶		حافظه اشتراکی مستطیلی
۳۶۷		دستیابی سطری در مقابل دستیابی ستونی
۳۶۸		نوشتن سطری و خواندن ستونی
۳۷۰		حافظه اشتراکی اعلان شده به صورت پویا
۳۷۱		لایه بندی حافظه اشتراکی اعلان شده به صورت ایستا
۳۷۳		لایه بندی حافظه اشتراکی اعلان شده به صورت پویا
۳۷۴		مقایسه عملکرد کرنل های حافظه اشتراکی مستطیلی
۳۷۵		کاهش دستیابی به حافظه سراسری
۳۷۶		کاهش موازی با حافظه اشتراکی
۳۸۱		کاهش موازی با باز کردن
۳۸۴		کاهش موازی با حافظه اشتراکی پویا
۳۸۴		پهنای باند موثر
۳۸۵		دستیابی های حافظه سراسری متصل
۳۸۶		کرنل ترانهاده پایه
۳۸۸		ترانهاده ماتریس با حافظه اشتراکی
۳۹۳		ترانهاده ماتریس با حافظه اشتراکی لایه بندی شده
۳۹۴		ترانهاده ماتریس با باز کردن
۳۹۸		نمایش موازات بیشتر

۴۰۰	حافظه ثابت
۴۰۱	پیاده سازی یک الگوی یک بعدی با حافظه ثابت
۴۰۴	مقایسه با کش فقط خواندنی
۴۰۷	دستورالعمل Warp Shuffle
۴۰۸	انواع دستورالعمل های warp shuffle
۴۱۱	اشتراک گذاری داده درون یک چند تار
۴۱۲	توزیع یک مقدار در میان یک چند تار
۴۱۲	بافت راست در یک چند تار
۴۱۳	شیفت به چپ در یک چند تار
۴۱۴	شیفت چرخشی در یک چند تار
۴۱۵	تعویض پروانه ای در میان چند تارها
۴۱۵	مبادله مقادیر یک آرایه در میان یک چند تار
۴۱۶	مبادله مقادیر با استفاده از اندیس های آرایه در میان یک چند تار
۴۱۸	کاهش توازی با استفاده از دستورالعمل warp shuffle
۴۲۱	خلاصه
۴۲۲	تمرین های فصل ۵

فصل ششم / جریان ها و همزمانی

۴۲۷	معرفی جریان ها و رویدادها
۴۲۹	جریان های کوبا
۴۳۳	زمانبندی جریان
۴۳۴	وابستگی های غلط
۴۳۴	Hyper-Q
۴۳۵	اولویت های جریان
۴۳۶	رویدادهای کوبا

۴۳۶	ایجاد و انحلال
۴۳۷	ثبت رویدادها و اندازه گیری زمان سپری شده
۴۳۹	همگام سازی جریان ها
۴۴۰	جریان های مسدود کننده و غیر مسدود کننده
۴۴۳	رویدادهای قابل پیکربندی
۴۴۴	اجرای کرنل همزمان
۴۴۴	کرنل های همزمان در جریان های Non-NULL
۴۴۷	وابستگی های غلط در GPUهای Fermi
۴۵۰	توزیع عملیات با OpenMP
۴۵۲	تنظیم فشار جریان با استفاده از متغیرهای محیطی
۴۵۴	همزمانی - محاسبه با GPU
۴۵۵	مسدود کردن رفتار جریان پس فرض
۴۵۶	ایجاد وابستگی ها در جریان
۴۵۸	همپوشانی اجرای کرنل و انتقال داده
۴۵۹	همپوشانی با استفاده از زمانبندی اول عمق
۴۶۳	همپوشانی با استفاده از زمانبندی اول عرض
۴۶۵	همپوشانی اجرای CPU و GPU
۴۶۶	Callback جریان ها
۴۶۹	خلاصه
۴۷۱	تمرین های فصل ۶

فصل هفتم / تنظیم اولویت های سطح دستورالعمل

۴۷۴	معرفی دستورالعمل های کوندا
۴۷۵	دستورالعمل های ممیز شناور
۴۷۸	توابع درونی و استاندارد

۴۷۹	 دستورالعمل های اتمیک
۴۸۳	 بهینه سازی دستورالعمل ها برای برنامه تان
۴۸۳	 دقت معمولی در مقابل دقت مضاعف
۴۸۵	 خلاصه
۴۸۶	 توابع استاندارد در برابر توابع درونی
۴۸۶	 تجسم توابع استاندارد و درونی
۴۸۹	 دستکشی توابع دستورالعمل
۴۹۵	 خلاصه
۴۹۶	 درک در خوارانه های اتمیک
۴۹۶	 اصول اولیه
۴۹۹	 توابع اتمیک درونی کودا
۵۰۰	 هزینه عملیات اتمیک
۵۰۲	 محدود کردن هزینه کارایی عملیات اتمیک
۵۰۳	 پشتیبانی از ممیز شناور اتمیک
۵۰۴	 خلاصه
۵۰۵	 ترکیب تمام این موارد با یکدیگر
۵۰۹	 خلاصه
۵۱۰	 تمرین های فصل ۷

فصل هشتم / OpenACC و کتابخانه های کودای تسریع دهنده GFortran

۵۱۴	 معرفی کتابخانه های کودا
۵۱۵	 حوزه های پشتیبانی شده برای کتابخانه های کودا
۵۱۷	 جریان کاری متداول کتابخانه
۵۲۱	 کتابخانه CUSPARSE
۵۲۳	 قالب های ذخیره سازی داده cuSPARSE

۵۲۲	متراکم
۵۲۳	مختصات (COO)
۵۲۴	قالب سطر تنک متراکم (CSR)
۵۲۷	خلاصه ای از قالب های ذخیره سازی cuSPARSE
۵۲۸	تغییر قالب بندی با cuSPARSE
۵۳۰	نمایش cuSPARSE
۵۳۲	عناوین مهم در توسعه cuSPARSE
۵۳۳	نماد cuSPARSE
۵۳۳	کتابخانه cuBLAS
۵۳۵	مدیریت داده cuBLAS
۵۳۷	تشریح cuBLAS
۵۳۷	مثالی ساده از cuBLAS
۵۳۹	تبدیل از BLAS
۵۴۰	عناوین مهم در توسعه cuBLAS
۵۴۱	خلاصه cuBLAS
۵۴۲	کتابخانه cuFFT
۵۴۳	استفاده از cuFFT API
۵۴۴	تشریح cuFFT
۵۴۵	خلاصه cuFFT
۵۴۶	کتابخانه cuRAND
۵۴۶	انتخاب اعداد تصادفی ساختگی یا اعداد شبه تصادفی
۵۴۸	بررسی کتابخانه cuRAND
۵۴۸	پوشش مفاهیم
۵۵۳	مقایسه API های میزبان و دستگاه

۵۵۳	تشریح cuRAND
۵۵۷	تولید مقادیر تصادفی برای کرنل های کودا
۵۵۹	عناوین مهم در توسعه cuRAND
۵۵۹	ویژگی های کتابخانه کودا معرفی شده در کودا ۶
۵۵۹	کتابخانه های کودا Drop-In
۵۶۱	کتابخانه های چند GPU یی
۵۶۴	بررسی کارای کتابخانه کودا
۵۶۴	cuSOLVER در مقابل MKL
۵۶۵	cuBLAS در مقابل MKL
۵۶۷	cuFFT در مقابل FFTW در مقابل MKL
۵۶۸	خلاصه کارایی کتابخانه کودا
۵۶۸	بکارگیری OpenACC
۵۷۳	بکارگیری دستورات محاسبات OpenACC
۵۷۳	بکارگیری دستورات کرنل ها
۵۷۷	بکارگیری دستور موازی سازی
۵۷۹	بکارگیری دستور حلقه
۵۸۴	خلاصه دستورات محاسباتی OpenACC
۵۸۴	بکارگیری دستورات داده OpenACC
۵۸۴	بکارگیری دستور داده
۵۹۰	افزودن شروط Data به دستورات parallel و kernels
۵۹۱	API زمان اجرای OpenACC
۵۹۴	ترکیب کتابخانه های کودا و OpenACC
۵۹۷	خلاصه ای از OpenACC
۵۹۸	خلاصه

فصل نهم / برنامه نویسی چند GPU ای

۶۰۳	حرکت به سمت چندین GPU
۶۰۶	اجرا بر روی چند GPU
۶۰۸	ارتباط نظیر به نظیر
۶۰۹	فعالسازی دسترسی نظیر
۶۱۰	کپی حافظه نظیر به نظیر
۶۱۰	همزمان سازی میان چند GPU
۶۱۱	تقسیم محاسبات بین چند GPU
۶۱۱	تخصیص حافظه بر چندین دستگاه
۶۱۲	توزیع کار از یک میزبان منحصر به فرد
۶۱۵	ارتباط نظیر به نظیر بر روی چندین GPU
۶۱۵	فعالسازی دسترسی نظیر به نظیر
۶۱۶	کپی حافظه نظیر به نظیر
۶۱۸	دسترسی حافظه نظیر به نظیر با آدرس دهی مجازی
۶۲۰	تفاوت محدود در چند GPU
۶۲۱	محاسبه الگو برای معادله موج ۲ بعدی
۶۲۲	الگوهای نمونه برای برنامه های چند GPU ای
۶۲۴	محاسبات الگو دو بعدی با چندین GPU
۶۲۷	همپوشانی محاسبه و ارتباط
۶۲۹	کامپایل و اجرا
۶۳۲	توزین برنامه ها در میان کلاسترهای GPU
۶۳۴	انتقال داده CPU به CPU
۶۳۴	پیاده سازی ارتباط MPI درون گره های

۶۳۷	وابستگی به CPU
۶۳۸	انتقال داده GPU به GPU با استفاده از MPI متداول
۶۳۸	وابستگی در برنامه های MPI-CUDA
۶۴۰	پایاده سازی ارتباط GPU به GPU با MPI
۶۴۲	انتقال داده GPU به GPU با MPI آگاه از کودا
۶۴۴	انتقال داده GPU به GPU درون گره با MPI آگاه از کودا
۶۴۵	تنظ. بدازه قطعه پیام
۶۴۶	انتقال داده GPU به GPU با GPUDirect RDMA
۶۵۰	خلاصه
۶۵۱	تمرین های فصل ۹

فصل نهم / ملاحظات پیاده سازی

۶۵۴	فرایند توسعه کودا C
۶۵۵	چرخه توسعه APOD
۶۵۶	ارزیابی
۶۵۷	موازی سازی
۶۵۸	بهینه سازی
۶۵۹	توسعه
۶۵۹	فرصت های بهینه سازی
۶۶۱	بهینه سازی دسترسی به حافظه
۶۶۳	بهینه سازی اجرای دستورالعمل
۶۶۴	کامپایل کد کودا
۶۶۶	کامپایل مجزا
۶۶۹	یکپارچه سازی فایل های کودا در یک پروژه C
۶۷۰	مدیریت خطاهای کودا

۶۷۲	 بهینه سازی مبتنی بر پروفایل
۶۷۳	 کشف فرصت‌های بهینه سازی با استفاده از nvprof
۶۷۳	 مدهای پروفایل
۶۷۴	 قلمرو پروفایل
۶۷۵	 پهنای باند حافظه
۶۷۵	 الگوی دسترسی به حافظه سراسری
۶۷۷	 ناسازگاری های بانک حافظه اشتراکی
۶۷۸	 سر، ز ثبات
۶۷۸	 توان عملیات، دستورالعمل
۶۷۹	 راهنمای بهینه سازی با استفاده از nvvp
۶۸۰	 تحلیل هدایت شده
۶۸۱	 تحلیل هدایت نشده
۶۸۲	 تعمیم ابزارهای انویدیا
۶۸۵	 اشکال زدایی کودا
۶۸۵	 اشکال زدایی کرنل و اشکال زدایی حافظه
۶۸۶	 اشکال زدایی کرنل
۶۸۶	 بکارگیری cuda-gdb
۶۸۷	 CUDA Focus
۶۸۸	 بررسی حافظه کودا
۶۸۹	 بدست آوردن اطلاعات محیط
۶۹۰	 پارامترهای قابل تنظیم اشکال زدایی کودا
۶۹۱	 تجربه CUDA-GDB
۶۹۴	 خلاصه CUDA-GDB
۶۹۵	 CUDA PRINF

۶۹۶	CUDA assert
۶۹۷	خلاصه اشکال زدایی کرنل
۶۹۸	اشکال زدایی حافظه
۶۹۸	کامپایل برای cuda-memcheck
۶۹۹	memcheck
۷۰۱	racecheck
۷۰۷	خلاصه اشکال زدایی
۷۰۸	مطالعه مورد: تبدیل برنامه های C به کود C
۷۰۹	تعیین رمز
۷۱۰	موازی سازی رمز
۷۱۲	بهینه سازی رمز
۷۲۲	توسعه رمز
۷۲۲	رمز چند GPU
۷۲۳	رمز پیوندی OpenMP-CUDA
۷۲۶	خلاصه ای از تبدیل رمز
۷۲۸	خلاصه
۷۲۹	تمرین های فصل ۱۰
۷۳۰	ضمیمه: مطالعات پیشنهادی

GPUها راه طولانی را طی کرده‌اند. در ابتدا به عنوان پردازنده‌های گرافیکی که می‌توانستند به سرعت تصاویر را به صورت خروجی برای واحد نمایش تولید کنند مطرح شدند، و رفته‌رفته به فناوری تبدیل شدند که برای پردازش‌های فراسریع نیاز می‌شوند. در چند سال گذشته، GPUها به طور فزاینده برای تسریع یک آرایه محاسباتی به نام محاسبات ناهمگن به CPU پیوستند. امروزه، GPUها بر روی بسیاری از سیستم‌های رومیزی، بر روی کلاسترهای محاسباتی، و حتی بر روی بسیاری از بزرگترین ابرکامپیورها، جهان پیکربندی می‌شوند. در نقش توسعه‌یافته آن‌ها به عنوان فراهم‌کننده حجم بزرگی از سرعت محاسباتی برای محاسبات فنی، GPUها در علم و مهندسی در حوزه‌های گوناگون پیشرفت کرده‌اند. آن‌ها همچنین در عین حفظ قدرت، تعداد بزرگی هسته محاسباتی را برای کار به صورت موازی ایجاد کرده‌اند.

خوشبختانه، رابط‌های برنامه‌نویسی GPUها با این تغییرات سریع حفظ می‌شود. در گذشته، تلاش اصلی برای استفاده از آن‌ها در خارج از محدوده برنامه‌ها نیازمند این بود که برنامه‌نویس GPU با بسیاری از مفاهیمی که برنامه‌نویس گرافیکی از آن‌ها آگاهی دارد آشنا باشد. سیستم‌های امروزی مفاهیم بسیار ساده‌تری برای ایجاد برنامه‌ها، نرم‌زاری که بر روی آن‌ها ایجاد می‌شوند دارند. به طور خلاصه، ما کودا را داریم.

کودا یکی از محبوب‌ترین رابط‌های برنامه‌نویسی کاربردی برای تسریع حوزه محاسبات کرنل در GPU است. که می‌تواند کد نوشته شده در C یا C++ را برای اجرا با GPU با تلاش برنامه‌نویسی خیلی مناسبی امکان‌پذیر کند. همچنین توازن بین نیاز به اشتراک‌گذاری به منظور بهره‌گیری کامل از آن، و نیاز به داشتن یک رابط برنامه‌نویسی که استفاده از آن آسان باشد و برنامه قابل خواندنی را نتیجه دهد، بر هم می‌زند.

این کتاب منبع ارزشمندی برای افرادی است که می‌خواهند GPUها را برای برنامه‌نویسی علمی و فنی استفاده کنند. و معرفی جامعی از رابط برنامه‌نویسی کودا و بکارگیری آن ارائه می‌کند. برای شروع، اصول محاسبات موازی بر روی معماری‌های ناهمگن را شرح می‌دهد و ویژگی‌های کودا را بیان می‌کند. سپس نحوه اجرای برنامه‌های کودا را توضیح می‌دهد. کودا مدل اجرایی و حافظه را به برنامه‌نویس نمایش می‌دهد؛ در نتیجه برنامه‌نویس کودا کنترل مستقیمی بر روی محیط به شدت موازی دارد. علاوه بر ارائه جزئیات مدل حافظه کودا، متن کتاب اطلاعاتی در مورد نحوه بهینه‌سازی آن ارائه می‌دهد. فصل بعدی در مورد جریان‌ها بحث می‌کند، همچنین نحوه اجرای همزمان و همپوشانی کرنل‌ها را مورد بحث قرار می‌دهد. سپس اطلاعاتی در مورد تنظیم، در استفاده از کتابخانه‌های کودا، و استفاده از رهنمودهای OpenACC برای GPUها ارائه می‌شود. بعد از فصل برنامه‌نویسی چند GPU، با بحث در مورد برخی ملاحظات پیکربندی کتاب به پایان می‌رسد. با این حال، گستره‌ای از مثال‌ها برای

کمک به خواننده به منظور شروع کار، که بسیاری از آن‌ها قابل دانلود و اجرا هستند ارائه می‌شود. کودا توازن خوبی بین صراحت و قابلیت برنامه‌نویسی ارائه می‌کند که در عمل اثبات شده است. با این حال، کسانی که آن را مأمور ساده‌سازی توسعه برنامه‌شان کرده‌اند بایستی بدانند که این یک داستان در حال پیشرفت است. در چند سال گذشته، پژوهشگران کودا بر روی بهبود ابزارهای برنامه‌نویسی ناهمگن کار کرده‌اند. کودا ۶ چندین ویژگی جدید، از جمله حافظه متحد و کتابخانه‌های پلاگین، برای آسان کردن برنامه‌نویسی GPU معرفی می‌کند. آن‌ها همچنین مجموعه‌ای از دستورات به نام OpenACC را نیز ارائه کرده‌اند، که در این کتاب معرفی می‌شود. OpenACC تکمیل کودا را به واسطه ارائه مفاهیم ساده‌تر برای بهره‌گیری از قدرت برنامه‌نویسی کودا برای اجرایی که نیاز به کنترل مستقیم کمتری دارند وعده می‌دهد. عجلتا نتایج نویدبخش‌تر هستند. OpenACC و کودا ۶، و دیگر عناوین پوشش داده شده در این کتاب به توسعه‌دهندگان کودا امکان می‌دهند تا برنامه‌هایشان را برای حداکثر نمودن کارایی ترغیب دهند. داشتن این کتاب در یک محل ثابت در قفسه کتاب‌هایتان ضروری است. شاد برنامه‌نویسی کنید!

BARBARA CHAPMAN
CACDS and Department of Computer Science
University of Houston