ERAD-SP

ERAD-SP 2021

Anais da
12ª Escola Regional de Alto Desempenho
de São Paulo

6 a 8 de maio de 2021

Organização

UFABC

EACH USP

Realização

SBC

Patrocínio

aws

Atos

LexisNexis® RISK SOLUTIONS

SBC

COMITÊS

Coordenação Geral

Daniel Cordeiro (EACH/USP)
Emilio Francesquini (UFABC)

Coordenação de Programa e Minicursos

Hélio Crestana Guardia (UFSCAR)
Ricardo Menotti (UFSCAR)

COMITÊ DE PROGRAMA

Aleardo Manacero Jr. (UNESP)
Alexandro Baldassin (UNESP)
Alfredo Goldman (USP)
Calebe Bianchini (Mackenzie)
Daniel Cordeiro (USP)
Denis Fantinato (UFABC)
Denise Stringhini (Unifesp)
Emilio Francesquini (UFABC)
Hermes Senger (UFSCAR)
Herve Yviquel (Unicamp)
Jairo Panetta (ITA)
João Vicente Ferreira Lima (UFSM)
Laércio Lima Pilla (CNRS)
Liria Sato (USP)
Lucas Mello Schnorr (UFRGS)
Lucas Wanner (Unicamp)
Luciana Arantes (Sorbonne Université)
Luiz Bovolenta (UNESP)
Luiz Fernando Bittencourt (Unicamp)
Marco Netto (IBM Research)
Marcos Amaris (UFPA)
Márcio Castro (UFSC)
Paulo Lopes de Souza (USP)
Paulo Souza (Atrio)
Philippe Olivier Alexandre Navaux (UFRGS)
Raphael Camargo (UFABC)
Renato Ishii (UFMS)
Ricardo Santos (UFMS)
Rogerio Iope (UNESP)
Sandro Rigo (Unicamp)
Sarita Bruschi (USP)
Vanderlei Bonato (USP)

SBC

CRAD-SP

Comissão Regional de Alto Desempenho do Estado de São Paulo

Região Capital I: Emilio Francesquini (UFABC, Santo André) **- Presidente**

Região Capital II: Calebe Bianchini (Mackenzie, São Paulo)

Região Interior I: Celso Luis Mendes (INPE, S.J. dos Campos)

Região Interior II: Sarita Mazzini Bruschi (ICMC-USP, São Carlos) **- Presidente**

Região Interior III: Aleardo Manacero (DCCE/UNESP, S. J. do Rio Preto)

Coordenadores Gerais da ERAD-SP'20: Calebe Bianchini (Mackenzie), Rogerio Iope (Unesp)

Coordenadores Gerais da ERAD-SP'21: Daniel Cordeiro (USP), Emilio Francesquini (UFABC)

Mensagem dos Coordenadores

A Escola Regional de Alto Desempenho de São Paulo (ERAD-SP) é um evento anual realizado pela Sociedade Brasileira de Computação (SBC) desde 2010. A ERAD-SP é organizada pela Comissão Regional de Alto Desempenho do Estado de São Paulo (CRAD-SP), com a chancela da Comissão Especial de Arquitetura de Computadores e Processamento de Alto Desempenho (CE-ACPAD) da SBC.

A XII edição da ERAD-SP foi organizada pela Universidade Federal do ABC e pela Universidade de São Paulo. O evento foi realizado no período de 6 a 8 de maio de 2021 e, pela segunda vez em sua história, suas sessões ocorreram de maneira inteiramente online, devido à pandemia de COVID-19.

A ERAD-SP tradicionalmente exerce dois papéis importantes, o de Escola de Computação e o de fórum de encontro da comunidade de Computação de Alto Desempenho do Estado de São Paulo.

Em seu papel de Escola de Computação, a ERAD-SP ofereceu uma programação dividida em 3 dias, com um total de 3 sessões técnicas de apresentação de artigos científicos, 4 palestras convidadas, 4 palestras empresariais e 5 minicursos. Todas as sessões foram transmitidas ao vivo pelo canal da ERAD-SP no YouTube e podem ser revistas por qualquer interessado na área.

Como fórum de encontro, os membros da comunidade puderam se reencontrar e interagir com os palestrantes, ao vivo, nas salas do sistema ConferênciaWeb da RNP e nas salas de bate-papo por texto e voz do sistema Discord. A realização do evento de maneira online expandiu a comunidade para membros de fora do Estado de São Paulo. Além dos inscritos que se identificaram como membros das instituições de ensino e pesquisa de SP, o evento teve entre seus 56 inscritos participantes de instituições de ensino de BA, DF, ES, GO, MA, MG, MS, PB, PI, PR, RJ, RO, RS e SE. Em seu canal do YouTube, a ERAD-SP alcançou pelo menos 324 espectadores únicos, com destaque para o fato de que cerca de 11,5% das visualizações do canal ocorreram a partir do Peru.

Os Anais da XII Escola Regional de Alto Desempenho de São Paulo (ERAD-SP 2021) trazem os artigos selecionados e apresentados na edição do evento realizada virtualmente. Nesta edição os anais incluem 10 artigos do Fórum de Iniciação Científica e 6 artigos do Fórum de Pós-Graduação. Esses trabalhos foram selecionados através de um processo de revisão por pares do tipo "blind". Todos os artigos receberam pelo menos 3 revisões cada. Os artigos que integram este volume foram submetidos em 19/03/2021, aceitos para publicação em 09/04/2021, tendo a versão final submetida em 18/04/2021.

O melhor artigo de cada um dos fóruns foi premiado com o Prêmio de Melhor Artigo da ERAD-SP. O melhor artigo do Fórum de Iniciação Científica foi atribuído ao artigo *Comparação dos modelos BERT e Snips para compreensão de linguagem natural para assistente virtual ADA*,de Leonardo Costa Santos, Antonio de Carvalho Jr. e Alfredo Goldman (USP). Recebeu o prêmio de melhor artigo do Fórum de Pós-graduação o artigo

Improved Failure Detection and Propagation Mechanisms for MPI, de Pedro Rosso e Emilio Francesquini (UFABC).

Para mais informações sobre a ERAD-SP 2021 e sobre os artigos aceitos visite o site desta edição do evento.

Ainda que a situação epidemiológica não tenha permitido a realização do evento nos moldes tradicionais, acreditamos que a edição 2021 da ERAD-SP cumpriu seu papel de divulgação das atividades científicas realizadas na área da Computação de Alto Desempenho por pesquisadores do Estado de São Paulo. Mesmo que a comunidade esteja distante fisicamente por conta da pandemia, a ERAD-SP 2021 aproximou a comunidade e foi capaz de atrair o interesse de pesquisadores e de praticantes de outros estados e países.

A coordenação agradece a todos os autores e ministrantes de minicursos por compartilharem com a comunidade seus conhecimentos. Agradece também todos os membros do comitê de programa, que exerceram em tempos de sobrecarga de trabalho o papel fundamental de formar ajudar a comunidade na formação de novos pesquisadores na área. Por fim, agradece pelo apoio financeiro proporcionado pelas empresas Amazon, Atos, LexisNexis Risk Solutions e SDC e pelo apoio técnico da Rede Nacional de Ensino e Pesquisa (RNP).

A XIII edição da ERAD-SP será organizada pela Universidade Federal de São Carlos (UFSCAR). Esperamos poder reencontrar toda a comunidade de Computação de Alto Desempenho do estado lá.

Daniel Cordeiro

Emilio Francesquini

Coordenação Geral

Hélio Crestana Guardia

Ricardo Menotti

Coordenação de Programa e Minicursos

Minicursos

Minicurso I: Otimização de Programas Paralelos com uso do OpenACC

Palestrantes: Evaldo B. Costa (UFRJ) e Gabriel P. Silva (UFRJ)

Resumo: Este minicurso tem por objetivo apresentar técnicas de otimização de programas paralelos com uso de diretivas do OpenACC através de ferramentas que executem uma análise completa de desempenho do código para identificação de regiões paralelizáveis e quais métodos podem ser aplicados. O OpenACC é um modelo de programação para computação paralela que pode ser executado em diversos tipos de arquiteturas: multicore, manycore e aceleradores. Assim, neste minicurso são avaliados os efeitos dos componentes de hardware como número de processadores, hierarquia de memória e aceleradores sobre o desempenho de programas paralelos. Ressaltam-se as modificações que devem ser feitas no código para explorar com vantagem as características dos recursos computacionais, avaliando os seus respectivos impactos no desempenho de um programa.

Minicurso II: Uma introdução ao processamento de fluxos de dados com Apache Flink

Palestrante: Pedro Silva (Hasso Plattner Institute, Alemanha)

Resumo: A proposta do minicurso é de apresentar os principais fundamentos teóricos relacionados ao processamento de fluxos de dados e permitir aos participantes um primeiro contato com o sistema de processamento de fluxos de dados Apache Flink, que atualmente é uma das principais ferramentas para esse fim. Utilizando dados provenientes de casos de utilização industriais (e.g., medidores de energia inteligentes, táxis conectados e sismógrafos), os participantes serão familiarizados com a arquitetura, restrições e principais características de aplicações de processamento de fluxo utilizando o Apache Flink como exemplo de implementação.

Minicurso III: Introdução à Programação com Memória Persistente

Palestrantes: Alexandro Baldassin (UNESP) e Emilio Francesquini (UFABC)

Resumo: Este minicurso é uma introdução à programação com memória persistente (PM). Nele apresentaremos a motivação para uso dessa nova tecnologia; o suporte atual disponibilizado por processadores e sistemas operacionais; os problemas de consistências de dados que podem acontecer; e como utilizar abstrações de mais alto nível (como transações) para resolvê-los. O minicurso também apresenta uma série de exemplos práticos utilizando o Intel PMDK para programação de várias estruturas de dados persistentes.

Minicurso IV: Programação em GPU no Ambiente Google Colaboratory

Palestrantes: Ricardo Ferreira (UFV), Michael Canesche (UFV) e Westerley Carvalho Oliveira (UFV)

Resumo: Este trabalho apresenta um minicurso sobre a programação de GPU no Google Colaboratory (Colab). Tem como objetivo fornecer um material introdutório para o ensino e pesquisa com GPU no ambiente Colab, democratizando e desmistificando o acesso às GPUs. O Colab disponibiliza 4 tipos de GPU (K80, P4, P100 e T4) de forma gratuita e pode ser acessado, inclusive, por celular. Inicialmente, uma introdução à programação GPU é apresentada, juntamente com a configuração do Colab. Em seguida, vários exemplos de código são apresentados para ilustrar a execução, os recursos e as técnicas básicas de programação para GPU. Os laboratórios do minicurso também incluem material adicional com exercícios e sugestões de atividades extras.

Minicurso V: Ciência Reprodutível para Experimentos em Computação de Alto Desempenho

Palestrantes: Pedro Bruel (USP), Lucas Mello Schnorr (UFRGS) e Alfredo Goldman (USP)

Resumo: Neste minicurso apresentaremos uma introdução à Ciência Reprodutível, orientada por discussões sobre os problemas comumente enfrentados na ciência experimental, e por apresentações de possíveis soluções que promovam a reprodutibilidade no contexto de experimentos computacionais.

Palestras

Palestra Convidada I: Memory, Power, and Reconfigurable Node-Level Resource Management for HPC

Palestrante: Swann Perarnau (Argonne National Laboratory)

Resumo: Several major HPC facilities have publicly announced that their exascale-class systems will be heterogeneous, featuring multiple accelerator devices linked to multicore processors serving as hosts. These complex node topologies and the increasingly complex scientific workloads targeting these systems have made node-level resource management an important challenge for the performance of the overall production platform, and is of major interest to the system software community. As usual, node-level system software must also compose with sometimes competing interests, from scientific users, to programming model design, runtimes, and system-level resource management policies.

This presentation will provide an overview of the challenge of node-level resource management at exascale, focusing on two critical resources (memory and power), and discuss the work done within the Argo team of the Exascale Computing Project towards building a more adaptable system software stack able to arbitrate between the many stakeholders involved.

Palestra Convidada II: Generating Code for the new IBM Power10: Hardware Matrix Multiply Assist and Pattern Matching in a Compiler (and the Story of an Alberta/Unicamp/IBM Collaboration)

Palestrante: J. Nelson Amaral (University of Alberta, Canadá)

Resumo: To support both Artificial Intelligence and High-Performance Computing workloads, the new IBM POWER10 processor introduces a hardware computing unit called the Matrix Multiply Assist (MMA). This talk describes a collaboration between the University of Alberta, the University of Campinas, and IBM to make the performance improvement enabled by MMA accessible to both HPC and AI workloads. We explore two ways to do so: integrating MMA into highly specialized linear-algebra libraries; and creating a compilation path that is aware of MMA and capable of using it effectively. Once MMA is properly integrated into a library, a performant solution is for an application to invoke the library to execute a numerical computation, such as matrix multiplication. However, could an application that contains a matrix-multiplication operation natively benefit from the library performance boost? Also, how can a compiler recognize that a segment of code is performing a matrix multiplication and thus generate the appropriate MMA instructions? The answer to both questions is to use efficient and robust pattern recognition in the compiler internal representation. This talk describes how we developed a much more robust pattern recognition methodology in the LLVM compilation system. It also discuss the implementation of a layered approach for efficient matrix multiplication, which were until now only available in numerical libraries, into an LLVM-based compiler-only software path. Making an efficient compiler-only path available is important for situations when invoking a function from a library is not an option.

Palestra Convidada III: Performance Engineering for HPC: the role of Source-to-Source Compilers and why we need them

Palestrante: João M. P. Cardoso (Universidade do Porto, Portugal)

Resumo: High-Performance Computing systems have become heterogeneous. In such systems, multicore CPUs coexist with hardware accelerators (such as GPUs and FPGAs). Although these systems can provide higher performance, their programming is more complex, prone to errors, requires longer development cycles and experts, especially when stringent performance figures need to be achieved. In this talk I will present some of the main challenges and current research avenues to mitigate development and help software programmers to take advantage of some of the heterogeneous resources present in the HPC systems. I will both consider HPC in embedded computing and in supercomputing and will focus the talk on the role of source-to-source compilers as first citizen tools to help and automate important performance engineering tasks.

Palestra Convidada IV: A Arquitetura RISC-V, suas Implementações e o Estado de seu Software Livre

Palestrante: Carlos E. de Paula (Red Hat, Inc.)

Resumo: A arquitetura RISC-V vem para abrir a última camada do Open Source, o hardware. Nessa palestra vamos falar sobre a arquitetura aberta RISC-V, suas implementações atuais em hardware e o suporte das aplicações a essa arquitetura. Linux, containers e sistemas operacionais embarcados já estão disponíveis e em uso por diversos grupos e sua evolução se dá de forma rápida no meio acadêmico e na indústria.

Palestra Empresarial I: HPC e AI/ML na AWS: clusters rápidos, simples, e baratos

Palestrante: Paulo Aragão (Amazon Web Services)

Palestra Empresarial II: GPUs, HPC, A.I. — Densidade e opções

Palestrante: Guilherme Friol (SDC)

Palestra Empresarial III: HPCC Systems: últimos avanços e oportunidades para uso e pesquisa em Big Data

Palestrantes: Hugo Watanuki e Artur Baruchi (LexisNexis Risk Solutions)

Palestra Empresarial IV: Supercomputadores e Computação Quântica – quais as tecnologias que temos disponíveis para criar computadores com alta capacidade de processamento

Palestrante: Genaro Costa (Atos Bull)