Dagstuhl Seminar 24201: Discrete Algorithms on Modern and Emerging Compute Infrastructure

May 12 – May 17, 2024

At-a-Glance Schedule

<table>
<tr><th></th><th>Monday</th><th>Tuesday</th><th>Wednesday</th><th>Thursday</th><th>Friday</th></tr>
<tr><td>7:30-8:45 AM</td><td colspan="5">Breakfast</td></tr>
<tr><td rowspan="2">09:00 AM</td><td>Welcome & Introductions</td><td>Plenary Talk: A Trail Guide to Parameterized Graph Algorithms in Practice (Blair Sullivan)</td><td>Plenary Talk: Challenges for Computational Graph Algorithms (Jon Gilbert)</td><td rowspan="2">Focus Session: Algorithms: Beyond the Static (Kathrin Hanauer, Helen Xu, Manuel Penschuck, Quanquan C. Liu)</td><td rowspan="2">Focus Session: Wafer Scale Computing: Fine Grain Parallelism and Rethinking Parallel Computing (Rob Schreiber)</td></tr>
<tr><td>Plenary Talk: Graph Algorithms in Unsettled Times (Alex Pothen)</td><td>Presentation of Working Groups</td><td>Working Groups</td></tr>
<tr><td>12:15 PM</td><td colspan="5">Lunch</td></tr>
<tr><td>02:00 PM</td><td>Focus Session: The Future of Computing (Bruce Hendrickson, Chris Goodyer, Oded Green, Jakob Engblom)</td><td>Focus Session: Quantum Computing (Ilya Safro, Eleanor Rieffel, Erik Boman)</td><td>Hike</td><td>Working Groups</td><td></td></tr>
<tr><td>06:00 PM</td><td colspan="4">Dinner</td><td></td></tr>
</table>

Morning coffee break: from 10:00 AM

Afternoon coffee break: from 03:00 PM

Monday, May 13, 2024

Morning (9am-12am including coffee/tea break between 10am and 11am)

- Welcome
- **Introductions** (max. 5 min each; single slide)
- Plenary Talk: **Graph Algorithms in Unsettled Times** (Alex Pothen, Purdue University, USA)

  We live in the time of a changing and uncertain computing environment. The end of Moore's law and Dennard scaling has led to searches for new computing substrates, from chiplets, accelerators, wafer scale chips, neuromorphic computers, quantum computers, etc. The growth of data science has led to graph models for unstructured data of increasing sizes for downstream inference tasks.
  The artificial intelligence revolution has led to the study of large scale graph neural networks, but also learning augmented algorithms, where machine learning concepts are used to provably improve the quality of the solution or the run time of the algorithm.
  All of these factors lead to the development of new models for algorithm design, including approximation algorithms, distributed algorithms, online algorithms, semi-streaming algorithms, dynamic algorithms, fixed parameter algorithms, etc.

  I will survey of some of these topics in this introductory talk.
  Several of the workshops and panels at this Dagstuhl seminar will consider these topics in more detail, and my hope is that these discussions could serve as a helpful vade mecum for algorithms researchers.

Afternoon (2pm-5:30pm including coffee/tea break between 3pm and 4pm)

Focus Session: **The Future of Computing** (Organizer: Bruce Hendrickson, Lawrence Livermore National Laboratory, USA)

A combination of technological and market forces are driving rapid changes in computer architectures and system designs. These changes will have significant impacts for algorithms and software. This session will review the drivers behind these changes and provide some thoughts on what the future might look like. The goal of this session is to provide context for much of the remainder of the Dagstuhl program.

Speakers:

Chris Goodyer, ARM: Building the future of computing

Oded Green, Nvidia: The Future of Accelerated Combinatorial and Sparse Applications

Jakob Engblom, Intel: Just Add Accelerators – The Answer to Everything?

Tuesday, May 14, 2024

Morning

Plenary Talk: **A Trail Guide to Parameterized Graph Algorithms in Practice** (Blair Sullivan, University of Utah, USA)

This talk will aim to introduce the audience to a mixture of classic and recent algorithmic techniques which originate primarily in the theoretical computer science community and exploit the non-uniformity of computational hardness. In particular, the focus will be on ideas that I think hold promise for real-world network analysis in the next decade -- despite often being completely impractical in their current form! I will also briefly discuss lessons learned from applications where some of these techniques have been engineered successfully to impact domain science, and highlight what I see as key challenges in the space.

Presentation of Working Groups (15min + Q&A each)

- Paolo Bientinesi: **Single-instance vs. Batched vs. Sequence of Problems**
- Chris Goodyer: **A New Standard for Cross-architecture Sparse Linear Algebra**
- Johannes Langguth: **Towards a Theory of Tile-centric Computation**
- Johannes Lotz: **Algorithmic Differentiation on Modern and Emerging Compute Infrastructure**

Afternoon

Focus Session: **Quantum Computing** (Organizer: Ilya Safro, University of Delaware, USA)

Speakers:

Ilya Safro: Introduction to "traditional" quantum computing at the graduate course level

Eleanor Rieffel: Overview of quantum optimization

Ilya Safro: Multilevel hybrid quantum-classical computing

Erik Boman: Neuromorphic computing

Wednesday, May 15, 2024

Morning

Plenary Talk: **Challenges for Computational Graph Algorithms** (John Gilbert, University of California, USA)

Though applications of graphs go back at least to Euler in 1736, the age of large-scale computation with graphs arguably began in the 1970s. Computing efficiently with graphs has always been hard, but the challenges have evolved quite a bit over the past 50 years. In this talk I will speculate on what key challenges the designers and users of high-performance graph computation will face during the next 10 years. I'll organize the challenges roughly into three categories: machine architecture; algorithms; and productivity.

Working Groups

Afternoon

Gentle Hike

Thursday, May 16, 2024

Morning

Focus Session: **Algorithms: Beyond the Static** (Organizer: Kathrin Hanauer, University of Vienna, Austria)

In the past, algorithms were often viewed as static entities, optimized for specific problems and processing constraints. However, with the explosion of big data, the advent of modern CPUs and the broad availability of computing clusters, the algorithm landscape has undergone a profound shift. Today, we see - among others - a growing demand for algorithms that can adapt dynamically to changing inputs, leverage parallel processing, harness the power of distributed computing, and boost performance by integrating techniques from machine learning.

In this session, we will give an introduction and overview over these modern algorithms and discuss both theoretical advancements and practical applications.

Speakers:

Kathrin Hanauer, University of Vienna: Dynamic Graph Algorithms in Theory and Practice

Helen Xu, Georgia Institute of Technology: Graph Representations and Optimization

Manuel Penschuck, Goethe University: Sampling Practical and Scalable Sampling Algorithms

Quanquan C. Liu, Yale University: Learning-Augmented Algorithms

Afternoon

Working Groups

Friday, May 17, 2024

Morning

Focus Session: **Wafer Scale Computing: Fine Grain Parallelism and Rethinking Parallel Computing** (Organizer: Rob Schreiber, Cerebras Systems, USA)

I will explain how wafer-scale computing currently works by detailing the hardware, architecture, and programming paradigms of the Cerebras machines, the only instance of commercial wafer-scale computers today.

The CS-3 incorporates all memory and processing on one wafer, a wafer that contains 900,000 processing elements. With 48KB of local memory, a PE cannot hold very much data. On the other hand, access to that data is at the same rate as peak speed computation. Most interesting, the mesh interconnect has single-clock latency for sending a message (of 4 bytes) to a mesh

neighboring PE, and the network can sustain a 4 byte message to and from each neighbor on every clock.

The wafer is therefore a working instance of processing co-located with memory. While it is distributed memory from the addressing perspective, the extreme interconnect performance allows programmers to treat distributed tensors as if they were shared – shared objects in a distributed memory substrate. This finds uses in graph computing, sparse matrix vector products, neutron transport applications, for some examples.

The absence of both memory walls and slow, high-overhead, high-latency interconnect permits very fine grained parallel applications that achieve excellent performance. This in turn allows strong scaling in which each PE holds only a few words of the problem data, taking full advantage of the easy accessibility of data on near neighbor PEs. Thus, strong scaling is quite successful, which reduces runtimes for problems of the scale that fit the wafer by two orders of magnitude, allowing applications that are impossible with conventional systems.

I will cover some use cases and give an outline of how the system can be programmed using the Cerebras SDK.

Following the talk, I will propose discussion topics for working groups around whether, and how, these changes in the computing substrate will affect what is achievable in discrete algorithms, how to modify algorithms if at all, and what might be desirable to have in the hardware or software of the system.